

PartGlot: Learning Shape Part Segmentation from Language Reference Games — Supplementary Material

Juil Koo¹ Ian Huang² Panos Achlioptas^{2,3} Leonidas Guibas² Minhyuk Sung¹
¹KAIST ²Stanford University ³Snap Inc.

In this supplementary material, we first further analyze the effect of using super-segments instead of points while varying the granularity of the super-segments (Section S.1). Then, we demonstrate how much the attention module in our network affects the target shape discrimination in the reference games (Section S.2). We also show more results of the out-of-distribution test with Airplanes and Cars (quantitatively in Section S.3 and qualitatively in Section S.10), and also analyze how much training data is needed to obtain meaning part segmentation results (Section S.4). The cross-part mIoUs are also reported in Section S.5. We also provide results when an additional regularization loss (group consistency loss introduced in AdaCoSeg [9]) is used (Section S.6) and also a more recent text encoder (ALBERT [4]) is used instead of LSTM, while these variations do not change the results much. We also experiment with finer-grained parts in PartNet [5] and synthetic referential language and report the results in Section S.8. At the end, we provide implementation details (Section S.9), and more qualitative results (Section S.10) and comparisons (Section S.11).

S.1. Effect of Granularity of Super-Segments

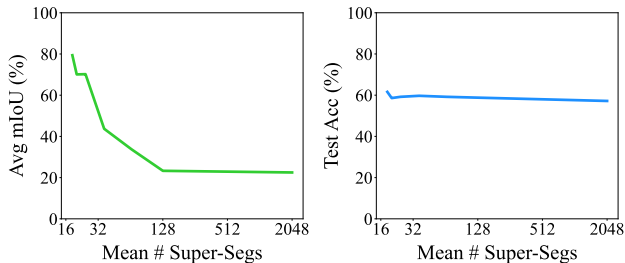


Figure S1. Results with different granularities of the super-segments. Left shows the average mIoUs, and right shows the target shape classification accuracy. The X-axis is the average number of super-segments in each shape in *log scale*; the higher, the smaller the super-segments are.

Our results in Section 4.2 in the main paper shows that using a set of super-segments as input instead of a point cloud is one of the crucial parts of our framework to achieve meaningful segmentation results through attention, while

the accuracy of the target shape classification is not affected much by the representation of shapes. We further analyze the effect of super-segments while varying their granularity. Although BSP-Net has parameters about the number of planes and the *maximum* number of convexes in training, increasing these numbers does not lead to producing more final convexes in practice. Thus, to achieve finer granularities, we use K-means clustering implemented in scikit-learn [6] to split each given super-segment into smaller pieces. For each super-segment s_i , we use K-means clustering for the points included in the super-segment (let \mathcal{P}_i denote the points) and set the number of subgroups K in the clustering to be $\lceil |\mathcal{P}_i|/N \rceil$, where N is our granularity parameter and $\lceil \cdot \rceil$ is the rounding function. When N is set to be the number of points in the entire point cloud (2048 in our experiments), it is the extreme case that the set of super-segments becomes the input point cloud itself. We test our network (with the PN-Aware setup) while varying the N from 16 to 256 (and 2048, which is the extreme case). Figure S1 illustrates the changes of the average mIoU in the part segmentation (left) and the accuracy of the target shape classification (right). The X-axis of the plots shows the average number of super-segments in each shape in *log scale*, and the Y-axis shows either the average mIoU or the accuracy. Interestingly, the granularity of the super-segments does not make any meaningful difference in the accuracy of the target shape classification but greatly affects the part segmentation mIoUs; more super-segments (smaller super-segments) results in a worse segmentation. This concludes that *pre-merging* the points as much as possible with geometric properties is the key to obtaining meaningful attention maps aligned with semantic parts.

S.2. Effect of Learning Attention

Table S1. Comparison with the cases of using uniform and random attention maps.

| Method | Classification Accuracy (%) |
|-------------------------------------|-----------------------------|
| Random $\{\mathbf{w}_i\}$ | 58.4 |
| Uniform $\{\mathbf{w}_i\}$ | 59.3 |
| Ours (Learning $\{\mathbf{w}_i\}$) | 61.5 |

To demonstrate whether our neural network learns the attention in a way to improve the discrimination of the target shape, we compare our attention module with two cases: using *uniform* attention maps and using *random* attention maps. As shown in Table S4, uniform attention maps provide better accuracy in the target shape classification compared with random attention maps, although its accuracy is still lower than the accuracy of our network learning the attention.

S.3. Out-of-Distribution Test — More Categories

Table S2. Quantitative results of the out-of-distribution test with Airplanes and Cars. The highest mIoU for each part of the target class is marked in **bold**.

| Other Classes | | Chair (w/ PN-Aware) | | | |
|---------------|--------|---------------------|-------------|-------------|------------|
| | | Back | Seat | Leg | Arm |
| Airplane | Body | 17.5 | 26.1 | 30.2 | 0.2 |
| | Wing | 3.1 | 47.5 | 3.5 | 6.3 |
| | Tail | 46.5 | 0.8 | 1.0 | 0.2 |
| | Engine | 5.4 | 11.6 | 6.1 | 7.5 |
| Car | Roof | 6.2 | 2.8 | 0.6 | 7.5 |
| | Hood | 0.1 | 17.5 | 1.5 | 0.6 |
| | Wheel | 9.9 | 12.5 | 21.3 | 1.6 |
| | Body | 45.3 | 29.5 | 2.4 | 10.5 |

In addition to the out-of-distribution test results in Section 4.3 in the main paper we provide more results testing our network trained with Chair shapes and utterances to Airplanes and Cars. The mIoUs across the parts are reported in Table S2, and qualitative results are in Section S.10. The model trained with the PN-Aware setup is used. Despite the big difference in the shapes, our model still recognizes some semantic parts such as Airplane *body*, *tail*, and *wing* and Car *body* and *wheel*.

S.4. Effect of Training Data Size

Table S3. Results when training with a subset of the training data. **Bold** indicates the best result for each column.

| Utterance Rate | Segmentation mIoU(%) | | | | | Classif. Acc.(%) |
|----------------|----------------------|-------------|-------------|-------------|-------------|------------------|
| | Back | Seat | Leg | Arm | Avg. | |
| 100% | 84.9 | 83.6 | 78.9 | 70.4 | 79.4 | 61.5 |
| 50% | 80.9 | 79.0 | 77.1 | 70.9 | 77.0 | 56.0 |
| 25% | 56.5 | 37.5 | 76.1 | 66.1 | 59.1 | 53.8 |

Table S3 illustrates results when only 50% and 25% of utterances are used in training (in the PN-Aware case). The part segmentation mIoUs are not changed much even when only 50% of the utterances are used, even when the target shape classification accuracy is decreased. In an extreme case using only 25% of the utterances in training, the mIoUs are decreased. This shows that our network does not necessitate a huge scale of data to obtain meaningful results.

S.5. Cross-Part mIoUs

Table S4. Part segmentation mIoUs across parts. The highest mIoU for each ground truth part is marked in **bold**.

| Ground Truth | Prediction | | | |
|------------------------|-------------|-------------|-------------|-------------|
| | Back | Seat | Leg | Arm |
| PN-Agnostic (Sec. 3.2) | | | | |
| Back | 82.2 | 4.2 | 1.6 | 3.5 |
| Seat | 0.8 | 78.8 | 1.5 | 5.2 |
| Leg | 0.5 | 4.2 | 75.5 | 3.1 |
| Arm | 0.2 | 0.7 | 0.8 | 40.6 |
| PN-Aware (Sec. 3.3) | | | | |
| Back | 84.9 | 2.5 | 1.5 | 1.7 |
| Seat | 1.8 | 83.6 | 2.6 | 1.4 |
| Leg | 1.1 | 2.4 | 78.9 | 0.7 |
| Arm | 0.4 | 0.6 | 1.3 | 70.4 |

We report the cross-part mIoUs for both PN-Agnostic and PN-Aware cases in Table S4. The diagonals are the same numbers reported in row 2 and 3 in Table 2 of the main paper. For both cases, the mIoUs of the corresponding parts are overwhelmingly higher than the ones of the other parts, indicating that there is almost no overlap across the attention maps of the parts.

S.6. Low Rank Regularization in AdaCoSeg [9]

Table S5. Results with the group consistency loss introduced by AdaCoSeg [9]. **Bold** indicates the best result for each column.

| Regularization | Segmentation mIoU(%) | | | | | Classif. Acc.(%) |
|--|----------------------|-------------|-------------|-------------|-------------|------------------|
| | Back | Seat | Leg | Arm | Avg. | |
| \mathcal{L}_{CE} (Ours) | 84.9 | 83.6 | 78.9 | 70.4 | 79.4 | 61.5 |
| $\mathcal{L}_{CE} + \mathcal{L}_{Coseg}$ | 83.4 | 82.2 | 79.2 | 72.0 | 79.2 | 60.2 |
| \mathcal{L}_{Coseg} | 79.2 | 80.1 | 78.1 | 72.5 | 77.5 | 60.6 |

AdaCoSeg [9] discussed in Section 2 in the main paper introduces a novel rank-based loss function improving the performance of the co-segmentation task. The loss function called *group consistency loss* (see Section 4.2 in the AdaCoSeg paper) maximizes the similarity of descriptors of the entities (super-segments in our case) included in the same group while differentiating the descriptors of entities assigned to different groups. The loss is computed with entities of *multiple* shapes in a minibatch, and thus it can enforce the consistency of the segmentation across multiple shapes.

We try to adapt this loss function to our network training. While this loss does not require labels, it still needs to *define* groups. Hence, to adapt the loss to our network training, we use the PN-Aware setup and consider the sets of super-segments belonging to each predefined part as the groups. For each super-segment s_i , we take the output of Per-Super Segment Encoder $g(s_i)$, the output of the last layer fed to predict the key $g_k(s_i)$ and value $g_v(s_i)$ vectors. We collate

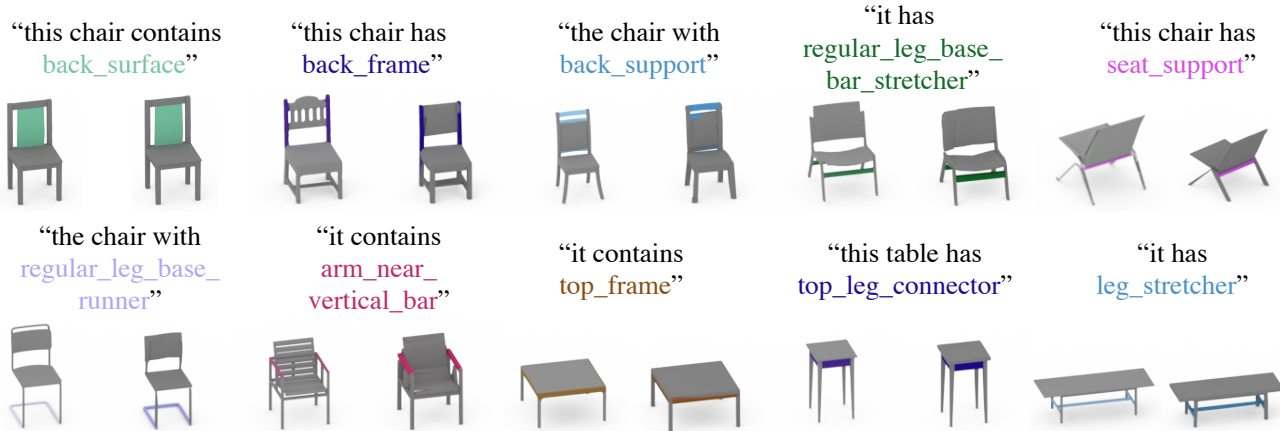


Figure S2. **PartNet Results.** For each pair, top is the utterance, left is the ground truth, and right is the predicted segment. The segmented parts of the given utterances are highlighted in color.

these descriptors for the super-segments assigned to each part (based on the attention outputs); let M_k denote the set of the descriptors (a matrix) for the k -th part. The group consistency loss of AdaCoSeg is then defined as follows:

$$\mathcal{L}_{\text{Coseg}} = 1 + \max_k \text{rank}(M_k) - \min_{k \neq l} \text{rank}([M_k, M_l]), \quad (1)$$

where rank indicates the second singular value of the matrix and $[\cdot]$ denotes the concatenation of two matrices.

Table S5 shows the results when using the group consistency loss in our network training. We find that the group consistency loss does not help improve the segmentation accuracy in our case. This result implies that the attention module in our network already learns consistent attention maps for each part.

S.7. Different Utterance Encoder — ALBERT [4]

Table S6. Results with ALBERT [4] as utterance encoders. **Bold** indicates the best result for each column.

| Utterance Encoder | Segmentation mIoU(%) | | | | | Classif. Acc.(%) |
|-------------------|----------------------|-------------|-------------|-------------|-------------|------------------|
| | Back | Seat | Leg | Arm | Avg. | |
| ALBERT (w/ FT) | 83.1 | 81.5 | 79.7 | 61.1 | 76.4 | 62.9 |
| ALBERT (w/o FT) | 80.9 | 80.8 | 78.7 | 72.6 | 78.2 | 57.8 |
| LSTM (Ours) | 84.9 | 83.6 | 78.9 | 70.4 | 79.4 | 61.5 |

For the utterance encoding, we also try the other Transformer-based encoder: ALBERT [4], which is a lite version of BERT [2]. We experimented with ALBERT in two ways: using a pretrained model and finetuning it, and without a pretrained model and training from scratch; the pretrained model is obtained from training BookCorpus [10] dataset with a masked language model objective. In the PN-Aware setup, the output of the *classification* encoder

$f_c(\mathbf{u})$ is obtained from the last hidden state of the [CLS] token, which is further processed through an MLP. The output of the *attention* encoder $f_a(\mathbf{l}_k)$ with a part name \mathbf{l}_k is obtained from the word embedding layer of ALBERT and also processed through an MLP. The results are compared in Table S6. Interestingly, the finetuned ALBERT does not improve the overall part segmentation mIoUs compared with our case of using a much simpler encoder, LSTM, while the target shape classification accuracy is higher. When training ALBERT from scratch, the classification accuracy decreases compared to using LSTM.

S.8. More Fine-grained Parts Segmentation Test

Table S7. The average mIoUs with *level 2* parts of Chair and Table in PartNet [5]. Due to the lack of utterances in the CiC dataset including words for the finer-grained parts, synthetic utterances are used. The results in the third column are the case shown in PartNet [5] when the network is *fully supervised* with the ground truth segments. Compared with that, our network learning only from referential language shows comparable results. For each category, the highest mIoU is marked in **bold**.

| Category | PN-Agnostic | PN-Aware | Supervised [5] |
|----------|-------------|----------|----------------|
| Chair | 35.3 | 32.7 | 38.2 |
| Table | 44.0 | 33.7 | 34.3 |

In our experiments so far, we used the four part classes (*back*, *seat*, *leg*, *arm*) of Chair in ShapeNet to match the majority of part names used in CiC utterances. To experiment with more parts, we would require a new reference game dataset that contains new part names in descriptions.

Here, we demonstrate experimental results with *synthetic* reference games created using the parts in PartNet [5]. We tested with two categories separately: Chair and Table. Among the parts in *level 2* of the PartNet hierarchy, we randomly sample one of them and synthesize an utterance with template sentences (shown in Figure S2) indicating the

existence or absence of a part. Given the utterance, target and distractor shapes are also randomly sampled based on part existence. Part classes which are present or absent in less than 10% of shapes are excluded. Figure S2 illustrates some qualitative results. Finer-grained parts such as frame, support, and stretcher are accurately discovered. Table S7 also shows the average mIoUs, which are comparable with the *supervised* segmentation result using PointNet, shown in Table 3 of PartNet [5]. The average mIoUs for both PN-Agnostic and PN-Aware network trained on this dataset are 35.3 and 32.7 respectively, which are comparable with the *supervised* segmentation result using PointNet (38.2) shown in Table 2 of PartNet [5]. (But we also note that the supervised segmentation result in PartNet is the case when using the entire set of level 2 parts without any exclusion.)

S.9. Implementation Details

For Per-Super-Segment Encoder g , we used a simplified version PointNet [7]. Our network takes a set of points included in each super-segment as input and processes the points using 64-dimensional linear layers with BatchNorms and ReLUs. The features of each point are then max-pooled to produce the feature of the super-segment.

In the utterance encoders $f_a(\cdot)$ and $f_c(\cdot)$, the dimensions of the word embedding and the LSTM hidden states are set to 100 and 64, respectively. The word attention method introduced in ShapeGlot [1] is used. In the cross attention module, a single attention layer is used, which is followed by an MLP and LayerNorm.

We train our networks for 30 epochs with batch size 64 and use the ADAM [3] optimizer. The initial learning rate is 10^{-3} and decayed by a polynomial scheduler (power=0.9). Both regularization losses \mathcal{L}_{CE} and \mathcal{L}_{Coseg} are weighted by 10^{-2} . When computing cross entropy for the target shape classification and also for the regularization loss \mathcal{L}_{CE} , we follow ShapeGlot [1] and use the label smoothing technique introduced by Szegedy *et al.* [8] with the same parameters.

Sections for more qualitative results are in the following pages.

S.10. More Segmentation Results

In the following, we provide more results of the part segmentation for Chairs, Tables, Lamps, Airplanes, and Cars, as shown in Figure 1 in the main paper. All the examples in the figure below are *randomly* sampled.

| Input (Super-Seg.) | Back | Seat | Leg | Arm | Output Segments | GT |
|---|---|---|---|---|--|---|
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |













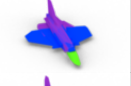
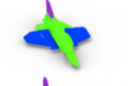






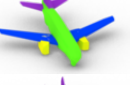


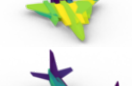




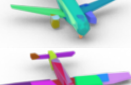





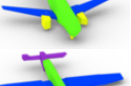







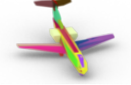





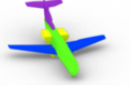













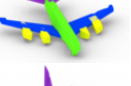









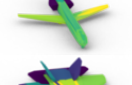
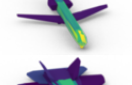


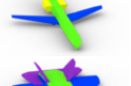





















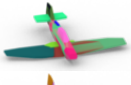




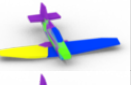

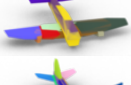
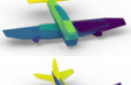










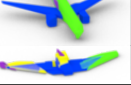
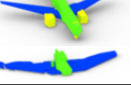










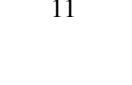










| Input (Super-Seg.) | Back | Seat | Leg | Arm | Output Segments | GT |
|-----------------------|------|------|-----|-----|--------------------|----|
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |





























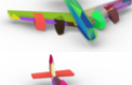











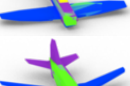















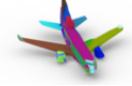



















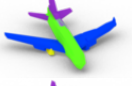







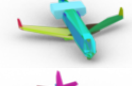
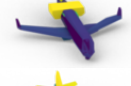
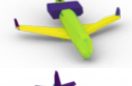
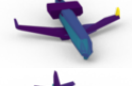

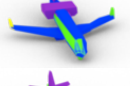
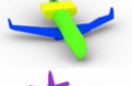














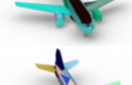




















| Input (Super-Seg.) | Back | Seat | Leg | Arm | Output Segments | GT |
|-----------------------|------|------|-----|-----|--------------------|----|
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |

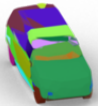
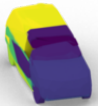



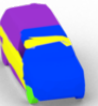
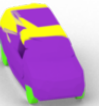
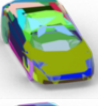
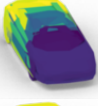
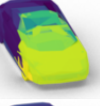
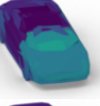


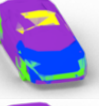
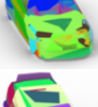
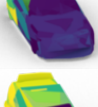
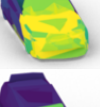
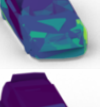
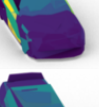
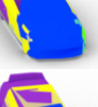
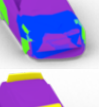
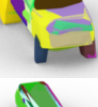
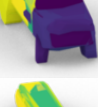
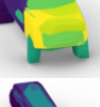
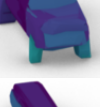
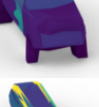
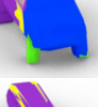
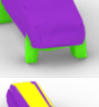
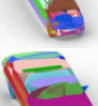
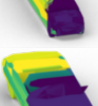
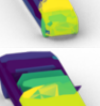
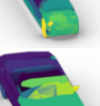
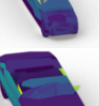
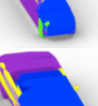
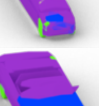
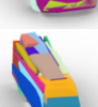
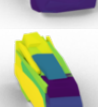
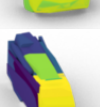
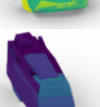
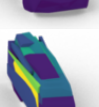
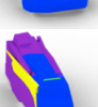
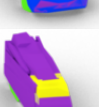
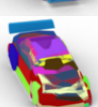
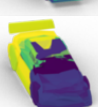
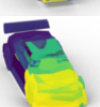
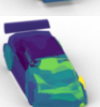



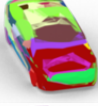
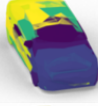
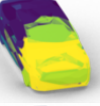


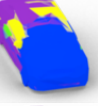
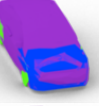
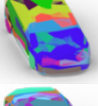
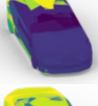
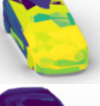
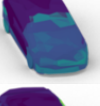

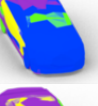

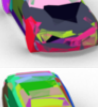
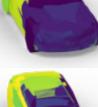
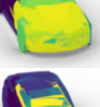
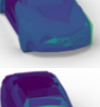
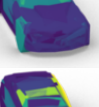
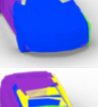

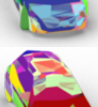
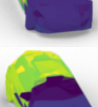
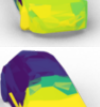
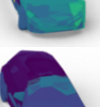


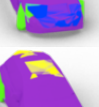
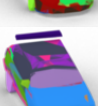
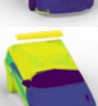

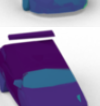
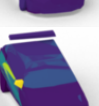
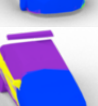
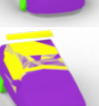
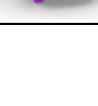
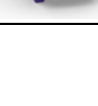



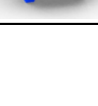
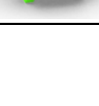







| Input (Super-Seg.) | Back | Seat | Leg | Arm | Output Segments | GT |
|-----------------------|------|------|-----|-----|--------------------|----|
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |





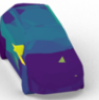




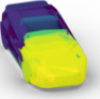
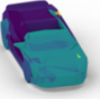



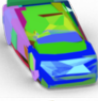
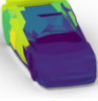
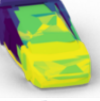

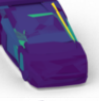
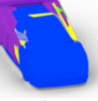
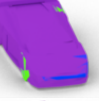
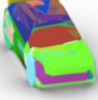

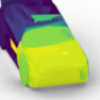
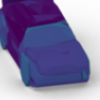


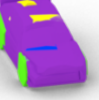
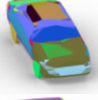
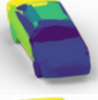
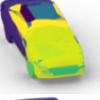
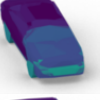

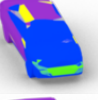


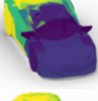



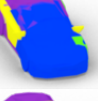



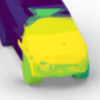
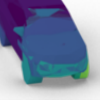


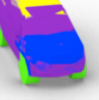
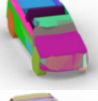
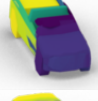
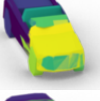
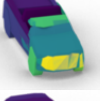
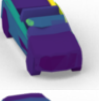
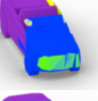


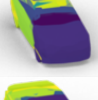
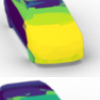
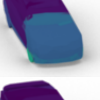

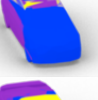
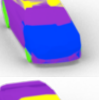
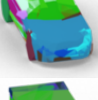
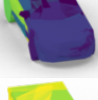
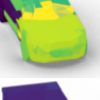
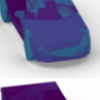
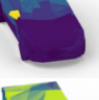

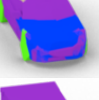
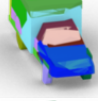

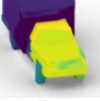
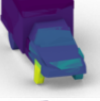
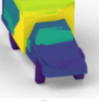
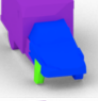



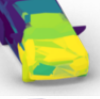




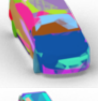
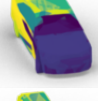
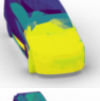

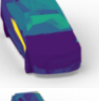
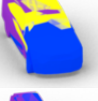

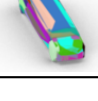
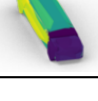
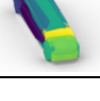


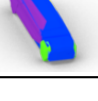
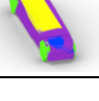
| Input (Super-Seg.) | Back | Seat | Leg | Arm | Output Segments | GT |
|-----------------------|------|------|-----|-----|--------------------|----|
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |

| Input (Super-Seg.) | Back | Seat | Leg | Arm | Output Segments | GT |
|-----------------------|------|------|-----|-----|--------------------|----|
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |

| Input (Super-Seg.) | Back | Seat | Leg | Arm | Output Segments | GT |
|---|---|---|---|--|---|---|
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

| Input (Super-Seg.) | Back | Seat | Leg | Arm | Output Segments | GT |
|---|---|---|---|---|---|---|
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

| Input (Super-Seg.) | Back | Seat | Leg | Arm | Output Segments | GT |
|---|---|---|---|---|---|---|
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

| Input (Super-Seg.) | Back | Seat | Leg | Arm | Output Segments | GT |
|---|---|---|---|---|---|---|
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

S.11. More Comparisons Results

We also provide more results of the comparison with the other methods below, as shown in Figure 4 in the main paper. All the examples in the figure below are *randomly* sampled.

| GT | PN-Agnostic (Ours) | PN-Aware (Ours) | Points | P \rightarrow Sp.-Seg. | w/o Unit Norm | $\sigma(\mathbf{X}) \rightarrow i$ | $\sigma(\mathbf{X}) \rightarrow k$ | $\frac{\sigma(\mathbf{X})}{\rightarrow i \rightarrow k}$ | w/ Global Feat. | w/o \mathcal{L}_{CE} |
|----|-----------------------|--------------------|--------|--------------------------|------------------|------------------------------------|------------------------------------|--|-----------------|------------------------|
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |

| GT | PN-Agnostic (Ours) | PN-Aware (Ours) | Points | P \rightarrow Sp-Seg. | w/o Unit Norm | $\sigma(\mathbf{X}) \rightarrow i$ | $\sigma(\mathbf{X}) \rightarrow k$ | $\frac{\sigma(\mathbf{X})}{\rightarrow i \rightarrow k}$ | w/ Global Feat. | w/o \mathcal{L}_{CE} |
|----|--------------------|-----------------|--------|-------------------------|---------------|------------------------------------|------------------------------------|--|-----------------|------------------------|
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |

| GT | PN-Agnostic (Ours) | PN-Aware (Ours) | Points | P \rightarrow Sp-Seg. | w/o Unit Norm | $\sigma(\mathbf{X}) \rightarrow i$ | $\sigma(\mathbf{X}) \rightarrow k$ | $\frac{\sigma(\mathbf{X})}{\rightarrow i \rightarrow k}$ | w/ Global Feat. | w/o \mathcal{L}_{CE} |
|----|--------------------|-----------------|--------|-------------------------|---------------|------------------------------------|------------------------------------|--|-----------------|------------------------|
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |

References

- [1] Panos Achlioptas, Judy Fan, X.D. Robert Hawkins, D. Noah Goodman, and J. Leonidas Guibas. ShapeGlot: Learning language for shape differentiation. In *ICCV*, 2019. 4
- [2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019. 3
- [3] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4
- [4] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite bert for self-supervised learning of language representations. In *ICLR*, 2020. 1, 3
- [5] Kaichun Mo, Shilin Zhu, Angel X. Chang, Li Yi, Subarna Tripathi, Leonidas J. Guibas, and Hao Su. PartNet: A large-scale benchmark for fine-grained and hierarchical part-level 3D object understanding. In *CVPR*, 2019. 1, 3, 4
- [6] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 2011. 1
- [7] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. PointNet: Deep learning on point sets for 3D classification and segmentation. In *CVPR*, 2017. 4
- [8] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016. 4
- [9] Chenyang Zhu, Kai Xu, Siddhartha Chaudhuri, Li Yi, Leonidas J. Guibas, and Hao Zhang. AdaCoSeg: Adaptive shape co-segmentation with group consistency loss. In *CVPR*, 2020. 1, 2
- [10] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *ICCV*, 2015. 3