Statistical Confidence Rescoring for Robust 3D Scene Graph Generation from Multi-View Images

Supplementary Material

A. Additional Experiments

A.1. Non-*None* Relationship Quantitative Results

As shown in Tab. 1, we additionally evaluate on the same 3RScan dataset without the $None$ class for predicate and relationship prediction. We use the same metrics of top-1 Recall and mRecall. Several works [2–4] consider the $None$ relationships crucial, while others [1] only consider annotated non-$None$ relationships. The latter approach avoids penalizing non-existent relationships that might otherwise be hallucinated while also preventing the model from overfitting to the prevalent $None$ relationships. Nonetheless, it serves as an effective mechanism to verify whether the strong performance of the model in the former setting is a result of overfitting to the $None$ relationship.

Our method performs better compared to other methods with less overfitting. Specifically, our predicate estimation surpasses all other methods even without the $None$ relationships. This suggests that our method robustly learns the other non-$None$ relationships.

Method	Recall%			mRecall%	
	Rel	Obj.	Pred.	Obj.	Pred
IMP [4]	18.3	51.8	19.3	30.0	23.0
VGfM [4]	20.8	53.3	22.1	31.6	24.4
3DSSG [4]	15.1	41.4	26.1	31.9	26.6
SGFN [4]	24.4	56.7	27.2	38.3	30.5
JointSSG [4]	25.5	58.1	27.3	43.0	33.3
IMP	14.4	48.1	16.8	35.8	19.9
VGfM	17.0	51.3	19.8	33.6	22.3
3DSSG	11.9	36.2	25.7	26.0	24.3
SGFN	21.5	53.8	24.6	35.9	27.1
JointSSG	23.1	55.1	26.6	45.4	35.2
JointSSG‡	22.4	56.8	26.9	52.9	36.2
Ours†	24.2	60.1	26.4	57.4	36.8
Ours	**25.7**	**61.1**	**27.6**	**60.5**	**39.2**

Table 1. Comparison with state-of-the-art methods on the 3RScan dataset with 20 object classes and 8 predicate classes. The top group of results are reported in [4]. The middle group of results are reproduced via 3DSSG GitHub repository. JointSSG‡ refers to JointSSG but with DINOv2 multi-view image features. Ours† refers to our method but with ResNet50 multi-view image features; Ours uses the DINOv2 multi-view image features instead. The **Best** and Second Best results are highlighted, respectively.

A.2. Quantitative Results on 160 Object and 26 Predicate Classes

We show the results of the evaluation on 160 object and 26 predicate classes for the 3RScan dataset similar to prior works. Direct comparisons on these metrics are unfair, they rely on ground truth point clouds and instance segmentation masks we do not use. We leverage only multi-view images with predicted instance masks and depth. Our method underperforms against the other methods as expected due to the inferior quality of predicted point clouds compared to the ground truth. However, it still performs comparably to SGFN for object-related metrics.

Method	Recall%			mRecall%	
	Rel	Obj.	Pred.	Obj.	Pred
SGFN(GT) [4]	64.7	36.9	48.4	16.2	14.4
JointSSG(GT) [4]	67.6	53.4	48.1	28.9	24.7
Ours(Pred)	57.7	32.1	10.9	19.6	2.3

Table 2. Comparison with state-of-the-art methods on the 3RScan dataset with 160 object classes and 26 predicate classes. The top group of results are reported in [4] which uses ground truth point clouds and instance segmentation masks. We leverage only multi-view images with predicted instance masks and depth

A.3. More Qualitative Results

For brevity in the main paper diagrams, we refer the reader to Tab. 3 for the nomenclature of the predicate notation.

We show the entire scene for scan 4d3d82b0 utilizing a similar layout format to scan 43b8cae1 in Fig. **??**. Furthermore, we show the partial scene for scan 43b8cae1 using the layout format of scan 4d3d82b0.

In Fig. 1, in addition to the incorrect node prediction for the curtain in the partial scene, we observe that our model

a	attached to
b	build in
c	connected to
h	hanging on
n	none
p	part of
s	standing on
u	supported by

Table 3. Nomenclature for predicates in the diagrams of our main paper.

also mis-classifies another furniture object as a door. This mis-classification likely arises due to the thin structure of the object, which closely resembles that of a door.

In Fig. 2, only our method successfully classifies the rare bookshelf class. Our method also correctly identifies an office table with a distinctive shape while other methods misclassify it as a cabinet. This demonstrates the superior ability of our method to recognize both rare object classes and uncommon shapes of common objects, such as tables.

We provide qualitative results on one additional scan with the same format of partial scene in Fig. 3 and full scene in Fig. 4, respectively.

In Fig. 3, SGFN fails to classify non-background object classes and all predicate classes. Although JointSSG outperforms SGFN, it misclassifies the rare refrigerator class and fails to detect the orange door. Our method correctly classifies the refrigerator, but also misses the door. As a result, the predicate linking the door to the wall is misclassified by all methods.

In the full scene shown in Fig. 4, our method correctly classifies most objects, but misses predictions on objects that are more difficult to discern. For example, counter, picture, and sink. Additionally, our method fails to classify the window, which is obscured beyond the bottom left of the scene. The light purple cabinet on the left is misclassified as other furniture, likely due to its atypical orientation compared to standard cabinets.

Our method also fails to predict $build\ in$ and $part\ of$ predicates since it does not recognize the presence of the counter and sink. However, it successfully predicts the challenging $connected\ to$ relationship between the dark green wall and the blue other furniture. Furthermore, it generates multiple plausible unseen predicate predictions for the $standing\ on$ relationship between various objects and the floor.

B. Analysis

B.1. Analysis on Logits

We analyze the impact of softmax logits on node estimation accuracy in Fig. 5 using the 3RScan test set. To assess this, we bin softmax probability predictions into 0.1 intervals and evaluate both the accuracy (represented by the bar chart) and the frequency of object predictions (represented by the line curve) within each bin. Each bin includes instances within the range $[x - 0.1, x)$, where x corresponds to the labeled bin on the x-axis. The softmax probabilities serve as a pseudo-confidence measure for each prediction.

For improved model performance, it is desirable to minimize the number of low-confidence predictions as indicated by the line curve for 3DSSG in Fig. 5. This is because lower-confidence predictions generally correspond to lower accuracy. 3DSSG exhibits a high frequency of low-confidence predictions with softmax probabilities below 0.5 and fewer high-confidence predictions. Similarly, IMP and VGFM also suffer from this issue. The unusually high performance of VGFM in the $[0.1, 0.2)$ bin is likely an outlier and can be ignored.

As evidenced by the rightward shift in the line curve, SGFN and JointSSG reduce the number of low-confidence predictions. Our method exhibits an even more pronounced shift, culminating in a sharp peak of nearly 400 object instances in the $[0.9, 1.0]$ bin. Additionally, our method achieves the highest accuracy in this bin, indicating that the majority of high-confidence predictions are correct. Combined with the insights from Fig. **??**, this suggests that the CR module plays a crucial role in increasing high-confidence predictions by enhancing the predicted node probability for the correct class.

Method	Recall%			mRecall%	
	Rel	Obj.	Pred.	Obj.	Pred
Ours	40.5	61.8	90.4	60.5	39.2
$\alpha = 0.8$	37.8	58.1	91.6	53.5	27.8
$\alpha = 0.2$	31.0	53.0	90.5	46.5	26.2

Table 4. Experiment on our method with a fixed confidence score threshold.

B.2. Analysis on CR Components

To investigate the impact of the components of the CR module, we conduct experiments where we fix the confidence score α.

As shown in Tab. 4, CR fails without the confidence score α modulating the input of the statistical prior, resulting in worse performance than our reproduced baseline for most metrics except predicate recall. For predicate recall, as established in A.1, the models might overfit to the $None$ relationship. As the mRecall for predicate estimation does not correspondingly improve with the predicate recall, this suggests that the CR module with a fixed α does not help the model robustly learn the other non-$None$ relationships. The adaptive value of α is crucial to attaining good performance on the task.

B.3. Analysis on CR Generalization Ability

The dependency of CR on fixed co-occurrence statistics from the training data is a limitation of our approach. To evaluate robustness to distribution shifts, we conduct a simulated experiment by removing frequently occurring relationships from the dataset.

Tab. 5 shows that there is a significant decrease in performance for predicate and relationship estimation while object estimation performance metrics only decreased slightly. As the $None$ relationship is included in the top relation-

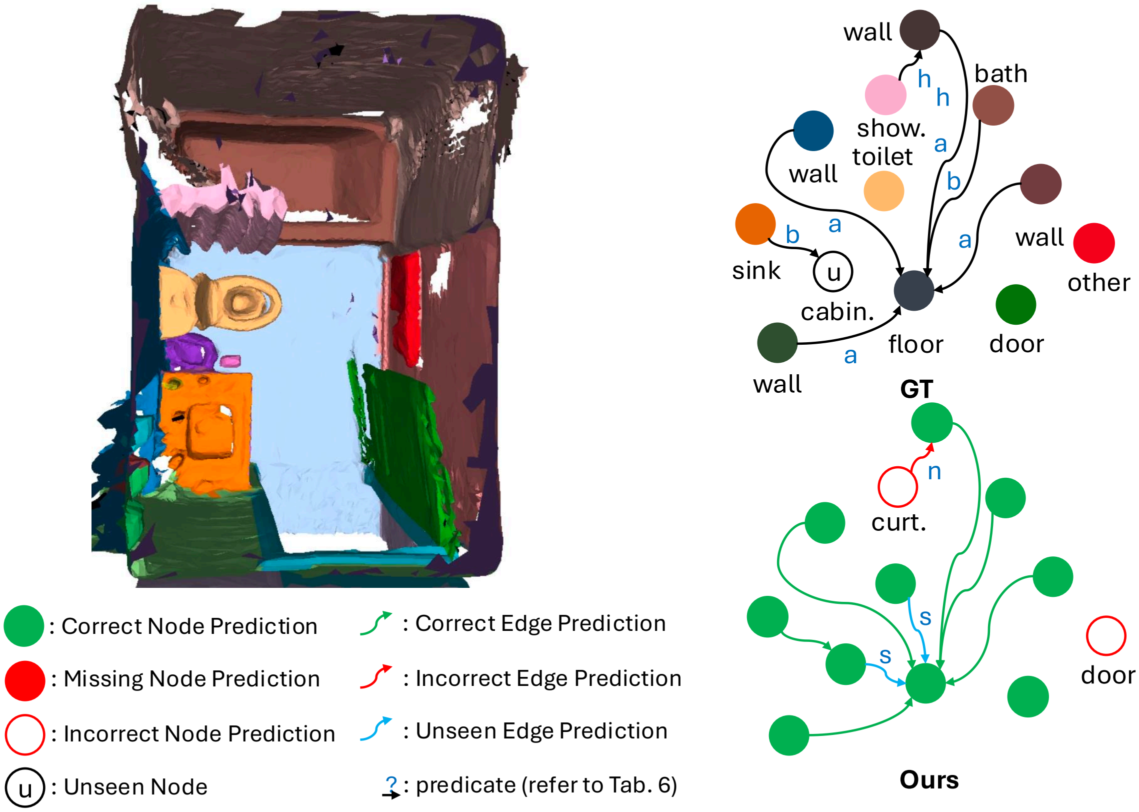
wall
h
h
bath
show.
a
toilet
wall
b
a
b
a
wall
sink
u
other
cabin.
floor
door
wall
a
GT
n
curt.
s
s
door
Ours
: Correct Node Prediction
: Missing Node Prediction
: Incorrect Node Prediction
u : Unseen Node
: Correct Edge Prediction
: Incorrect Edge Prediction
: Unseen Edge Prediction
?: predicate (refer to Tab. 6)

Figure 1. The qualitative results on 3RScan dataset on the entire scene of scan 43b8cae1. Our method can correctly classify most objects in the scene except the misclassified pink shower curtain and other furniture depicted in red. Our method can reliably predict all predicate classes in the scene except for the predicate between the wall and the shower curtain.

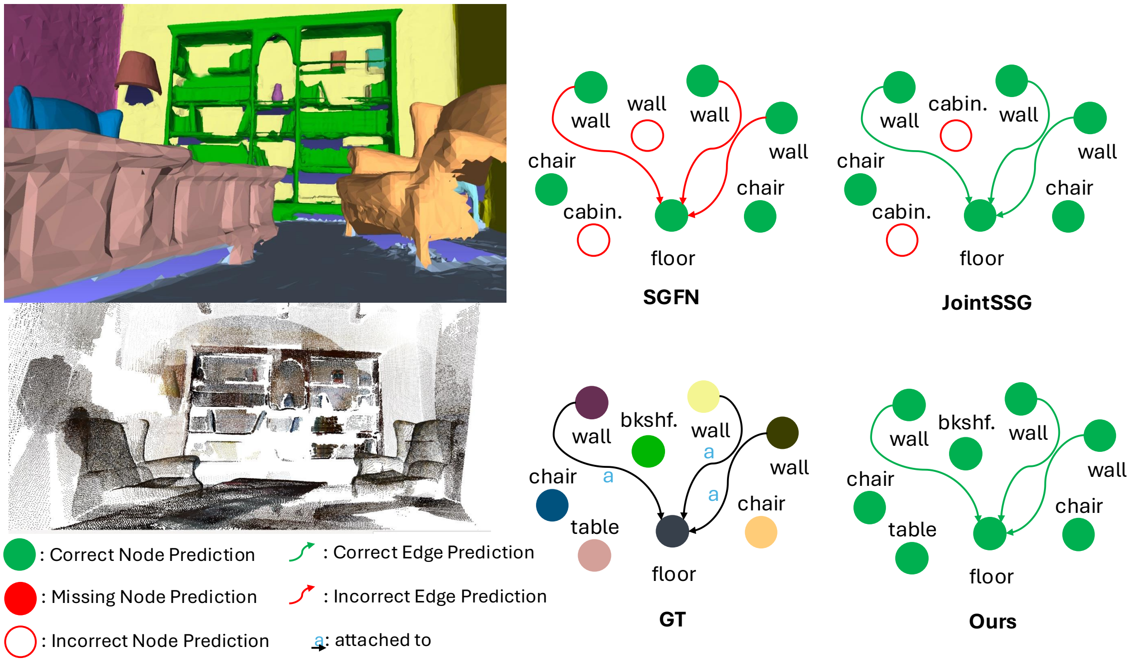

Figure 2. The qualitative results on 3RScan dataset of previous works and our proposed framework on scan 43b8cae1. Our method can correctly classify all objects in the partial scene, including the green bookshelf and pink table that competing methods misclassify.

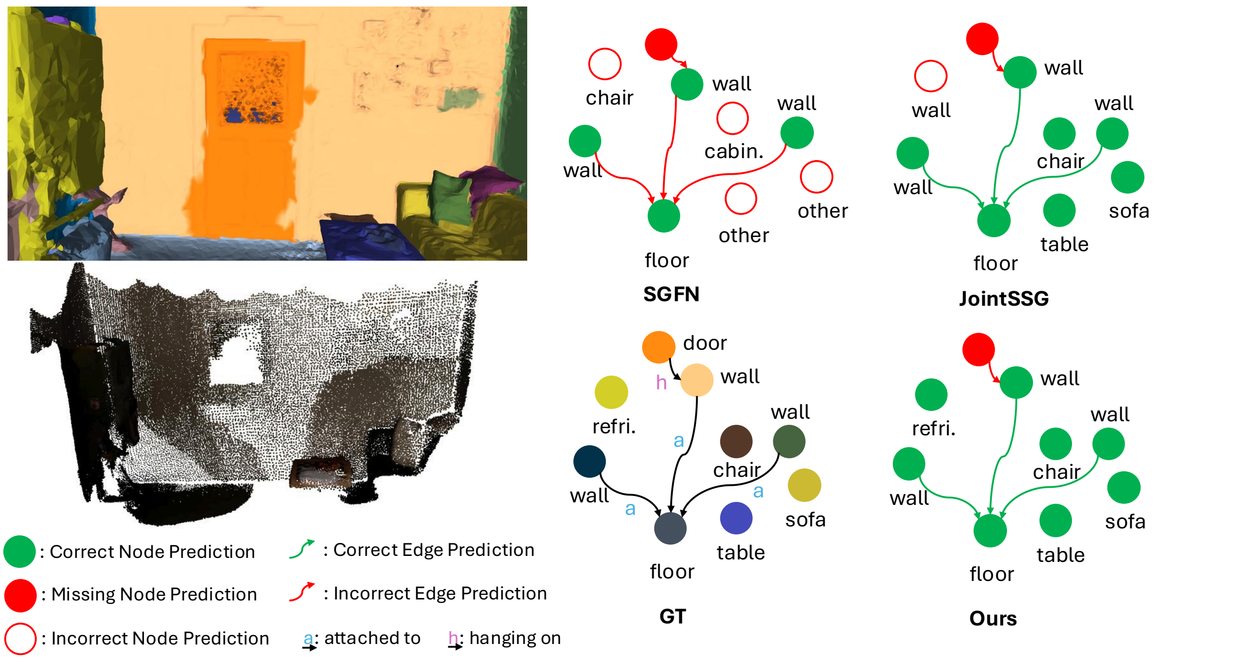

Figure 3. The qualitative results on 3RScan dataset of previous works and our proposed framework on scan 4a9a43. SGFN fails to classify non-background object classes and all predicate classes. JointSSG performs better but misclassifies the refrigerator as a wall. The method also fails to detect the door. Our method correctly classifies the refrigerator but it also fails to detect the door.

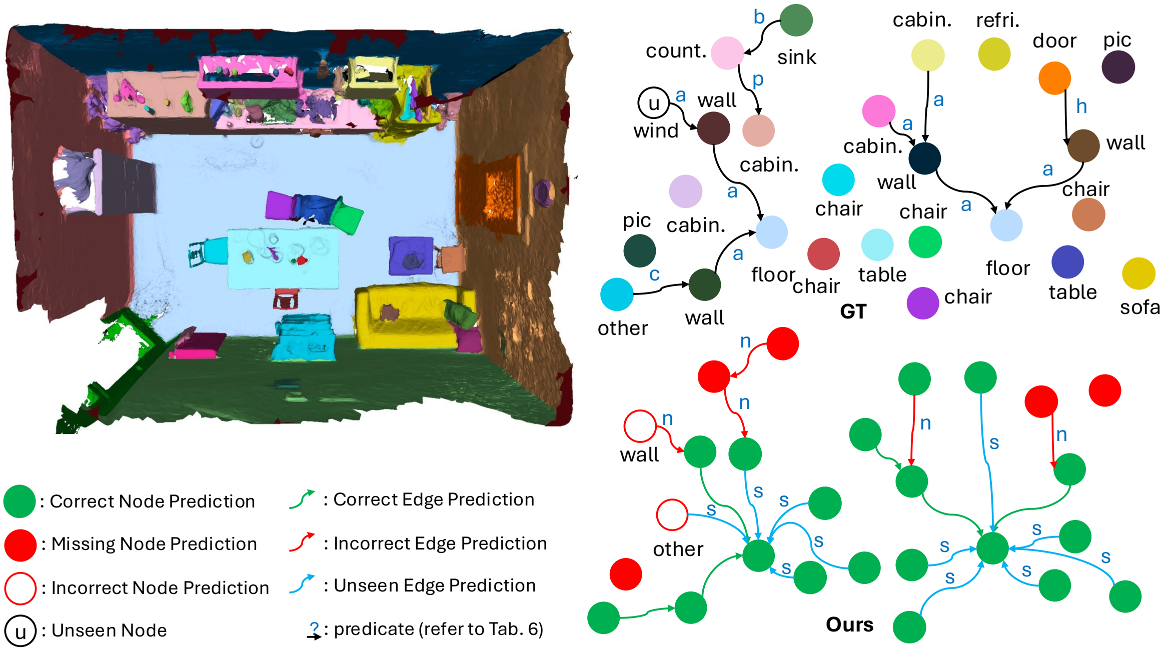

Figure 4. The qualitative results on 3RScan dataset of our proposed framework on the entire scene of scan 43b8cae1. Our method can correctly classify most objects in the scene, though it may fail to predict the presence of some objects such as counter, picture, and sink. Our method can predict most predicate classes in the scene except for predicates with undetected neighboring objects.

ships to be removed, we attached the experiment without the *None* relationship for reference. The decrease in performance for predicate and relationship estimation compared to the experiment referenced above is less significant, which suggests some level of robustness of our method to distribution shifts.

For a broader generalization, adaptive co-occurrence statistics from visual foundation models could be explored,

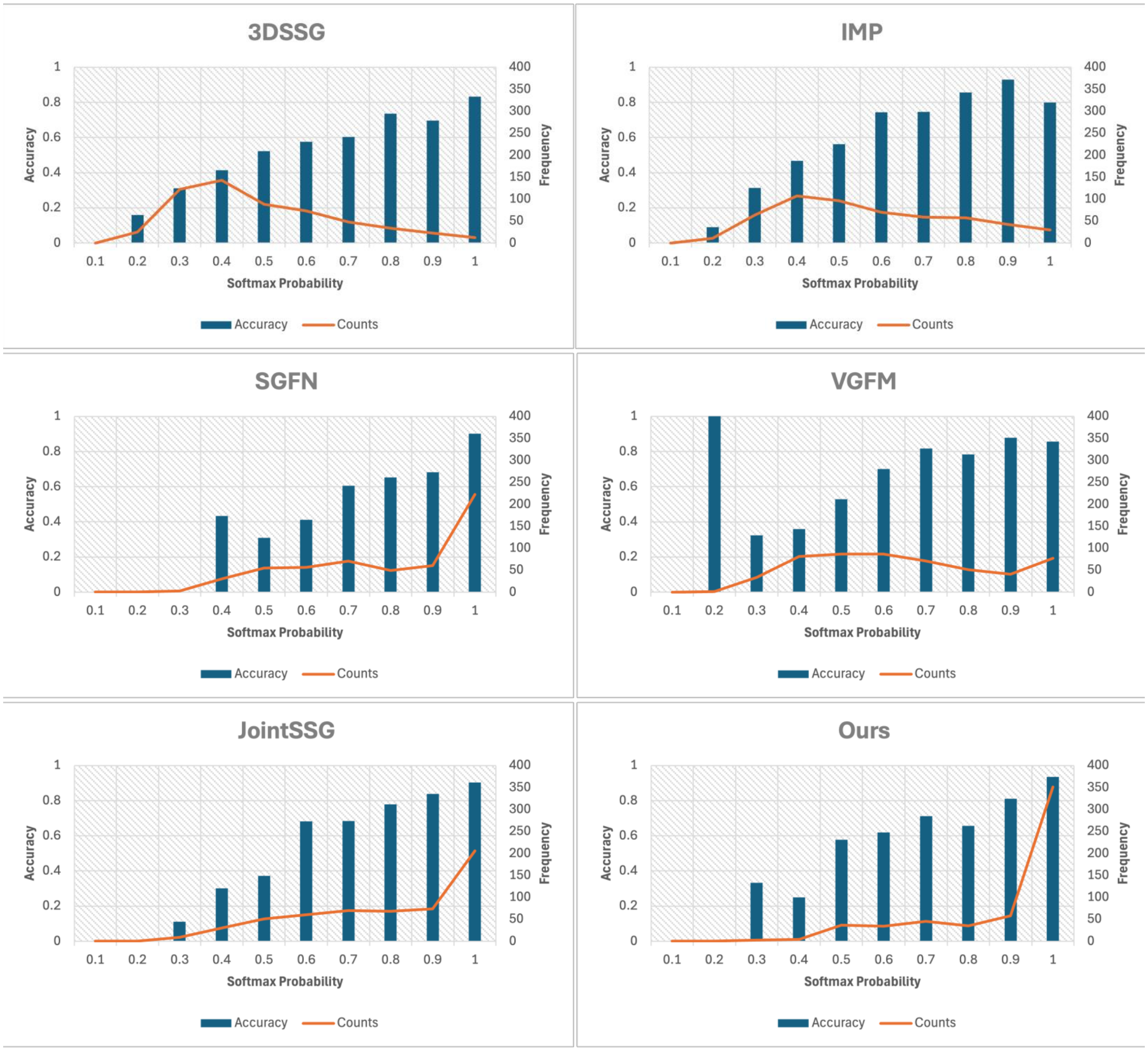

Figure 5. Histogram of accuracy of object predictions with pseudo confidence bins for previous methods. More of our object predictions falls within the bins to the right as shown by the line graph suggesting more object predictions achieves higher pseudo confidence levels. Our method obtains higher accuracy at higher pseudo confidence levels (softmax probability) as shown by the bar graph.

but this lies beyond the scope of our current closed-set setting.

References

[1] Paul Gay, James Stuart, and Alessio Del Bue. Visual graphs from motion (vgfm): Scene understanding with object geometry reasoning. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*, pages 330–346. Springer, 2019. 1

[2] Johanna Wald, Helisa Dhamo, Nassir Navab, and Federico Tombari. Learning 3d semantic scene graphs from 3d indoor reconstructions. In *Proceedings of the IEEE/CVF Conference*

Method	Recall%			mRecall%	
	Rel	Obj.	Pred.	Obj.	Pred
Ours	40.5	61.8	90.4	60.5	39.2
Ours (w/o $None$)	25.7	61.1	27.6	60.5	39.2
Remove Top 20%	24.5	61.1	27.0	59.5	26.6
Remove Top 50%	18.9	60.1	20.6	56.5	19.1

Table 5. Simulated experiment on our method's robustness to distribution shifts. We remove the most frequently occuring relationships according to the arbitrarily determined percentage in the first column of the table.

on Computer Vision and Pattern Recognition, pages 3961–3970, 2020. 1

[3] Shun-Cheng Wu, Johanna Wald, Keisuke Tateno, Nassir Navab, and Federico Tombari. Scenegraphfusion: Incremental 3d scene graph prediction from rgb-d sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7515–7525, 2021.

[4] Shun-Cheng Wu, Keisuke Tateno, Nassir Navab, and Federico Tombari. Incremental 3d semantic scene graph prediction from rgb sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5064–5074, 2023. 1