# Training semantic segmentation networks with temporal consistency

1st Feiyu Qin
*University of Science and Technology of China*
Heifei, China
feiyuqin@mail.ustc.edu.cn

2rd Xuejin Chen
*University of Science and Technology of China*
Heifei, China
xjchen99@ustc.edu.cn

***Abstract*—Semantic segmentation is a fundamental task in indoor scene understanding. Most previous supervised approaches rely on densely annotated image datasets. Due to the limited amount of images with segmentation labels, the performance and robustness of existing networks is greatly limited. In this work, we exploit spatical and temporal correlation in video frames to improve the performance of segmentation networks. Two effective semi-supervised learning strategies are proposed to propagate the information from a few labeled frames to adjacent video frames. First, we augment training data for supervised semantic segmentation networks by generating pseudo ground-truth for neighboring frames from a labeled frame using filtered homography transformation. Second, we introduce a self-supervised loss function to ensure temporal consistency between the segmentation results of adjacent frames. The experimental results show that our proposed method outperforms state-of-the-art techniques for semantic segmentation on NYU-Depth V2 dataset.**

***Index Terms*—Indoor scene, semantic segmentation, temporal consistency, label propagetion**

## I. Introduction

Semantic segmentation is a fundamental part of scene understanding. For indoor scenes, the cluttered backgrounds, large variety of scenes, object occlusions and various illumination pose a series of challenges for accurate semantic segmentation. In recent years, a great deal of studies have been conducted for indoor scene semantic segmentation, and they can be mainly divided into three groups: semantic segmentation of a single RGB image, semantic segmentation of RGBD images, and multi-task learning methods.

**Semantic Segmentation of a Single Image.** Fully convolutional network (FCN) [1] is a pioneering work for pixel-wise segmentation. It first converts existing convolutional neural networks (CNN) constructed from classification to semantic segmentation. To overcome the limitations of FCN that the network limited by a fixed-size receptive field, Noh *et al.* [2] proposed a novel deconvolution algorithm to segment finer object structures. Bayesian SegNet [3] performs visual scene understanding with a measure of model uncertainty to produce a probabilistic segmentation result. To capture semantic correlations between neighboring patches and exploit patch-patch contextual information, Lin *et al.* [4] formulated conditional random fields (CRFs) with CNN-based pairwise potential function. For generating fine prediction, a multi-path refinement network is proposed to effectively exploit multi-level features and refine the prediction step by step [5].

**Semantic Segmentation of RGB-Depth Images.** With the popularity of affordable depth-cameras, many techniques have been proposed for semantic segmentation of RGBD images. Gupta *et al.* [7] extracted features from RGB and depth data and integrated them for object detection and segmentation. A novel long short-term memorized context fusion (LSTM-CF) model is proposed [8] to fuse contextual information from multiple sources such as RGB images and depth data. Cheng *et al.* [9] rethink the relationship between RGB images and depth, and propose a locality-sensitive deconvolution network to refine object boundaries of segmentation result as well as a gated fusion layer to combine features of two modes. Park *et al.* [10] propose a multi-modal feature fusion block for fusing the multi-level RGB-D features and integrate these blocks into RefineNet.

**Multi-task Learning.** While semantic segmentation, depth estimation, and other tasks for indoor scene understanding share features, especially at shallow layers, many techniques have been proposed to train networks for multiple tasks in order to complement each other. Eigen and Fergus [11] address three different tasks including depth prediction, surface normal estimation, and semantic labeling using a single multi-scale convolutional network. Prediction-and-distillationn network (PAD-Net) [12] predicts a set of intermediate tasks including depth, surface normal, semantic and contour estimations, and then uses the predictions from these tasks as multi-modal distillation modules' inputs for final tasks. A novel joint Task-Recursive Learning (TRL) [13] framework for semantic segmentation and depth estimation is proposed by Zhang *et al*. In TRL, two tasks are alternately processed in the decoder to improve each other. Jiao *et al.* [14] present an attention-driven loss for network supervision and a synergy network to learn the information sharing strategies. Although the multi-task learning methods improve the segmentation results, they require a lot of supplementary information or even dense-annotated data for other tasks.

Above approaches provide a lot of novel ideas to improve the segmentation results. However, the performance and robustness of these methods is greatly suppressed by limited densely annotated training data. Obtaining a large amount of annotated semantic labels for images is expensive and time-

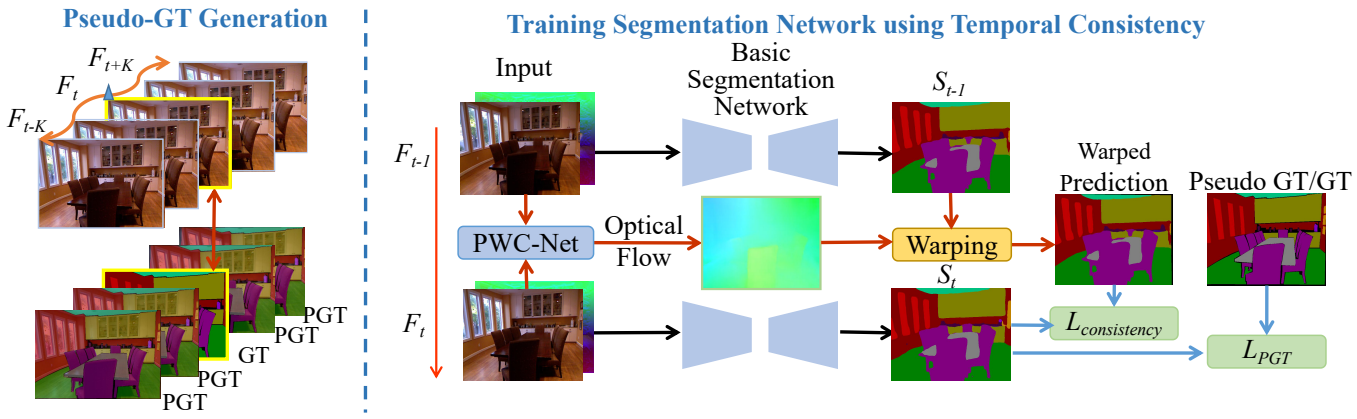



Fig. 1. The diagram of our proposed method. Left: we propagate the ground truth from a labeled frame to its adjacent frames by image warping. Right: the flowchart of our training strategy using temporal consistency. It mainly consists of three steps: 1) The current frame $F_t$ passes through a basic semantic segmentation network to generate the prediction $S_t$ 2) Warping the prediction $S_{t-1}$ of previous frame $F_{t-1}$ depends on optical flow generated from PWC-Net [6] 3) Warped prediction of $F_{t-1}$ and generated PGT joint action on the network by $L_{Consistency}$ and $L_{PGT}$ respectively.

consuming. In contrast, it is economical to capture videos, and the temporal correlation between adjacent frames naturally provides more information and constraints for the semantic labels. Therefore, we propose to improve existing networks for semantic segmentation by exploiting the temporal correlation between video frames in two aspects.

First, instead of computing loss for each annotated image individually, we train a semantic segmentation network with a self-supervised loss function to ensure the temporal consistency between adjacent frames in video sequences. Second, with only a few labeled video frames, we propagate these labels to its neighboring frames with geometric transformations and generate a large number of new frames with pseudo ground truth (PGT) for network training. Compared to common data augmentation operations, it significantly increases the diversity of training data. Many image alignment methods could be used to build correspondence of two images for label propagetion. The correspondence problem between two adjacent frames in a video is often formulated as optical flow estimation. However, due to optical flow eatimation depends on the low-level features of the image to find dense correspondence, it is often inaccurate and computationally expensive. Hence, in our work, the method of parametric model is selected for label propagetion. The commonly used parameterized model methods are rigid body transforms and homography transformation. Because there are obvious holes in the depth data collected by Kinect, and the data set does not provide the corresponding camera external parameters, we refer to homography transformation for label propagation.

With above two semi-supervised training strategies, our approach improves the results of fully supervised baseline method and achieves state-of-the-art performance on semantic segmentation for indoor scenes. The rest of this paper is organized as follows. Our methodology will be introduced firstly. Then, we will report the extensive experimental results as well as analyses. Finally we will draw a conclusion.

## II. Methodology

In this section, we begin by formulating the problem. And then we put forward our main methods. Fig. 1 shows the diagram of our proposed method. First, propageting the labels from the labeled frames to their adjacent frames to generate pseudo ground truth. Then we train a segmentation network with both the manually labeled frames and frames with pseudo labels, as well as a self-supervised loss function to ensure the temporal consistency between adjacent frames.

### A. Problem formulation

In a set of video sequences $X := (x_1, ..., x_m, x_{m+1}, ..., x_n)$ of which only a few frames $X_M := (x_1, ..., x_m)$ are manually labeled with pixel-wise semantic labels $Y_M := (y_1, ..., y_m)$. The remaining frames $X_U := (x_{m+1}, ..., x_n)$ are unlabeled. The goal in our semi-supervised learning is to use both labeled frames and part of unlabeled frames to train a semantic segmentation network that provides the prediction of unseen frames.

### B. Pseudo Ground Truth Generation

In a video sequence of which one frame $F_t$ is manually annotated with pixel-wise semantic labels $G_t$, we propagate these labels to its adjacent $2K$ frames $\mathcal{F} = \{F_{t-K}, \ldots, F_{t+K}\}$ via two steps. First, for each frame $F_k$ in $\mathcal{F}$, a homography matrix $\mathbf{H}_{tk}$ is first computed by looking for matched SIFT keypoints with RANSAC verification. In order to avoid large distortion or content changes, we only keep the video frames whose transformation matrix to the reference frame $F_t$

$$\mathbf{H}_{tk} = \begin{bmatrix} h_{00} & h_{01} & h_{02} \\ h_{10} & h_{11} & h_{12} \\ h_{20} & h_{21} & 1 \end{bmatrix} \quad (1)$$
$$s.t.\ \ 0.95 < \max\{|h_{00}|, |h_{01}|, |h_{10}|, |h_{11}|\} < 1.05,$$
$$\max\{|h_{02}|, |h_{12}|, |h_{20}|, |h_{21}|\} < 15.$$

Fig. 3 shows an example of the generated pseudo ground truth labels from a reference frame $F_t$. It can be seen that most

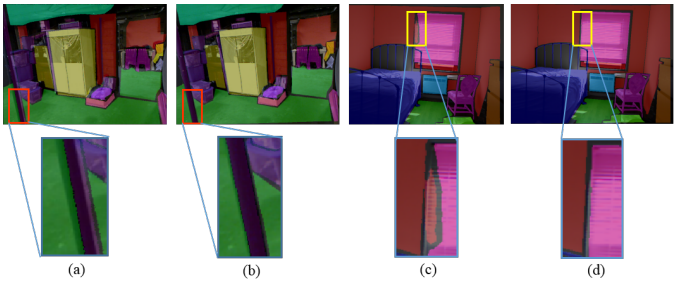



Fig. 2. Comparison of different methods of annotation propagetion. (a) PGT generated by rigid body transformation. The camera external parameters is evaluated by Bundler and the depth is filled using NYU-v2 official code. (b)(d) PGT of our method. (c) PGT generated by optical flow using PWC-Net [6].

of the semantic labels are well propagated and can be used as extra data for network training. However, there are still some noisy labels in the propagated PGT, especially at the object boundary regions. Therefore, we add a manual filtering process to remove the frames where there are obvious visual artifacts in the generated pseudo labels.

A toy example showing the results of label propagetion by our method, rigid body transformation and flow estimation is shown in Fig. 2. Compared to other PGT generated methods, our approach keeps the boundary of PGT better.

### *C. Training Segmentation Network using Temporal Consistency*

As Fig. 1 shows, we train a semantic segmentation network using two adjacent frames in a video instead of using a single labeled frame. First, two adjacent frames $F_{t-1}$ and $F_t$ are passed through a basic segmentation network to predict semantic labels $S_{t-1}$ and $S_t$. The basic segmentation network could be any existing network for semantic segmentation. We use RDFNet [10], which is a high-precision network for semantic segmentation of a RGBD image.

Second, the predicted semantic score map $S_{t-1}$ for frame $F_{t-1}$ is warped to $\hat{S}_t$ according to the optical flow $O_{t-1,t}$ between $F_t$ and $F_{t-1}$, which is estimated using an efficient and accurate model PWC-Net [6] from two frames. We implement a bilinear interpolation for 40 dimensional score vector of each pixel in $S_{t-1}$, where 40 is the total number of object classes in the dataset.

$$\hat{S}_t = Warp(S_{t-1}, O_{t-1,t}). \tag{2}$$

An argmax layer is then added on the warped score map $\hat{S}_t$ to generate the warped semantic prediction $W_t$.

Finally, the warped prediction $W_t$ acts as a self supervised item to constrain the consistency of prediction results between neighboring image frames. Meanwhile, the PGT of current frame $F_t$ acts as another supervisory item to guarantee that the prediction results are as accurate as possible. Combining these two items, our loss function is defined as

$$L_{seg} = \lambda L_{Consistency} + (1-\lambda) L_{PGT}, \tag{3}$$

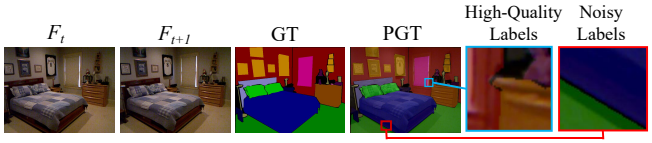



Fig. 3. Propagated pseudo ground truth from a reference frame $F_t$ to $F_{t+1}$. Most of the pseudo labels are reliable, while there are some noisy around object boundaries.

where, $L_{PGT}$ and $L_{Consistency}$ are cross-entropy loss, defined as

$$\begin{aligned} L_{Consistency} &= -\sum_{i=0} l(\mathbf{w}_i, \mathbf{y}_i), \\ L_{PGT} &= -\sum_{i=0} l(\mathbf{p}_i, \mathbf{y}_i), \\ l(\mathbf{p}_i, \mathbf{y}_i) &= -\sum_{k=0}^{c} \mathbf{p}_{ik} \log \mathbf{y}_{ik} \ , \ \mathbf{y}_i = f(x_i), \end{aligned} \tag{4}$$

where $x_i$ and $\mathbf{y}_i$ are a pixel of an input frame and the corresponding predicted output of the semantic segmentation model respectively. $\mathbf{y}_i$ is a $c$ dimensional probability vector for each pixel. $c = 40$ is number of categories. $\mathbf{p}_i$ and $\mathbf{w}_i$ are an one-hot vector for each pixel of the pseudo ground truth or ground truth label, and warped prediction $W_t$ from previous frame, respectively. $l(\cdot)$ is the softmax loss of one pixel, and $f(\cdot)$ is our basic segmentation network. By these two losses, we can constrain the accuracy and temporal consistency of the network prediction simultaneously.

## III. EXPERIMENTS AND EVALUATION

In order to evaluate the proposed method for indoor scene segmentation, we conduct experiments on a publicly available benchmark dataset (NYUD-v2) and show the superiority of our method.

**Dataset.** The NYUD-v2 dataset contains in total 464 diverse indoor scenes and the corresponding video sequences. For the task of semantic segmentation, 1449 images are manually annotated, among which 795 images are used for training and the remaining 654 images for testing. In our experiment, we map the semantic labels into 40 categories, similar as [7]. From the 795 manually labeled images for training, we propagated these GT labels to 14344 unlabeled frames, and filtered out 5897 frames by homography matrices, and filtered out 821 frames by manual check. We finally obtained 7626 frames with pseudo ground truth. Comparing with dense image annotation, filtering PGT for 273 videos took about 8 hours for a graduate student. Although the generated pseudo labels still have a certain amount of noise, the diversity of training set is greatly enriched.

The training images, including both manually labeled images and propagated images, are augmented by random horizontal flips with a possibility of 0.5, scaling with a randomly selected ratio in $\{0.7, 0.8, 0.9, 1.0, 1.1, 1.2, 1.3\}$ and random cropping. All our experiments are conducted on a single Nvidia Titan X Pascal GPU, using the Caffe framework.

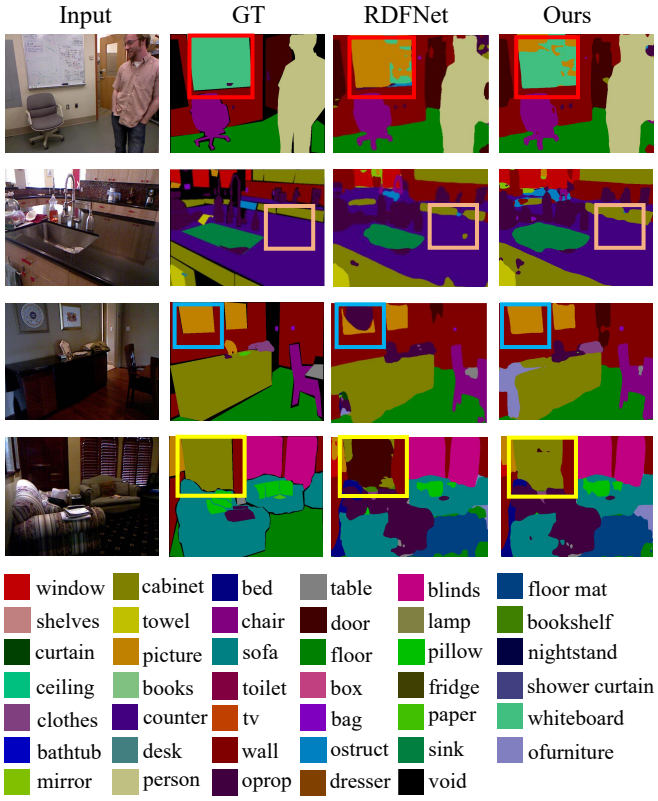



Fig. 4. Qualitative results on the NYUD-v2 dataset. In each row, we show the input image, the ground truth of semantic segmentation, the result of RDFNet, and the result of our method, from left to right, respectively. Our method predicts more accurate segmentation results, especially at hhe highlighted regions.

**Evaluation Metrics.** Following previous work [3]–[5], [7]–[14], we use three quantitative metrics to evaluate our semantic segmentation results, including pixel accuracy, mean accuracy and mean IoU. Denote $m_{kj}$ as the number of pixels of class $k$ classified as class $j$. The number of categories is $c$, and $M_k = \sum_j m_{kj}$ is the total number of pixels belonging to class $k$. And $N = \sum_k M_k$ denotes the number of all pixels in the test set. The three metrics are defined as follows:

$$\begin{aligned} \text{pixel accuray}: &\ \frac{1}{N}\sum_k m_{kk}, \\ \text{mean accuracy}: &\ \frac{1}{c}\sum_k \frac{m_{kk}}{M_k}, \\ \text{mean IoU}: &\ \frac{1}{c}\sum_k \frac{m_{kk}}{M_k+\sum_j m_{jk}-m_{kk}}. \end{aligned} \tag{5}$$

We set the RDFNet [10] as our baseline. However, due to hardware difference, using the official code, we get results (Baseline) that is slightly different from that reported in their paper (RDFNet), as shown in Table I. We compare different variants of our approach to demonstrate the effectiveness of each component, and the results are reported in Table I.

**Analysis of Temporal Consistency.** While training RDFNet with the 795 manually labeled images, and our temporal consistency constraint (+$L_C$ for short), the mean accuracy increases to 63.3% from 63.0%, and IoU increases to 50.1% from 50.0%. The results show that training the network with self-supervised loss function could effectively imporve the performance of fully supervised baseline method. The main reason why the performance improvement is not siginificant is that the information supplement of adjacent frames is not enough.

TABLE I
COMPARISON OF DIFFERENT METHODS. THE THREE COLUMNS REPRESENT THREE DIFFERENT TYPES OF METHODS: SEMANTIC SEGMENTATION OF A SINGLE IMAGE, SEMANTIC SEGMENTATION OF RGBD IMAGES AND MULTI-TASK LEARNING METHODS.

| Method | Data | Pixel accuracy | Mean accuray | Mean IoU |
|---|---|---|---|---|
| B-SegNet [3] | RGB | 68.0 | 45.8 | 32.4 |
| Context [4] | RGB | 70.0 | 53.6 | 40.6 |
| RefineNet [5] | RGB | 73.6 | 58.9 | 46.5 |
| Gupta et al. [7] | RGBD | – | 35.1 | – |
| Eigen et al. [11] | RGBD | 65.6 | 45.1 | 34.1 |
| FCN [1] | RGBD | 65.4 | 46.1 | 34.0 |
| LSTM-CF [8] | RGBD | – | 49.4 | – |
| Cheng et al. [9] | RGBD | 71.9 | 60.7 | 45.9 |
| RDFNet [10] | RGBD | **76.0** | 62.8 | 50.1 |
| Baseline | RGBD | 75.8 | 63.0 | 50.0 |
| Ours (+$L_C$) | RGBD | 75.8 | 63.3 | 50.1 |
| Ours (+$Pr$) | RGBD | 75.5 | 63.5 | 49.8 |
| Ours(+$Hm$) | RGBD | | | |
| Ours (+$PGT$) | RGBD | 75.9 | 63.6 | 50.2 |
| Ours (+$L_C$+$PGT$) | RGBD | 75.8 | **63.8** | **50.2** |
| PAD-Net [12] | RGB | 75.2 | 62.3 | 50.2 |
| TRL [13] | RGB | 76.2 | 56.3 | 46.4 |
| Jiao et al. [14] | RGB | *81.1* | 62.2 | *50.9* |

**Analysis of Pseudo labels.** Though the pseudo labels still contain a certain amount of noises, most of them are reliable, especially inside an object region. While training the RDFNet with extra images of pseudo labels, the performance (Ours+PGT) greatly increases, especially at mean accuracy (63.6%), compared with the baseline. To facilitate future study, we will release our dataset with pseudo labels. In addition, we investigate the effect of quality of PGT on network performance. Simply limiting the range of annotation propagation, the performance of baseline method begins to decline, largely due to the poor quality of PGT. But on the other hand, with the increase of the number of samples, the performance (Ours+Pr) of abtained model for each category (Mean accuracy) is improved.

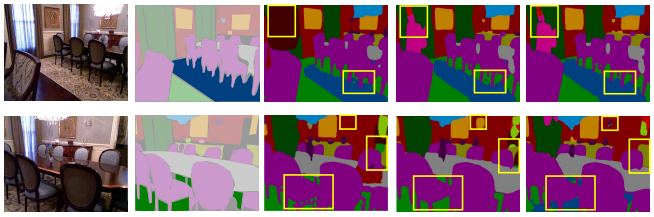
Combining both the pseudo ground truth and temporal consistency, our approach outperforms existing semantic segmentation networks, especially at mean accuracy and IoU metrics. We also compare our method with three multi-task learning techniques and our method achieves comparable performance. Though Jiao *et al.* achieve the best performance on pixel accuracy and IoU, they require additional dense ground-truth for depth estimation and much heavier computational cost. Fig. 4 shows a group of segmentation results. It can be seen that our method performs well on objects such as pictures, whiteboards and so on. And compared with baseline method, there is less over-segmentation in our apprrach, such as picture on the walls (Fig. 4, line 3) and cabinet in the cornor (Fig. 4, line 4). Class-wise mean IoU of our results compared with RDFNet [10] are shown in Table II. The experimental results in Table II show that our method effecitively improves segmentation accuracy on most categories (24/40) by using

TABLE II
CLASS-WISE IOU ON NYUD-V2

| | wall | floor | cabinet | bed | chair | sofa | table | door | window | bkshelf |
|---|---|---|---|---|---|---|---|---|---|---|
| RDF-152 | 79.7 | 87.0 | **60.9** | **73.4** | **64.6** | **65.4** | **50.7** | 39.9 | 49.6 | 44.9 |
| Ours | **80.8** | **87.9** | 60.1 | 72.1 | 64.1 | 64.6 | 49.2 | **41.6** | **49.8** | **45.3** |
| | counter | blind | desk | shelf | curtain | dresser | pillow | mirror | mat | cloths |
| RDF-152 | **67.1** | **63.9** | **28.6** | 14.2 | 59.7 | **49.0** | 49.9 | 54.3 | 39.4 | 26.9 |
| Ours | 66.6 | 61.9 | 25.1 | **14.4** | **59.9** | 46.2 | **50.1** | **55.9** | **41.5** | **27.4** |
| | books | refridge | tv | paper | tower | shower | box | board | person | stand |
| RDF-152 | **35.0** | 58.9 | **63.8** | **34.1** | 41.6 | **38.5** | 11.6 | 54.0 | 80.0 | **45.3** |
| Ours | 34.7 | **61.4** | 61.9 | 32.9 | **41.7** | 37.0 | **11.9** | **56.8** | **82.4** | 44.5 |
| | sink | lamp | bathtub | bag | othstr | othfurn | othprop | picture | ceiling | toilet |
| RDF-152 | 62.1 | 47.1 | 57.3 | 19.1 | 30.0 | 20.6 | **39.0** | 61.2 | **69.1** | 65.7 |
| Ours | **65.3** | **47.2** | **58.5** | **19.9** | **32.3** | **20.8** | 38.1 | **61.7** | 68.9 | **66.5** |

label propageting and self-supervised loss function. It's not only in categories with large planes such as wall, floor and door, but also in classes with small size such as bag and box. Although our method imporves the performance of small objects, the mean IoU of box and bag are still under 30%. The main reason is that the size of these two kinds of objects is too small and the occurrence frequency of them is very low ($< 25\%$).

**Dicussion.** By involving temporal consistency, we also found that there are conflicts between manual labels of the same scene, as shown in Fig. 5. Compared to other methods, Our method produces much more accurate and consistent results. And the feature extracted by our model is much more robust.



(a) (b) (c) (d) (e)

Fig. 5. Mislabeled images. (a) Two frames of the same scene. (b) Their corresponding ground truth labels. Note the difference between the labels of the floor pixels. Blue as "floor mat" and green as "floor". The mislabled pixels are highlighted. (c)(d)(e) The prediction results of RefineNet, RDFNet and our method, respectively.

## IV. CONCLUSION

In this paper, we introduce two effective semi-supervised approaches for exploiting unlabeled frames in videos to improve semantic segmentation performance of indoor scenes. First, we propagate the labeled data to neighboring frames and generate a large number of reliable pseudo labels to enrich the training set. Second, our training policy takes advantage of temporal correlations between adjacent frames to enhance the semantic segmentation performance. Experimental results on NYUD-v2 dataset demonstrate the superiority of our proposed method.

## REFERENCES


[1] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2015, pp. 3431–3440.

[2] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han, "Learning deconvolution network for semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 2015, pp. 1520–1528.

[3] Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla, "Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding," in *arXiv preprint arXiv:1511.02680*, 2015.

[4] Guosheng Lin, Chunhua Shen, Anton Van Den Hengel, and Ian Reid, "Efficient piecewise training of deep structured models for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2016, pp. 3194–3203.

[5] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017, pp. 5168–5177.

[6] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz, "Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2018, pp. 8934–8943.

[7] Saurabh Gupta, Ross Girshick, Pablo Arbeláez, and Jitendra Malik, "Learning rich features from rgb-d images for object detection and segmentation," in *European Conference on Computer Vision*. Springer, 2014, pp. 345–360.

[8] Zhen Li, Yukang Gan, Xiaodan Liang, Yizhou Yu, Hui Cheng, and Liang Lin, "Lstm-cf: Unifying context modeling and fusion with lstms for rgb-d scene labeling," in *European Conference on Computer Vision*. Springer, 2016, pp. 541–557.

[9] Yanhua Cheng, Rui Cai, Zhiwei Li, Xin Zhao, and Kaiqi Huang, "Locality -sensitive deconvolution networks with gated fusion for rgb-d indoor semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017, pp. 1475–1483.

[10] Seong-Jin Park, Ki-Sang Hong, and Seungyong Lee, "Rdfnet: Rgb-d multi-level residual feature fusion for indoor semantic segmentation," in *The IEEE International Conference on Computer Vision*, 2017, pp. 4990–4999.

[11] David Eigen and Rob Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 2015, pp. 2650–2658.

[12] D. Xu, W. Ouyang, X. Wang, and N. Sebe, "Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing," in *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 2018, pp. 675–684.

[13] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Zequn Jie, Xiang Li, and Jian Yang, "Joint task-recursive learning for semantic segmentation and depth estimation," in *European Conference on Computer Vision*. Springer, 2018, pp. 238–255.

[14] Jianbo Jiao, Ying Cao, Yibing Song, and Rynson Lau, "Look deeper into depth: Monocular depth estimation with semantic booster and attention-driven loss," in *European Conference on Computer Vision*. Springer, 2018, pp. 55–71.