Manuscript version: Author’s Accepted Manuscript
The version presented in WRAP is the author’s accepted manuscript and may differ from the published version or Version of Record.

Persistent WRAP URL:
http://wrap.warwick.ac.uk/135683

How to cite:
Please refer to published version for the most recent bibliographic citation information. If a published version is known of, the repository item page linked to above, will contain details on accessing it.

Copyright and reuse:
The Warwick Research Archive Portal (WRAP) makes this work by researchers of the University of Warwick available open access under the following conditions.

Copyright © and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable the material made available in WRAP has been checked for eligibility before being made available.

Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

Publisher’s statement:
Please refer to the repository item page, publisher’s statement section, for further information.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk.
Abstract—We propose a novel fast and robust 3D point clouds segmentation framework via coupled feature selection, named 3DCFS, that jointly performs semantic and instance segmentation. Inspired by the human scene perception process, we design a novel coupled feature selection module, named CFSM, that adaptively selects and fuses the reciprocal semantic and instance features from two tasks in a coupled manner. To further boost the performance of the instance segmentation task in our 3DCFS, we investigate a loss function that helps the model learn to balance the magnitudes of the output embedding dimensions during training, which makes calculating the Euclidean distance more reliable and enhances the generalizability of the model. Extensive experiments demonstrate that our 3DCFS outperforms state-of-the-art methods on benchmark datasets in terms of accuracy, speed and computational cost.

I. INTRODUCTION

3D scene understanding based on LiDAR, RGB-D and stereo cameras has received increasing attention from both academia and industry because of its critical role in robotic scene perception, robotic manipulation and autonomous driving [1, 2]. Instance and semantic segmentation are the most widely used tasks in this research field. Building on the great success achieved in recent years [3]–[6] for each single task, joint learning methods for both tasks [7], [8] have opened up a more effective way to improve performance and promote further developments.

The two tasks have some common ground that can be associatively utilized to boost their performance. For example, points with different classes must be from different instances, and points from the same instance must be of the same class. The simplest but most naive methods to jointly perform instance and semantic segmentation are progressively using the predicted semantic labels to further cluster instances or utilizing the predicted instance results as prior knowledge for semantic segmentation. Nevertheless, using the unreliable upstream prediction as prior information may affect the downstream task; consequently, such approaches may be suboptimal. Another approach is to directly combine the high-level features of both tasks to perform information integration [7]. Although semantic and instance segmentation share the same goal of detecting specific informative regions, they have different individual learning orientations, and some part of the information they contain may be contradictory. Instance segmentation focuses on extracting point features from different objects to distinguish them, while the features extracted by semantic segmentation are used to classify points with different categories; as a result, the features of both tasks definitely contain different task-oriented parts. Therefore, feature selection is an essential step in the reciprocal process.

Actually, such selection within a mutually aided process is consistent with human scene perception. Semantic and instance segmentation are the most important visual tasks in human scene perception. For humans, semantic perception mainly abstracts the advanced semantic features from the objects in the scene, while instance segmentation pays more attention to exploiting the primary features. These two processes can complement each other. Specifically, the mapping from advanced features to primary features can be beneficial for instance segmentation. For example, if we know the category of an object, we will obtain the blur shape information (primary features), which can help to correct errors in the instance segmentation result if it is difficult to see the

† The first two authors contributed equally to this work.
* This work was supported by the 111 Project (NO.B18015), the National Natural Science Foundation of China (No.91630314), the key project of Science & Technology for Brain-Inspired Intelligence (primary features), which can help to correct errors in the instance segmentation result if it is difficult to see the

Fig. 1. An illustration of CFSM and a qualitative comparison of our method and the baseline. The baseline is a traditional multitask framework without CFSM, as introduced in Section III. Our framework via coupled feature selection is able to exploit and integrate the reciprocal information from both tasks based on gate mechanism to boost their performance, as shown in the marked regions.
whole object because of environmental light. By contrast, if we know the primary features of objects of the same category that are unknown, we can establish links between those that belong to the same category, which can help us to quickly and accurately accomplish the semantic segmentation. Consequently, these two processes are not independent but coupled. However, humans are rarely disturbed by such different task-oriented information because the human brain has the capability to quickly and adaptively select useful information instead of the entire set of information. This characteristics of human scene perception inspired us to build a multitask-coupled framework to simulate the information-selection process of human scene perception via gate control units for robotics. The coupled and gate-based training pipeline is shown in Figure 1. For the encoder, we use the PointNet/PointNet++ utilized by [7], [9]. For the decoder, we investigate a novel coupled feature selection module (CFSM) that contains two coupled instance-to-semantic and semantic-to-instance streams to extract useful information while filtering useless information.

Our 3DCFS uses the Euclidean distance to calculate the similarity of different embeddings among all points for clustering. However, the Euclidean distance is sensitive to the magnitudes of different embedding dimensions, which makes the clustering result depend on only a small number of dimensions and reduces the generalizability of the model. We therefore propose a novel loss function, named $\mathcal{E}_{EMED}$, that helps the model learn to maintain equilibrium among the magnitudes of the instance embedding dimensions. In summary, the main contributions of our work are as follows:

- We propose a fast, yet effective end-to-end point clouds segmentation framework that simultaneously performs semantic and instance segmentation inspired by the human scene perception process.
- We introduce a novel coupled feature selection module (CFSM) to exploit the potential reciprocal information in semantic and instance segmentation tasks to seamlessly fuse the heterogeneous features, allowing these two tasks to benefit from each other.
- We design a novel loss for instance segmentation in 3DCFS, which helps the model learn to balance the magnitudes of the embedding dimensions to maintain the stability of the Euclidean distance calculation during training.
- We achieve state-of-the-art performance for 3D semantic and instance segmentation on benchmark datasets in terms of accuracy, speed and computational cost.

II. RELATED WORK

2D Semantic and Instance Segmentation. The great advances in semantic and instance segmentation have largely been driven by the success of fully convolutional neural networks (FCNs) [5]. Numerous approaches [10–16] based on FCNs have dominated semantic segmentation tasks. [4], [17], [18] learned to segment instances by proposing segmentation candidates based on the region-based CNN (R-CNN) [19]. A top-down detector-based Mask R-CNN framework was first introduced by He et al. [3] to simultaneously perform mask and class label prediction. By contrast, bottom-up methods such as [20], [21] aim to assign per-pixel predictions to instances.

3D Point Clouds Segmentation. Recent advances in deep neural networks have also led to various cutting-edge 3D semantic [22–28] and instance segmentation [9], [29–31] approaches. Using voxelized volumes to represent 3D point clouds is a popular and effective strategy. [32–35] transferred 3D point clouds data into regular volumetric occupancy grids and applied 3D CNNs to perform voxel-level predictions. Based on the MLP, PointNet [36] was the first to directly process raw point clouds and perform point-level predictions, demonstrating high performance on both segmentation and classification tasks. Following that pioneering work, PointNet++ [37], PointCNN [38], GB-RCU [39] and RSNet [40] were developed through investigations of the local context and hierarchical learning structures. Graph neural networks have opened up more efficient and flexible ways to handle 3D segmentation [6], [41], [42]. Recently, by advancing a joint semantic and instance learning framework, [7], [8] proposed methods that achieve superior performance on both tasks.

III. METHOD

As depicted in Figure 2, the framework with CFSM and $\mathcal{E}_{EMED}$ removed is the baseline method. First, point clouds of size $L_P$ are encoded into a feature matrix $F_{SHARE} \in \mathbb{R}^{L_P \times L_P}$ by the encoder (PointNet/PointNet++). Next, two tasks separately decode the shared encoded feature for their own missions. $F_{SHARE}$ is decoded by the semantic segmentation branch into the semantic feature matrix $F_{SEM} \in \mathbb{R}^{L_P \times L_P}$ and then outputs the semantic predictions $F_{SEM} \in \mathbb{R}^{L_P \times L_C}$, where $L_C$ is the semantic class number. The instance segmentation branch decodes $F_{SHARE}$ into the instance feature matrix $F_{INS} \in \mathbb{R}^{L_P \times L_P}$, which is utilized to predict the per-point instance embeddings $E_{INS} \in \mathbb{R}^{L_P \times L_E}$, where $L_E$ denotes the length of the output embedding dimensions. These embeddings are used to calculate the Euclidean distances between points for instance clustering. During the training process, the semantic branch is supervised by cross-entropy loss and the instance branch is supervised by the instance loss following [7], and the specific loss formula is detailed in [7]. In our paper, we denote this loss as $E_{INS}$. For inference, we use mean-shift clustering [43] on the instance embeddings to obtain the final instance labels. The mode of the semantic labels of the points within the same instance is assigned as the predicted semantic class.

A. CFSM

Reciprocal Feature Selection and Integration. As illustrated in Figure 2, our CFSM contains two branches: $C_{I \rightarrow S}$ for instance-fused semantic segmentation and $C_{S \rightarrow I}$ for semantic-aware instance segmentation. Both $C_{I \rightarrow S}$ and $C_{S \rightarrow I}$ can be separately integrated into baseline model, when
the other branch is replaced by the MLP. In our method, there are two types of gates: attention gates (A-gates) and selection gates (S-gates). Both are learnable modules that implement several $1 \times 1$ convolutions and activation functions. An A-gate is used for reweighting the semantic and instance features themselves before fusion, while the S-gate is used to filter or select the information from the other task. The cell is a decoding unit that contains convolutions and activation functions.

CI-S. As depicted in Figure 2, we denote $F_{INS}$ and $F_{SEM}$ as the instance and semantic features decoded by the MLP from $F_{SHARE}$. The semantic branch $C_{I-S}$ contains three units: an A-gate called “A-gate-S”, an S-gate named “S-gate-S” and a cell termed “Cell-S”. The units with the same name share the weight parameters. For the $C_{I-S}$ branch, the red $F_{INS}$ pass through the Cell-S and A-gate-S to obtain the output $O_I$; the $F_{SEM}$ in green are fed into all three units to obtain the output $O_{IS}$. These outputs are calculated by the dot product $\otimes$ and summation $\oplus$ operations as illustrated in Figure 2 which are formulated as follows:

\[
\begin{align*}
O_I &= \sigma(W_{cell} \ast F_{INS}) \cdot \varsigma(W_A \ast F_{INS}) \quad (1) \\
O_i &= \xi(W_S \ast F_{SEM} \cdot O_I) \quad (2) \\
O_{IS} &= \sigma(W_{cell} \ast F_{SEM}) \cdot \varsigma(W_A \ast F_{SEM}) + O_i \quad (3) \\
F_{I-S} &= O_{IS} \quad (4)
\end{align*}
\]

where $W_{cell}$, $W_A$ and $W_S$ are the weights of Cell-S, A-gate-S and S-gate-S, respectively, and $\sigma$, $\varsigma$ and $\xi$ indicate their activation functions. The symbol $\ast$ denotes the convolution operation, and $\cdot$ represents the dot product. The final output of $C_{I-S}$ for the semantic segmentation task is $F_{I-S} \in \mathbb{R}^{L_P \times L_F}$. Based on the attention mechanism, our A-gate has the capability to reweight the features of the task itself to facilitate extracting the crucial internal information of both tasks. Then, the representations in $F_{INS}$, which are useful to the semantic segmentation task, are exploited and reserved through our S-gate. This selection process is guided and controlled by the semantic features. For example, the S-gate-S in $C_{I-S}$ can select features from the same instances and discover their general characteristics, which helps the model to recognize their category.

CS-I. The $C_{S-I}$ architecture is the same as the $C_{I-S}$ structure, except that the instance features are assisted by the semantic features. $F_{SEM}$ is passed to $C_{S-I}$ as complementary information to help improve the performance of the instance segmentation. S-gate-I in $C_{S-I}$ is able to block the useless information and filter the features that blur the differences between instances as well as select more valuable representations to indicate the differences between categories for instance segmentation. The output of $C_{S-I}$ for the semantic segmentation task is $F_{S-I} \in \mathbb{R}^{L_P \times L_F}$.

CFSM. As illustrated in Figure 2, $C_{I-S}$ and $C_{S-I}$ can be simultaneously integrated into the baseline model. As the complementary information, $F_{SEM}$ and $F_{INS}$ are taken as the input to the other branches $C_{S-I}$ and $C_{I-S}$, respectively. At the same time, $F_{SEM}$ and $F_{INS}$ are passed to their own branches $C_{I-S}$ and $C_{S-I}$, respectively.

B. Learn to Balance the Embedding Dimension Magnitude

To further improve the instance segmentation performance of our framework, we design a loss function to learn to balance the magnitudes of the output embedding dimensions. The Euclidean distance is sensitive to magnitude differences, which makes the cluster results dependent on only a few dimensions of the embedding and reduces the generalizability of the model. A traditional trick to solve this issue is to apply a mean-removal strategy on the output embeddings before instance clustering during inference. Rather than employ this post-process method, we directly apply our proposed loss function to the model to stabilize the Euclidean distance calculation during training. Specifically, we denote $\mathcal{E}_{EMED}$ as the equilibrium loss for the magnitude. The loss term can be written as follows:

\[
\begin{align*}
\mathcal{E} &= \frac{1}{L_P} \sum_{i=1}^{L_P} E_i \\
\mathcal{E}_{EMED} &= \frac{1}{L_E} \sum_{d=1}^{L_E} (\mathcal{E}_d - \mu)^2 \\
\mathcal{E}_{INS} &= \mathcal{E}_{INS} + \alpha \mathcal{E}_{EMED}
\end{align*}
\]
where $E_i$ is the embedding of each point, $\mathcal{E}_{\text{INS}}$ is the total instance loss of our 3DCFS, $\mu$ denotes the mean value of $\bar{E}$, and $\alpha$ is the balanced weight of $\mathcal{E}_{\text{INS}}$ and $\mathcal{E}_{\text{EMED}}$.

IV. EXPERIMENTS

A. Datasets and Experimental Setup

Dataset. Our experiments are conducted on two benchmark datasets: the Stanford 3D Indoor Semantics Dataset (S3DIS) [44] and ShapeNet Dataset [45].

- S3DIS is a 3D scene dataset that contains large-scale scans of indoor spaces. Each point is annotated with an instance label and a semantic label from 13 semantic classes. S3DIS embeds each point into a 9-dimensional feature vector including XYZ, RGB and normalized coordinates. Following [36], we split the rooms into 1 m $\times$ 1 m overlapping blocks with stride 0.5 m on the ground plane and sample 4,096 points from each block.

- The ShapeNet part dataset contains 16,881 3D shapes from 16 semantic classes. Each point is associated with one of the 50 different parts. We utilize the instance annotations from [9] as the ground-truth labels. Each shape is represented by point clouds with 2,048 points following [36], and each point is represented by an XYZ 3-dimensional vector. The point clouds are sampled for the input of our framework following [7].

Evaluation. Following [7], we conduct experiments involving S3DIS on Area5. The performance on the sixth fold cross validation with microaveraging [4] is also measured. For semantic segmentation, we calculate the overall accuracy (oAcc), mean accuracy (mAcc) and mean IoU (mIoU). To evaluate the performance of instance segmentation, we use the coverage (Cov) and weighted coverage (WCov), the specific calculation formulas are detailed in [21], [46], [47].

Implementation Details. For instance segmentation, we train 3DCFS with $\lambda = 0.001$. We use five output embeddings following [7] and set $\alpha$ to 0.01. We train the network for 50 epochs for PointNet and PointNet++ with a batch size of 12 and the base learning rate set to 0.001 and divided by 2 every 300 k iterations. We select the Adam optimizer to optimize the network on a single GPU (Tesla P100) and set the momentum to 0.9 for the training process. During the inference process, we set the bandwidth to 0.6 for mean-shift clustering and apply the BlockMerging algorithm [9] to merge instances from different blocks. The code will be available at GitHub, which contains more details.

B. S3DIS

Following [7], we conducted experiments on the S3DIS dataset based on the PointNet and PointNet++ backbone networks.

Quantitative Results. For Area5, the quantitative results of 3DCFS on the instance and semantic segmentation tasks are shown in Table I. Using the PointNet backbone, 3DCFS achieves 44.4 mWCov, which dramatically outperforms the state-of-the-art method ASIS [7] by 2.8 and significantly improves the mPrec by 3.3. After replacing the backbone with PointNet++, we still achieve 4.5 mWCov gains and 2.1 mIoU gains on the instance and semantic segmentation tasks, respectively. Table II shows the performance of our 3DCFS on semantic segmentation in all 6 areas. 3DCFS outperforms ASIS by 1.4 for mWCov and 1.8 for mPrec. Using the PointNet++ backbone on all 6 areas, 3DCFS improves the mAcc by 2.3 and the mIoU by 1.0. Clearly, our 3DCFS method outperforms the SOTA method ASIS [7] by a large margin. As reported in Table I and II, whether constructed upon the PointNet or PointNet++ backbones, evaluated in Area5 or 6-fold CV, our method consistently obtains better performance on both instance and semantic segmentation tasks than the state-of-the-art methods. The stable improvement demonstrates that our 3DCFS is a general and effective framework that can be built upon different network backbones. Table III shows the instance and semantic segmentation results for specific categories. We reproduced the results of ASIS [8] and JSIS3D [7] using the code at GitHub published by the respective authors to make a full class comparison with the same PointNet backbone.

Ablation Study. The ablation study results are shown in Table III. Compared with the baseline, our 3DCFS achieves obvious improvements. We obtain 3.1 mWCov and 4.2 mPrec gains for the instance segmentation task. For semantic
Real Scene Baseline 3DCFS Ground Truth

Fig. 3. Comparison of the baseline method and 3DCFS on instance segmentation. The different colors represent different instances.

Real Scene Baseline 3DCFS Ground Truth

Fig. 4. Comparison of the baseline method and 3DCFS on semantic segmentation.

For instance segmentation, different colors represent different instances. As depicted in Figure 3, the baseline approach incorrectly clusters two nearby different class instances into one instance (e.g., board and wall). After applying 3DCFS, the instances are correctly clustered. For semantic segmentation, each color refers to a particular class. The qualitative comparisons are shown in Figure 4. 3DCFS performs better on classifying the gains provided by the individual CS-I and CI-S. The results demonstrate that improving one task can also help improve the other because each task learns better reciprocal features.

Table III also compares our \( E_{EMED} \) with the post-processing method mentioned in Section III-B. Figure 6 shows the comparison of the mean and variance of 5 dimensions of the embeddings with or without \( E_{EMED} \). The statistical analysis of the results reveals that the mean values are balanced and the variances are not influenced by \( E_{EMED} \), which indicates that it maintains the representational ability of the instance feature. The superior performance shows that our \( E_{EMED} \) successfully helps the model learn to balance the dimension magnitude. Table V shows that incorporating our proposed \( E_{EMED} \) boosts the performance with embedding lengths 5 and 10. Note that the improvement is much more significant as the embedding length increases.

Qualitative Results. For instance segmentation, different colors represent different instances. As depicted in Figure 3, the baseline approach incorrectly clusters two nearby different class instances into one instance (e.g., board and wall). After applying 3DCFS, the instances are correctly clustered. For semantic segmentation, each color refers to a particular class. The qualitative comparisons are shown in Figure 4. 3DCFS performs better on classifying the
TABLE IV

PER CLASS RESULTS ON THE S3DIS DATASET.

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Method</th>
<th>train</th>
<th>ceiling</th>
<th>floor</th>
<th>wall</th>
<th>beam</th>
<th>column</th>
<th>window</th>
<th>door</th>
<th>table</th>
<th>chair</th>
<th>sofa</th>
<th>bookcase</th>
<th>board</th>
<th>clutter</th>
</tr>
</thead>
<tbody>
<tr>
<td>mPrec</td>
<td>JSIS3D</td>
<td>41.5</td>
<td>82.7</td>
<td>85.1</td>
<td>44.2</td>
<td>0.0</td>
<td>15.4</td>
<td>57.1</td>
<td>35.7</td>
<td>68.3</td>
<td>39.1</td>
<td>43.2</td>
<td>58.4</td>
<td>73.4</td>
<td>34.0</td>
</tr>
<tr>
<td></td>
<td>ASIS</td>
<td>53.7</td>
<td>88.6</td>
<td>87.5</td>
<td>57.8</td>
<td>5.1</td>
<td>41.1</td>
<td>63.3</td>
<td>41.1</td>
<td>68.3</td>
<td>59.1</td>
<td>43.2</td>
<td>58.4</td>
<td>73.4</td>
<td>34.0</td>
</tr>
<tr>
<td></td>
<td>3DCFS</td>
<td>55.3</td>
<td>89.1</td>
<td>91.9</td>
<td>59.3</td>
<td>6.3</td>
<td>41.1</td>
<td>63.3</td>
<td>41.1</td>
<td>68.3</td>
<td>59.1</td>
<td>43.2</td>
<td>58.4</td>
<td>73.4</td>
<td>34.0</td>
</tr>
</tbody>
</table>

| mAcc    | JSIS3D  | 64.3  | 20.0    | 47.5  | 9.1  | 30.3  | 59.1   | 43.2   | 58.4 | 7.3   | 36.0  | 64.8 | 34.4      |       |         |
|         | ASIS    | 64.3  | 20.0    | 47.5  | 9.1  | 30.3  | 59.1   | 43.2   | 58.4 | 7.3   | 36.0  | 64.8 | 34.4      |       |         |
|         | 3DCFS   | 64.3  | 20.0    | 47.5  | 9.1  | 30.3  | 59.1   | 43.2   | 58.4 | 7.3   | 36.0  | 64.8 | 34.4      |       |         |

TABLE V

ABLATION STUDY ON THE S3DIS DATASET IN AREA5. CFMS5 AND CFMS10 REPRESENT THE INSTANCE SEGMENTATION WITH OUTPUT EMBEDDING LENGTH OF 5 AND 10, RESPECTIVELY. 3DCFS5 AND 3DCFS10 DENOTE THE METHOD EQUIPPED WITH E:\EMEM\.

<table>
<thead>
<tr>
<th>Method</th>
<th>mCov</th>
<th>mWconv</th>
<th>mRec</th>
<th>mAcc</th>
<th>mIoU</th>
<th>oAcc</th>
</tr>
</thead>
<tbody>
<tr>
<td>CFMS5</td>
<td>48.0</td>
<td>30.8</td>
<td>54.7</td>
<td>44.6</td>
<td>62.3</td>
<td>54.5</td>
</tr>
<tr>
<td>3DCFS5</td>
<td>49.0</td>
<td>52.1</td>
<td>55.5</td>
<td>45.9</td>
<td>62.7</td>
<td>55.5</td>
</tr>
<tr>
<td>CFMS10</td>
<td>45.4</td>
<td>48.3</td>
<td>52.2</td>
<td>40.5</td>
<td>61.2</td>
<td>53.6</td>
</tr>
<tr>
<td>3DCFS10</td>
<td>48.4</td>
<td>51.7</td>
<td>55.0</td>
<td>44.1</td>
<td>63.1</td>
<td>55.2</td>
</tr>
</tbody>
</table>

TABLE VI

COMPARISONS OF COMPUTATION SPEED, GPU MEMORY AND PERFORMANCE. THE UNITS FOR TIME AND MEMORY ARE MINUTES AND MB RESPECTIVELY.

<table>
<thead>
<tr>
<th>Method</th>
<th>Metrics</th>
<th>Train</th>
<th>Test</th>
<th>mWconv</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>time</td>
<td>memory</td>
<td></td>
</tr>
<tr>
<td>SGPN [9]</td>
<td>59.3</td>
<td>7549</td>
<td>209.5</td>
<td>420</td>
</tr>
<tr>
<td>ASIS [7]</td>
<td>64.7</td>
<td>4275</td>
<td>54.2</td>
<td>1235</td>
</tr>
<tr>
<td>3DCFS</td>
<td>26.4</td>
<td>2227</td>
<td>36.3</td>
<td>307</td>
</tr>
</tbody>
</table>

TABLE VII

SEMANTIC SEGMENTATION RESULTS ON SHAPENET DATASETS.

<table>
<thead>
<tr>
<th>Method</th>
<th>mIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASIS [7]</td>
<td>85.0</td>
</tr>
<tr>
<td>SGPN [9]</td>
<td>85.8</td>
</tr>
<tr>
<td>3DCFS</td>
<td>87.6</td>
</tr>
</tbody>
</table>

C. ShapeNet

We conducted experiments on the ShapeNet dataset using instance segmentation annotations generated by [9]. For instance segmentation, only the qualitative results are provided following [9] because no true ground truth exists. As shown in Figure 7, the tires of the car and legs of the chair and the table are properly grouped into individual instances. Both the semantic and instance segmentation results are accurate and clear. The semantic segmentation results are shown in Table VII. Our 3DCFS further outperforms the state-of-the-art method SGPN by 1.8 mIoU based on PointNet++. These results reveal that our proposed 3DCFS also has the capability to boost part segmentation performance.

V. CONCLUSIONS

In this paper, we proposed a fast and robust joint 3D semantic-instance segmentation framework. A novel CFSM was introduced to exploit the reciprocal information from two different tasks in a coupled manner. We also proposed a novel loss function that helped our 3DCFS learn to balance the magnitudes of the instance embedding dimensions to make the Euclidean distance calculation more reliable. Experimental results on the S3DIS and ShapeNet part datasets demonstrated the effectiveness and efficiency of 3DCFS.

entire semantic information, especially at the boundaries of different categories. Figure 5 shows qualitative examples of 3DCFS on both instance and semantic segmentation. Our results are essentially the same as the ground truth, especially for instance segmentation.

Speed and Computing Resources. Table VI shows a comparison of the memory cost and computation time measured on a single GTX 1080 GPU. For a fair comparison, we conducted the experiments in the same environment, including the same GPU, batch size (4) and data (Area5). Note that all time units are minutes and all memory units are MB. In the training process, the result is the time and memory cost for one epoch. Our approach takes only 26.4 minutes and 2,227 MB, which is significantly faster and more efficient than the state-of-the-art methods. In the test process, the results show the resource consumption for inferencing. Here, our method is also found to be superior to the state-of-the-arts in terms of accuracy, speed and computational cost.

TABLE VII

TABLE VI

Comparison of Computation Speed, GPU Memory and Performance. The units for time and memory are minutes and MB respectively.

<table>
<thead>
<tr>
<th>Method</th>
<th>Metrics</th>
<th>Train</th>
<th>Test</th>
<th>mWconv</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>time</td>
<td>memory</td>
<td></td>
</tr>
<tr>
<td>SGPN [9]</td>
<td>59.3</td>
<td>7549</td>
<td>209.5</td>
<td>420</td>
</tr>
<tr>
<td>ASIS [7]</td>
<td>64.7</td>
<td>4275</td>
<td>54.2</td>
<td>1235</td>
</tr>
<tr>
<td>3DCFS</td>
<td>26.4</td>
<td>2227</td>
<td>36.3</td>
<td>307</td>
</tr>
</tbody>
</table>


