Pacific Graphics (2022)
N. Umetani, E. Vouga, and C. Wojtan (Editors)

Multi-instance Referring Image Segmentation of Scene Sketches based on Global Reference Mechanism -Supplemental Material-

Peng Ling, Haoran Mo, and Chengying Gao†

Sun Yat-sen University

1. Comparison with Baseline Methods

1.1. Multi-instance Referring Image Segmentation

Figure 1 and Figure 2 show the results of our GRM-Net and the baseline method by Zou *et al*. [ZMG*19].

1.2. Standard Referring Image Segmentation

Figure 3 shows the results of our GRM-Net and the baseline methods (*e.g*., DMN [MTPBA18], CMSA [YRLW19], CMPC [HHL*20], and CEFNet [FHZL21]) for standard task of referring expression segmentation.

2. Ablation Study

2.1. Fusion Stage(s) of Language Features

We introduce a two-step fusion scheme for the language features in a coarse-to-fine manner. As illustrated in the Pipeline (shown in the Fig.2 in the main paper), the expression information is fused into the segmentation pipeline both in the RPN Fusion Step and the RoI Fusion Step. We thus compare with the performance of fusion in each step individually.

The quantitative comparisons are shown in Table 1, from which we can see our approach with two-step fusion works the best compared with the models with a single fusion step. The qualitative results in Fig. 4 are in line with the quantitative results. When the fusion is done in the RPN Fusion Step only, a rough filtering is performed and there remain a certain number of proposals irrelevant to the expression. Without the language information for a finer selection in the latter RoI Fusion Step, the model has no guidance on discarding the undesired candidates, and tends to predict all these proposals as target objects. On the other hand, when the language information is fused in the RoI Fusion Step only without a rough filtering in the RPN Fusion Step, some irrelevant instances which should be rejected originally serve as references for the RoIs in the GRM module. This probably introduces noise and reduces the performance by producing incorrect instances as shown in Fig. 4. In contrast, our method with two-step fusion produces correct results.

Fusion stage(s)	AP	AP_{50}	AP_{75}
RPN	31.24	38.73	32.57
RoI	49.12	61.71	49.47
RPN & RoI (ours)	**59.39**	**71.37**	**62.07**

Table 1: *Ablation studies of fusion stage(s) in the segmentation pipelines for the language features.*

2.2. Configurations of Global Reference Mechanism

Figure 5 shows the comparison results between models with and without global reference mechanism (GRM).

3. Effectiveness of Global Reference Mechanism

Figure 6 shows the effectiveness of global reference mechanism (GRM).

4. Conclusion and Limitations

Figure 7 and 8 show the failure cases in complicated scene sketches and complicated expressions, respectively.

References

[FHZL21] FENG G., HU Z., ZHANG L., LU H.: Encoder fusion network with co-attention embedding for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 15506–15515. 1

[HHL*20] HUANG S., HUI T., LIU S., LI G., WEI Y., HAN J., LIU L., LI B.: Referring image segmentation via cross-modal progressive comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 10488–10497. 1

[MTPBA18] MARGFFOY-TUAY E., PÉREZ J. C., BOTERO E., ARBELÁEZ P.: Dynamic multimodal instance segmentation guided by natural language queries. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 630–645. 1

[YRLW19] YE L., ROCHAN M., LIU Z., WANG Y.: Cross-modal self-attention network for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 10502–10511. 1

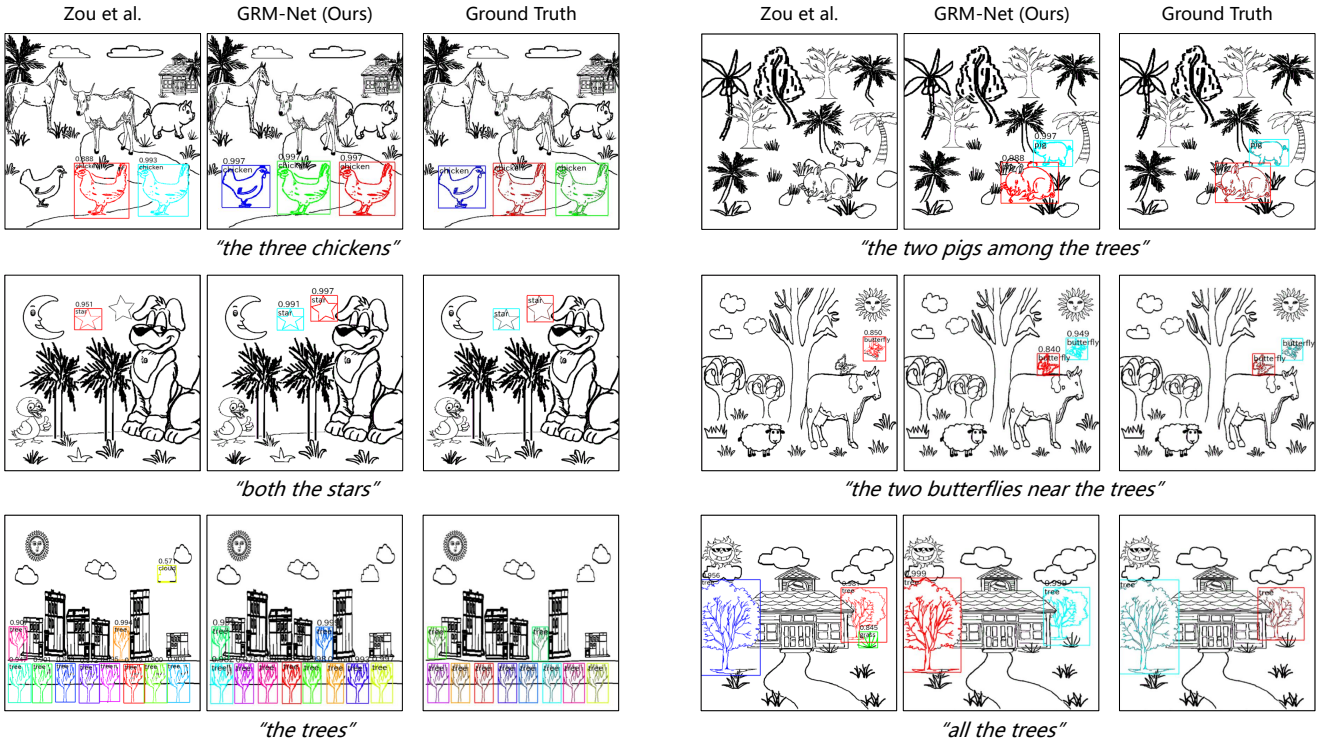

Figure 1: *Comparisons with the baseline method by Zou et al. [ZMG*19].*

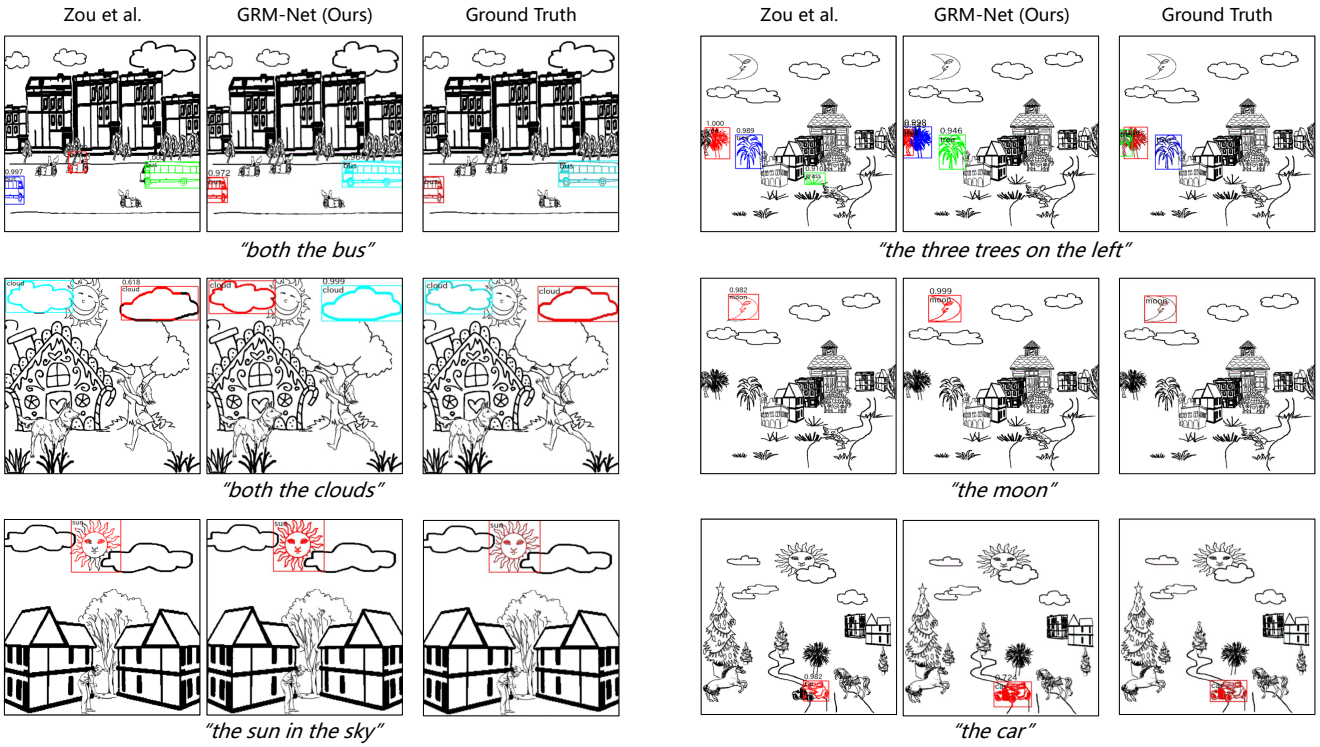

Figure 2: *Comparisons with the baseline method by Zou et al. [ZMG*19].*

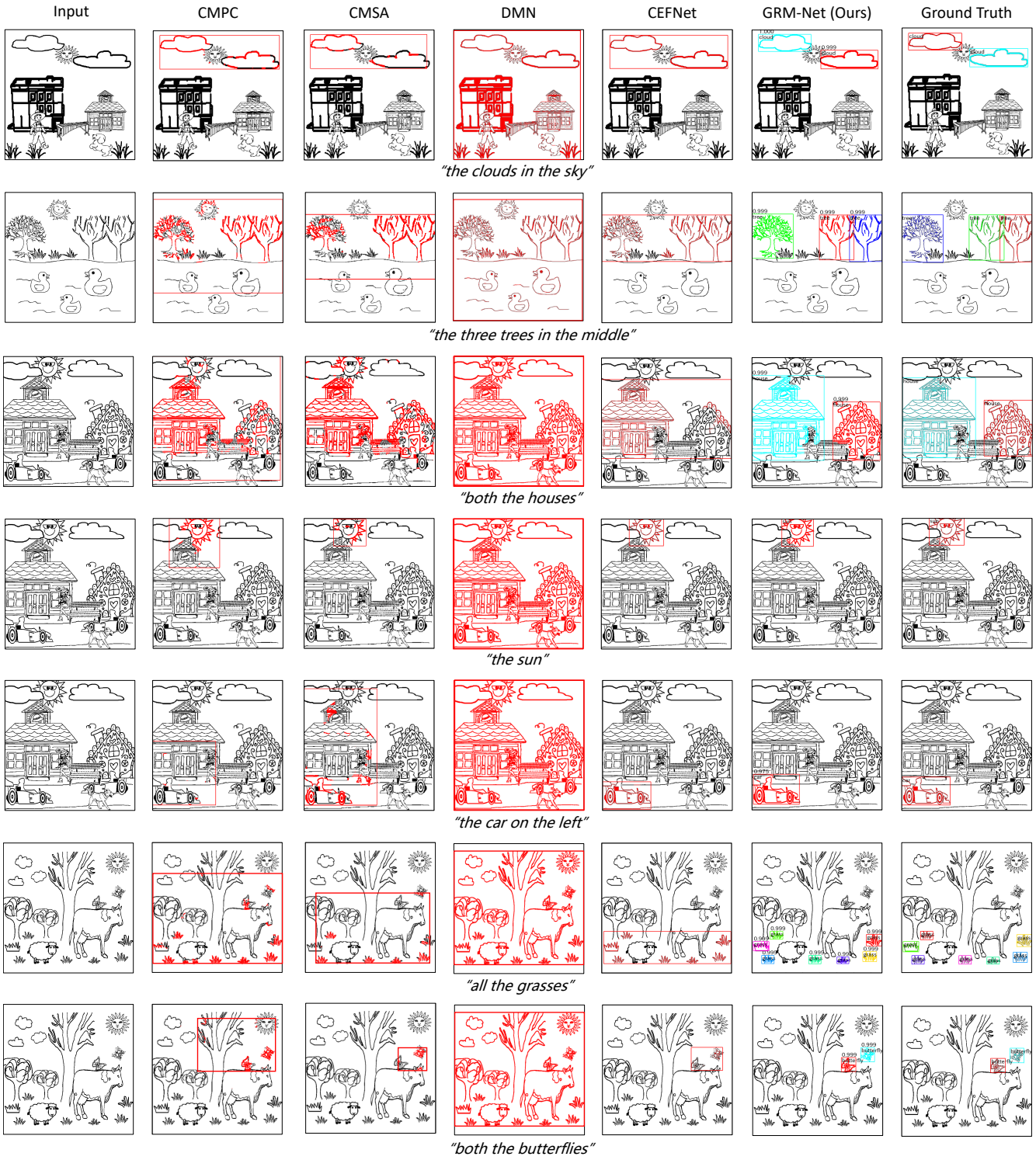

Figure 3: *Comparisons with the standard RES methods.*

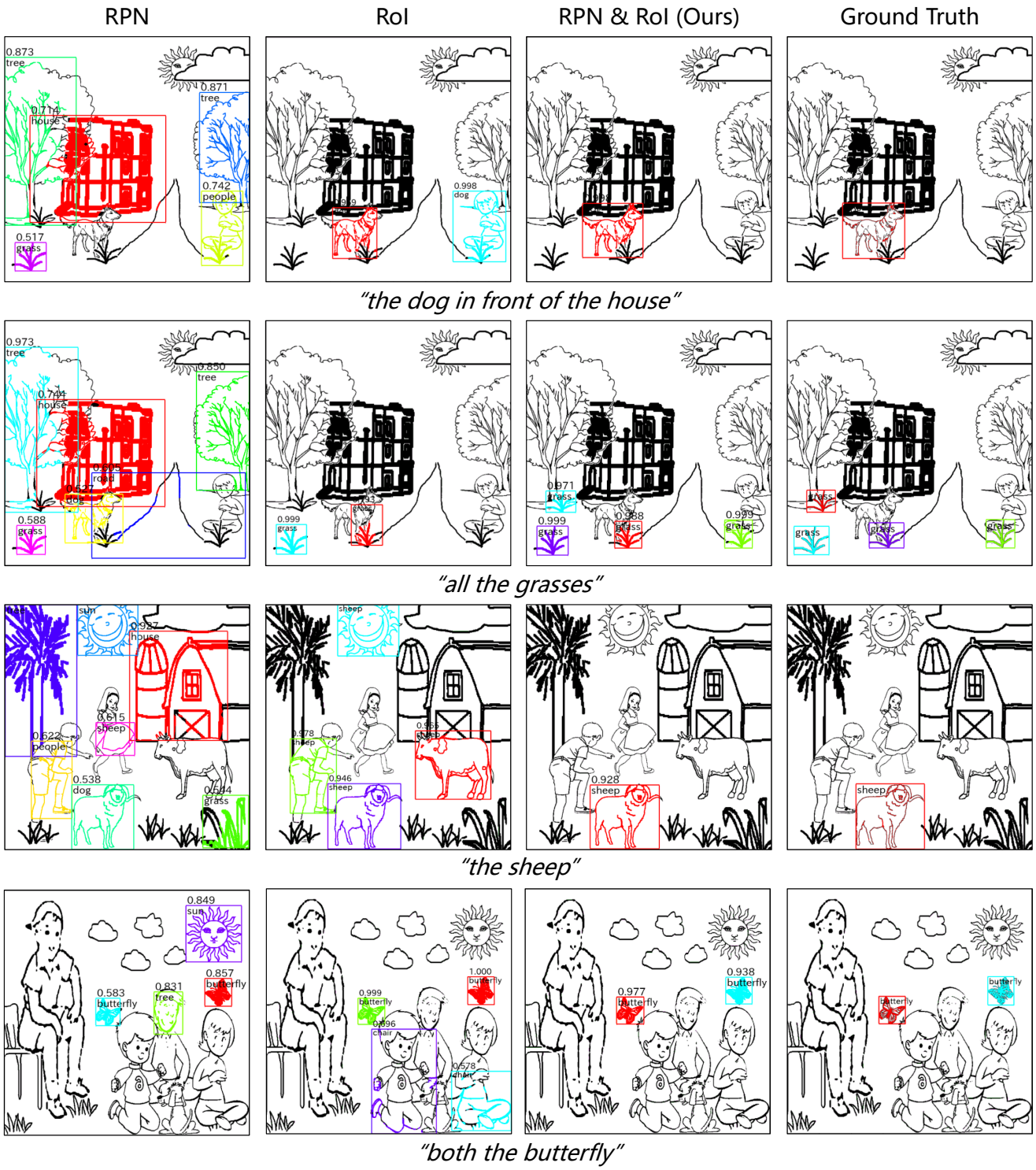

Figure 4: *Comparisons of fusion stage(s) in the segmentation pipelines for the language features.*

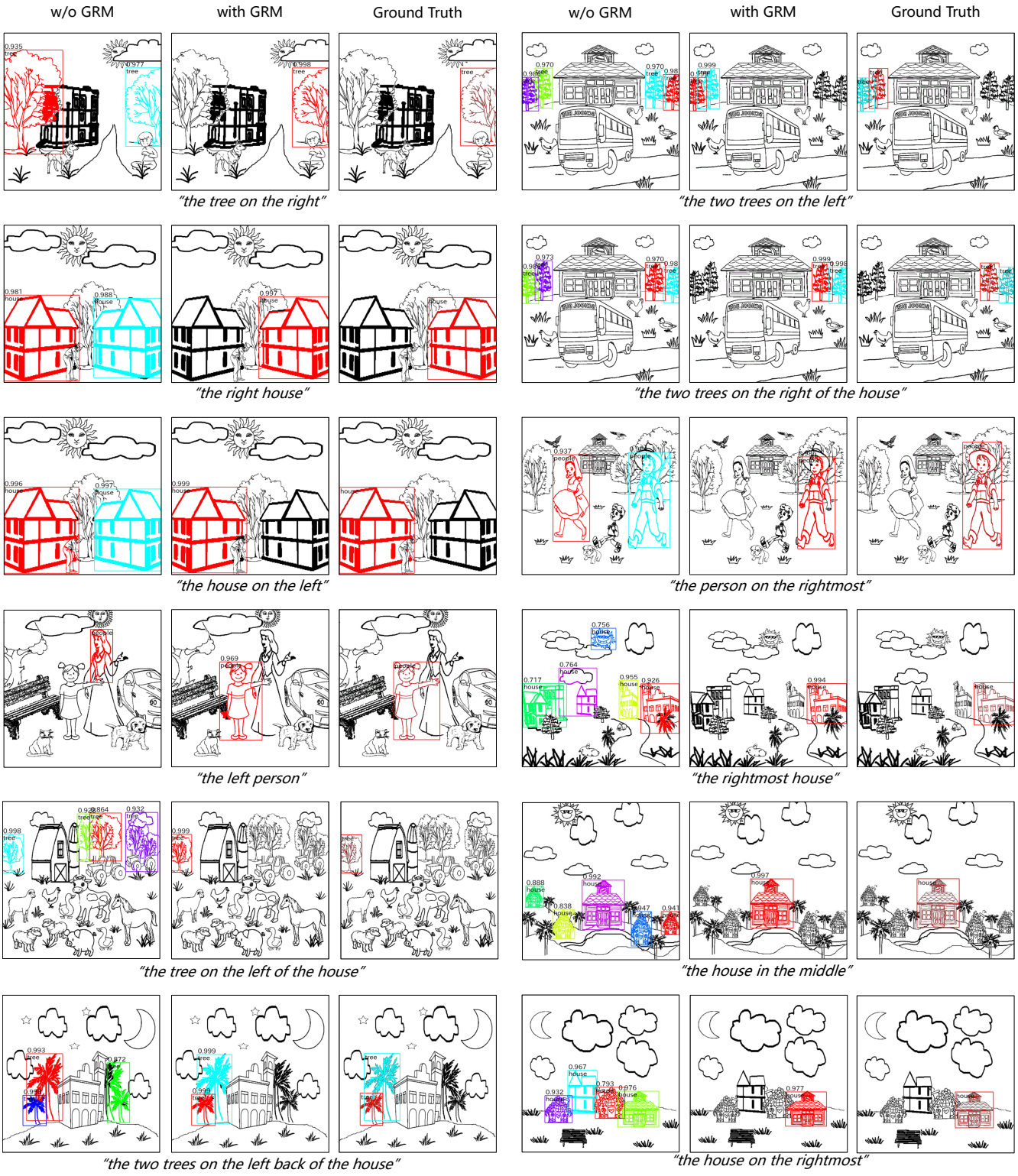

Figure 5: *Comparisons between models with and without global reference mechanism (GRM).*

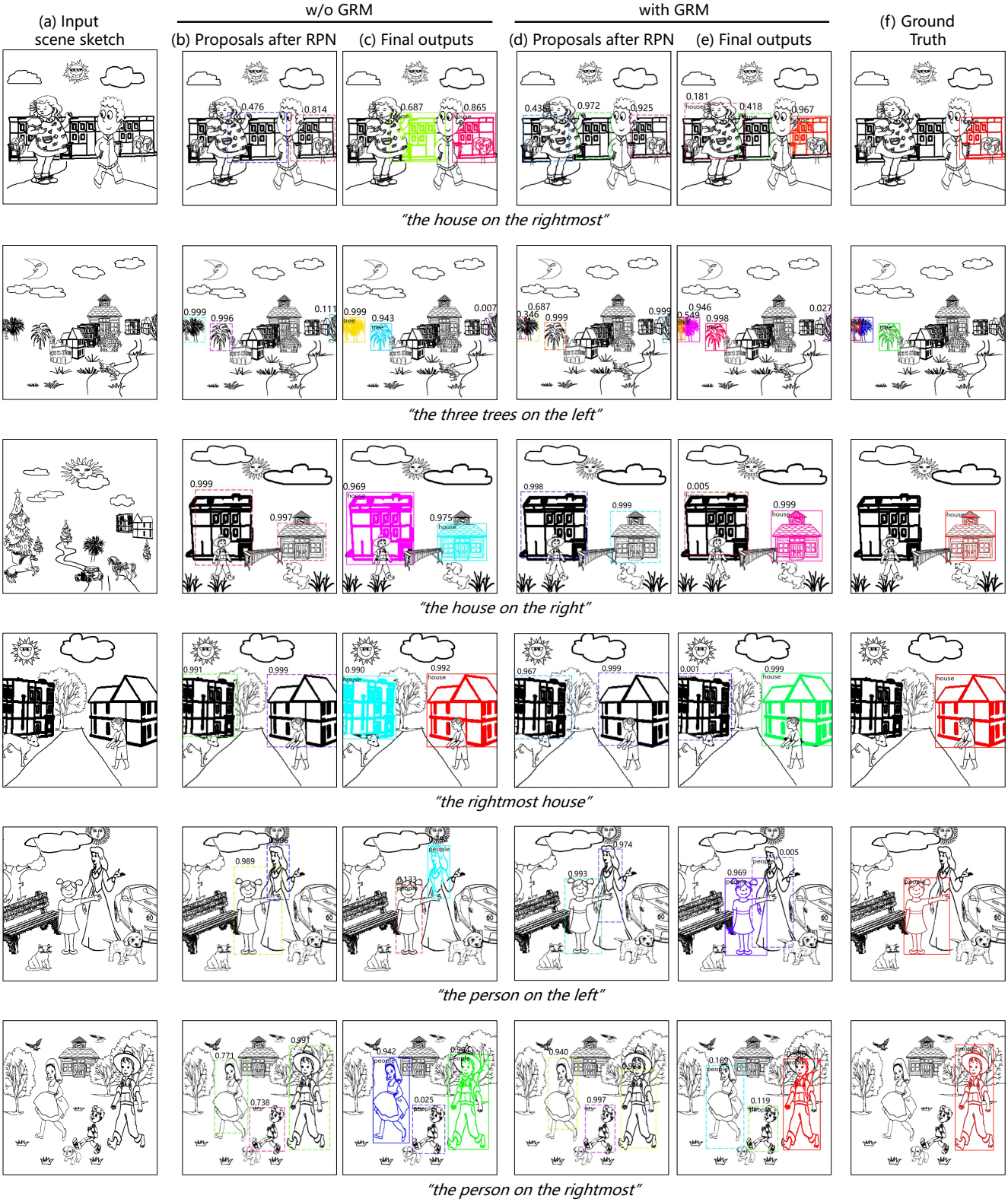

Figure 6: *Effectiveness of global reference mechanism. Boxes in dashed line in (b) and (d) are the proposals after the RPN stage (other proposals are omitted for brevity). Boxes in solid line in (c) and (e) are the final instance segmentation. In (e), the boxes in dashed line are assigned non-object labels and thus are* ***not*** *the final output instances; the numbers and category labels around them are the largest confidence except the non-object class and the corresponding classes.*

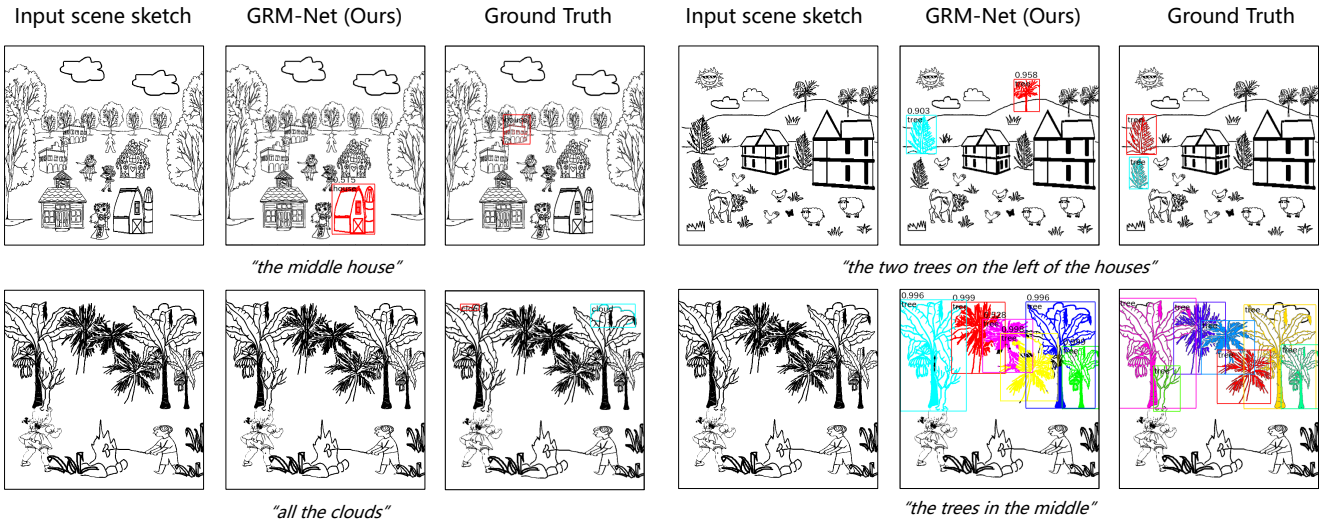

Figure 7: *Visualization of the failure cases in the complicated sketch scene.*

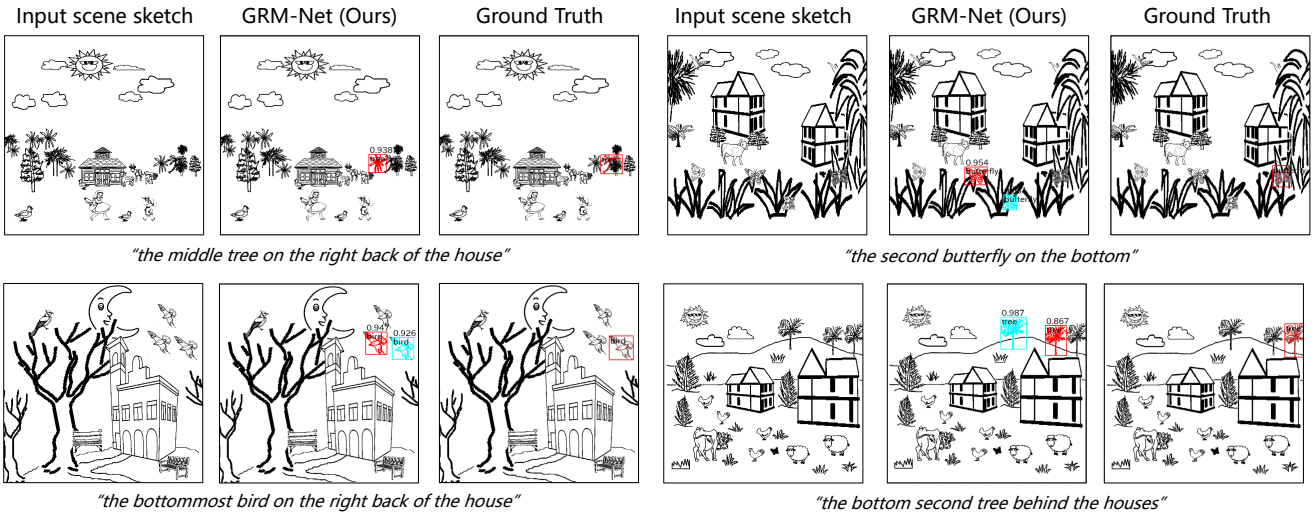

Figure 8: *Visualization of the failure cases in the complicated expression.*

[ZMG*19] ZOU C., MO H., GAO C., DU R., FU H.: Language-based colorization of scene sketches. *ACM Transactions on Graphics (TOG) 38*, 6 (2019), 1–16. 1, 2