# Supplementary Materials

Ali Ayub
aja5755@psu.edu
Alan R. Wagner
alan.r.wagner@psu.edu
The Pennsylvania State University
State College, PA, USA

## 1 Algorithms

Algorithm 1 describes the *Agg-Var* clustering algorithm used as part of CBCL to generate centroid pairs for the RGB-D images. Algorithm 2 explains the category merging process using the Silhouette indices of the training images of all the categories.

**Algorithm 1** *Agg-Var* Clustering

**Inputs:** $F_{RD}$ **:** RGB and depth feature map pairs from a training dataset with $m$ categories and $N$ samples
$D$**:** Distance threshold hyperparamter
$w_{rgb}$, $w_D$ **:** RGB and depth fusion weights hyperparameters
**Output:** A collection containing a set of centroid pairs for each category of indoor scene, $C_{RD} = C^1_{RD}, C^2_{RD}, ..., C^m_{RD}$

$F^j_{RD}$**:** the set of RGB and depth feature map pairs labeled as category $j$, with $N_j$ samples, where $f^j_{rd_i}$ represents the $i$-th sample in $F^j_{RD}$.
$C^j_{RD}$**:** set of RGB and depth centroid pairs for category $j$, where $c^j_{rd_i}$ represents the $i$th centroid pair in $C^j_{RD}$.
**for** $j = 1; j \leq m$ **do** $C^j_{RD} \leftarrow \{f^j_{rd_i}\}$
**for** $j = 1; j \leq m$ **do**
    **for** $i = 2; i \leq N_j$ **do**
        $d_{min} \leftarrow \min_{l=1,..,size(C^j_{RD})} dist_{RD}(c^j_{rd_i}, f^j_{rd_i})$
        $x \leftarrow \text{argmin}_{l=1,..,size(C^j_{RD})} dist_{RD}(c^j_{rd_i}, f^j_{rd_i})$
        **Set** $c^j_{rd_x}$ to be the nearest centroid pair
        **Set** $w^j_x$ to be the number of images clustered
        in the $x$th centroid pair of category $j$
        **if** $d_{min} < D$ **then**
            Use Eq. 2 to update centroid pair $c^j_{rd_x}$
        **else**
            $C^j_{RD}.append(f^j_{rd_i})$

**Algorithm 2** Category Merging

**Input:** $C = \{C^1_{RD}, ..., C^m_{RD}\}$ ▷ current class centroid pair sets
$F_{RD}$ **:** RGB and depth feature map pairs from a training dataset with $m$ categories and $N$ samples
**Output:** $Y_{merged} = \{1, ..., t\}$ ▷ New labels of $N$ images belonging $t \leq m$ categories

Repeat Until $z^j_{conf} < 0.25 \vee Y^j_{conf} \neq Y^{Y^j_{conf}}_{conf}$
Calculate $S = \{s^1_1, ..., s^j_i, ..., s^m_N\}$ ▷ silhouette values for $N$ RGB-D images for all $m$ categories
Calculate $Y = \{y_1, ..., y_N\}$ ▷ Label of closest centroid corresponding to each image
**for** $j = 1; j \leq m$ **do**
    $z^j_{conf} \leftarrow \frac{\sum^N_{i=1, y=j, s^y_i \leq 0} s^y_i}{N_j}$ ▷ percentage of images in class $j$ with Silhouette value <=0
    $Y^j_{conf} \leftarrow \max_{i=1,...,m} \sum^N_{k=1, s^i_k \leq 0, y_k = i} 1$ ▷ most common category among images with $s \leq 0$
    **if** $z^j_{conf} > 0.25 \wedge Y^j_{conf} = Y^{Y^j_{conf}}_{conf}$ **then**
        $Y^j_{conf} = j$ ▷ Merge categories by assigning the same label

## 2 Hyperparameter Values

In order to train VGG16 on depth data, for both of the datasets, we train it for 200 epochs with an initial learning rate of 0.01 and divide it by 50 after 60, 120 and 160 epochs. Stochastic gradient descent with momentum of 0.9 was used with minibatches of size 64 and trained with cross-entropy loss with weight decay of 0.0005. The hyperparameters for CBCL (distance threshold $D$, *fusion weight* $w_D$ and number of closest centroids for classification $n$) were tuned using cross-validation on the training set. For the SUN RGB-D dataset, the values of the hyperparameters $D$, $n$, and $w_D$ are tuned to 85, 17, and 0.73 respectively. For NYU Depth V2 dataset, $D$, $n$, and $w_D$ are tuned to 95, 7, and 0.76, respectively. For both the datasets, $w_R$ is set to 1.0. Even though the hyperparameters for the two datasets are a little different, CBCL is robust to changes in the hyperparameter values for a large range (Section 4.3.2 in the paper).

# 3 Confusion Matrices

The confusion matrices for SUN RGB-D and NYU Depth-V2 datasets are shown below:

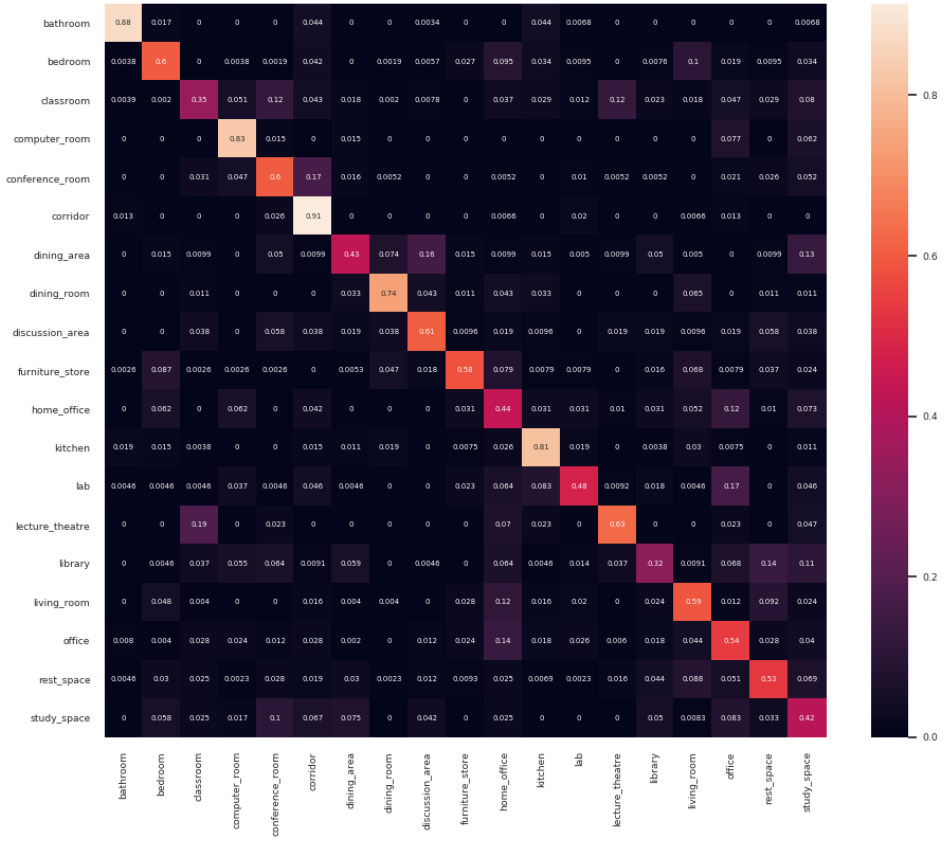


Figure 1: The confusion matrix for CBCL on the SUNRGB-D dataset. The vertical axis depicts the ground truth and the horizontal axis dpicts the predicted labels.

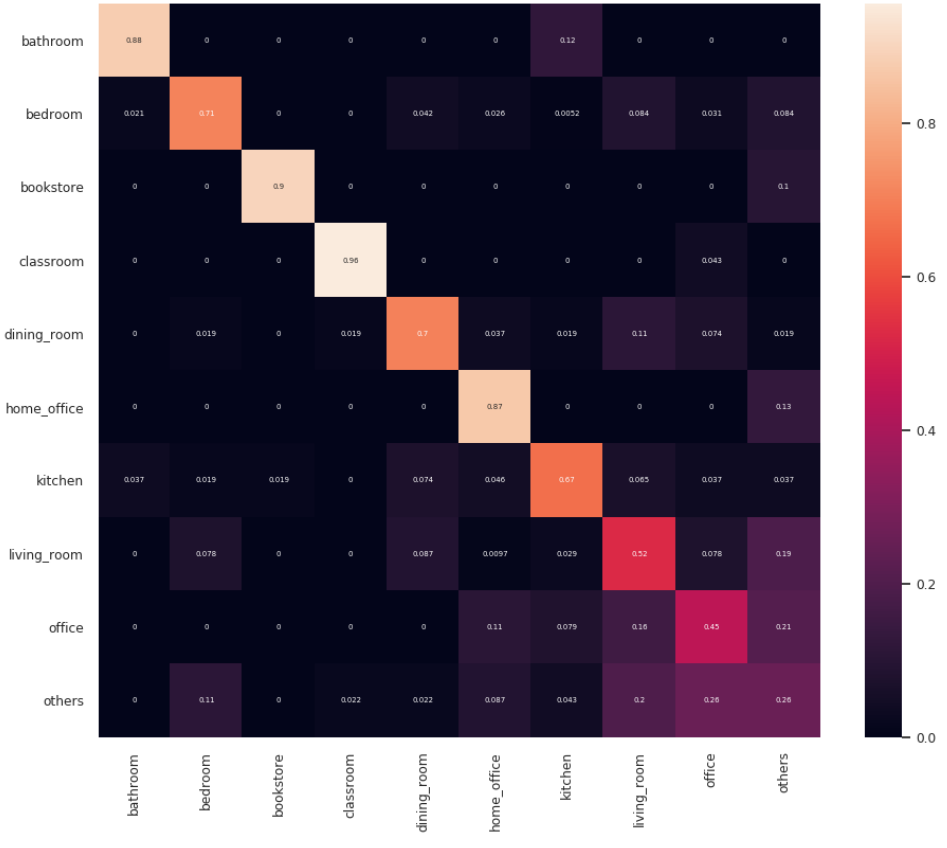


Figure 2: The confusion matrix for CBCL on the NYU Depth V2 dataset. The vertical axis depicts the ground truth and the horizontal axis depicts the predicted labels.

# 4 Ablation Study on NYU Depth-V2 Dataset

Table 1 presents the results for an ablation study on the NYU Depth-V2 dataset. We compare CBCL against the same three baselines (VGG, Alexnet and ResNet18) and two clustering hybrids (Agglomerative and K-means) as for SUN RGB-D dataset. Similar to the ablation study for SUN RGB-D dataset, the results indicate that most of the performance gain for our approach comes from better classification of RGB features and *Agg-Var* clustering.

| **Methods** | **RGB** | **Depth** | **Fusion** |
|---|---|---|---|
| VGG Baseline | 57.34 | 49.15 | 60.18 |
| Alexnet Baseline | 59.50 | 49.30 | 60.60 |
| ResNet18 Baseline | 59.80 | 52.30 | 63.80 |
| Agglomerative RGB-D | 63.27 | 45.46 | 66.30 |
| K-means RGB-D | 61.13 | 43.02 | 63.97 |
| **CBCL** | **66.40** | 49.50 | **70.91** |

Table 1: Ablation study results on the NYU Depth-V2 dataset

# 5 Examples of Conceptually Similar Categories


Figure 3: Additional examples of images from conceptually similar scene categories. These images are from the SUN RGB-D dataset.

# 6 Silhouette Index Values and Category Merging on the SUN RGB-D Dataset

Table 2 presents the result from calculating the silhouette index values for each image in each category of the SUN RGB-D dataset. The table lists the percentage of images with silhouette value less than or equal to zero and the category most often confused. The italicize categories have greater than 25% of images with $s <= 0$ and were confused bi-directionally, i.e. images with $s \leq 0$ of one category of the pair were mostly confused with the other category in the pair and vice versa. The algorithm for merging the categories is described in Section 1. The procedure is repeated recursively and stopped if any of the two criteria are not met: categories should have more than 25% images with $s \leq 0$ and they should be confused bi-directionally.

| **Categories** | **Percent of images with $s$<=0** | **Confused Category** |
|---|---|---|
| bathroom | 4.0 | bedroom |
| bedroom | 8.1 | living_room |
| *classroom* | *28.5* | *lecture_theatre* |
| computer_room | 7.6 | classroom |
| *conference_room* | *27.9* | *study_space* |
| corridor | 11.7 | bedroom |
| dining_area | 12.8 | rest_space |
| dining_room | 13.8 | discussion_area |
| discussion_area | 12.4 | dining_room |
| furniture_store | 7.4 | bedroom |
| *home_office* | *29.2* | *office* |
| kitchen | 6.8 | furniture_store |
| lab | 9.5 | kitchen |
| *lecture_theatre* | *32.0* | *classroom* |
| library | 22.3 | rest_space |
| *living_room* | *30.4* | *rest_space* |
| *office* | *28.4* | *home_office* |
| *rest_space* | *29.7* | *living_room* |
| *study_space* | *31.9* | *conference_room* |

Table 2: Percentage of images in each of the 19 categories of SUN RGB-D dataset that have silhouette's index $s <= 0$ and the corresponding categories with which most of these low silhouette value images are closer to.

# 7 Category Merging for NYU Depth V2 Dataset

The NYU Depth V2 dataset includes a large number of categories with a few training images which traditionally are combined into one "others" category [1] to reduce the total number of categories from 27 to 10. The 17 categories that typically compose the "others" category, however, are conceptually different from each other. The silhouette index values were calculated on all 27 categories. The same criteria was used to identify mergeable categories for the SUN RGB-D dataset. The recursive analysis indicated that the following category should be merged: (home_office, foyer, study, basement), (dining_room, dinette), (cafe, bookstore, furniture_store), (kitchen, office_kitchen), (study_room, office, printer_room), (computer_lab, conference_room, classroom), (living_room, playroom, student_lounge, reception_room). The categories exercise_room, home_storage, indoor_balcony and laundry_room did not meet the criteria for merging so they remained in the "others" category. The result was a total of 10 categories but under a different merging scheme from that used by [1].

On this updated dataset, CBCL achieves **78.95%** mean class accuracy which is 8.04% higher than on the original dataset. The VGG baseline achieves 67.24% accuracy which is 7.06% higher than on the original dataset, although still 11.71% lower than CBCL. These results are similar to the SUN RGB-D results.

# References

[1] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.