Pacific Graphics (2022)
Y. Yang, A. D. Parakkat, B. Deng, and S.T. Noh (Editors)

Short Paper

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

Peng Ling, Haoran Mo, and Chengying Gao†

Sun Yat-sen University

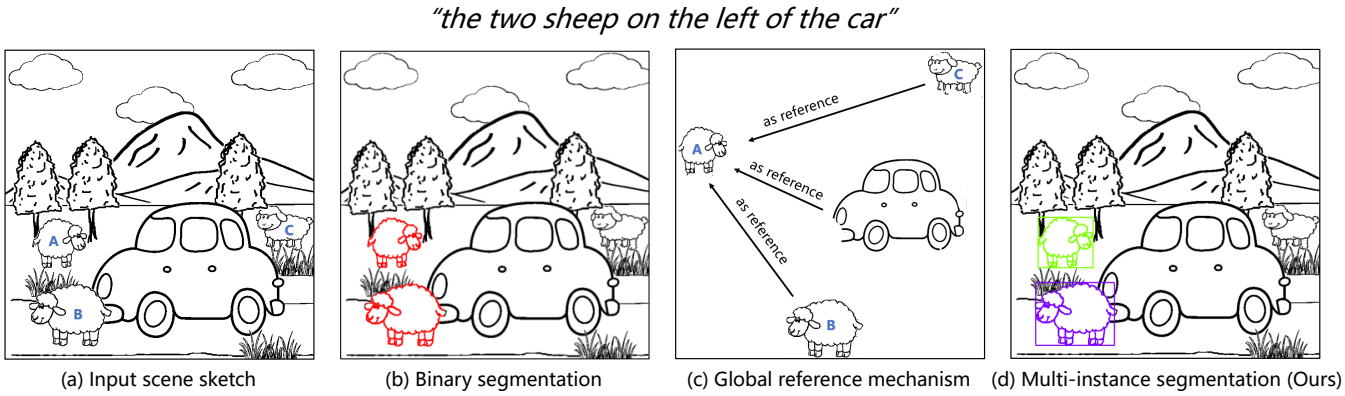

Figure 1: *Comparison between conventional binary segmentation (b) and multi-instance segmentation (d) of scene sketches based on input referring expression. In our framework, we propose a global reference mechanism (c), where each candidate is assigned references to identify its global position. The sheep instances are indexed to avoid confusion.*

Abstract

Scene sketch segmentation based on referring expression plays an important role in sketch editing of anime industry. While most existing referring image segmentation approaches are designed for the standard task of generating a binary segmentation mask for a single or a group of target(s), we think it necessary to equip these models with the ability of multi-instance segmentation. To this end, we propose GRM-Net, a one-stage framework tailored for multi-instance referring image segmentation of scene sketches. We extract the language features from the expression and fuse it into a conventional instance segmentation pipeline for filtering out the undesired instances in a coarse-to-fine manner and keeping the matched ones. To model the relative arrangement of the objects and the relationship among them from a global view, we propose a global reference mechanism (GRM) to assign references to each detected candidate to identify its position. We compare with existing methods designed for multi-instance referring image segmentation of scene sketches and for the standard task of referring image segmentation, and the results demonstrate the effectiveness and superiority of our approach.

CCS Concepts

• ***Computing methodologies*** → *Scene understanding; Image Segmentation;*

1. Introduction

Scene-level sketch editing plays a fundamental role in the industry of anime production. Efficient ways for the manipulation have been explored, and language expression is generally considered as the most user-friendly and convenient one. Given that segmentation serves as a critical step in the editing pipeline, the task of scene sketch segmentation based on referring expression has come into being. While there exist plenty of studies of the referring image segmentation task [LLS*17, HFS*20, FHZL21, YRLW19, YLS*18, DLWJ21], the majority of them work with natural images, and very few focus on sketches that have unique characteristics, *e.g.*, spar-

† Corresponding author: mcsgcy@mail.sysu.edu.cn

DOI: 10.2312/pg.20221238

sity and abstraction. Scene sketches, theoretically, can be treated as normal images and processed by the models for natural images, but most methods are tailored for binary mask generation with only one output mask covering a single target instance (in most cases) or a group of objects, as shown in Fig. 1-(b). In the pipeline of anime production where finer-grained manipulation is required, the individual segmentation for each of the multiple instances (Fig. 1-(d)) should be produced, including a bounding box, a segmentation mask and a category label. Therefore, it is critical to promote the models to develop the ability of multi-instance segmentation based on expressions, in order to meet the demands for more forms of human-computer interaction in the future.

To the best of our knowledge, Zou *et al*. [ZMG*19] propose the only method for *multi-instance referring image segmentation of scene sketches*, where all instances are segmented first and then the desired targets are selected through a filtering algorithm. With a pre-training stage and another step to obtain the final results, it disables end-to-end joint training, which tends to cause error accumulation when the performance of the latter stage is particularly dependent on the former. To overcome this issue, we propose a one-stage framework, named *GRM-Net*, for multi-instance referring image segmentation of scene sketches. We integrate the expression information into an instance segmentation pipeline to filter out the mismatched instances via a one-stage training. Therefore, the weights of the entire model are optimized jointly, which avoids error accumulation and helps to improve the performance.

Similar to the classic instance segmentation workflow [HGDG17], candidate regions of interest (RoIs) go through a coarse-to-fine filtering process. Each of them stores regional information and lacks a global view of its spatial relationship with other instances and the overall arrangement of all objects. As a result, it increases the difficulty of the model to judge whether the candidate matches the spatial information conveyed by the expression. To provide each detected candidate with a global view, we propose a *global reference mechanism (GRM)* which assigns object references for each candidate as shown in Fig. 1-(c). The references for each candidate are found in a global manner, that is, selected largely according to the consistency to the expression learned by the model instead of the local spatial distance to the candidate. For example, when sheep A is the candidate, not only the car, but also sheep B and even sheep C that is not spatially close to sheep A are considered as references because "sheep" is mentioned in the expression. This mechanism offers more meaningful references to the candidate, which helps the segmentation model to identify the positions of the candidates globally and improves its ability of responding to the location information in the expression.

We evaluate our proposed approach on a scene sketch dataset [ZMG*19] through comprehensive ablation studies and comparisons with the advanced methods for both multi-instance referring image segmentation task and standard referring image segmentation task. The results show that our approach outperforms the advanced methods both quantitatively and qualitatively.

The main contributions of this work are as follows:

- A one-stage framework *GRM-Net* for multi-instance referring image segmentation of scene sketches, with a two-step language fusion in a coarse-to-fine manner.
- A global reference mechanism (GRM) that offers a global view of object arrangement and helps to identify the positions for the candidate instances.
- Visualization of how the proposed global reference mechanism (GRM) works and in-depth comparisons with existing approaches that show the efficiency of our method.

2. Method

2.1. Overview

As illustrated in Fig. 2, our framework *GRM-Net* designed for multi-instance referring image segmentation of scene sketches is built on the combination of a conventional instance segmentation model Mask R-CNN [HGDG17], a language model, and a global reference mechanism (GRM) module. The language model encodes the input expression and serves as a filtering module to offer information of the instances specified by the expression. Mask R-CNN originally outputs all the instances, and when fused with the filtering information, it is able to learn to discard the outliers and keep the correct ones. Given that modeling spatial relationships among objects is a non-trivial task based on language information only, we propose a global reference mechanism to assign references to each object proposal and provide the segmentation model with a global view on the relative positioning among objects. With this mechanism, the model learns to better distinguish between correct (*e.g*., "the two trees on the left of house" in Fig. 2) and incorrect instances (*e.g*., the rightmost tree) based on the expression.

2.2. Language Model

The original Mask R-CNN [HGDG17] model segments all the object instances in the image, while the multi-instance referring image segmentation task aims at a subset of the output from the conventional instance segmentation model. Therefore, it is straightforward to inject the language information into the pipeline of Mask R-CNN to filter out the incorrect instances that are not specified by the language expression.

As shown in Fig. 2, we process the input referring expression with a language model, in which a recurrent neural network built with LSTM cells is employed to extract the language features. The encoded language features are fused with both the image features in the RPN (referred to as *RPN Fusion Step*) and the RoI features after the RoIAlign operation (referred to as *RoI Fusion Step*). The two-step fusion enables the filtering in a *coarse-to-fine* manner.

In the *RPN Fusion Step*, anchors are classified into positive and negative ones, which serves as the first round of region filtering by using the positive proposals for the following processes. Different from the original Mask R-CNN that distinguishes foreground and background regions in this step, we incorporate the language information to allow a *rough* filtering to reject the regions irrelevant to the expression. In the *RoI Fusion Step*, the classification head determines the categories of the RoIs, which forms the second round of filtering by discarding the RoIs with non-object classes. Besides a naive classification as in Mask R-CNN, our model is also required to assign non-object labels to the RoIs mismatched with the language expression even if they contain objects inside (*e.g*., the rightmost tree in Fig. 2). Thus, the language features facilitate a *finer*

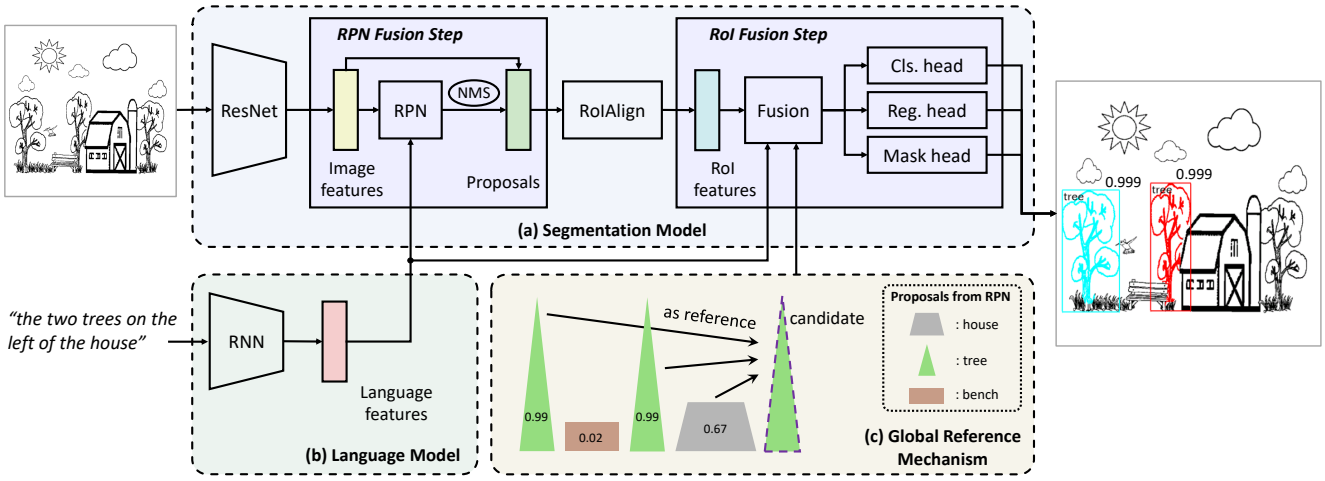

Figure 2: *Our framework GRM-Net for multi-instance referring image segmentation of scene sketches, consisting of a segmentation model, a language model and a global reference mechanism module. The encoded language features (b) and the global reference information produced in (c) are fused to the segmentation pipeline (a). In (c), the candidate uses proposals from the RPN as global references, and numbers in the proposals indicate the probability of being positive. The inputs to the GRM (c) are all the proposals output from the RPN.*

selection between the matched and mismatched candidates. Essentially, in both steps, the segmentation model incorporated with the language guidance learns to filter out the undesired proposals/RoIs by decreasing their probability of being a positive proposal or an object-class RoI.

2.3. Global Reference Mechanism

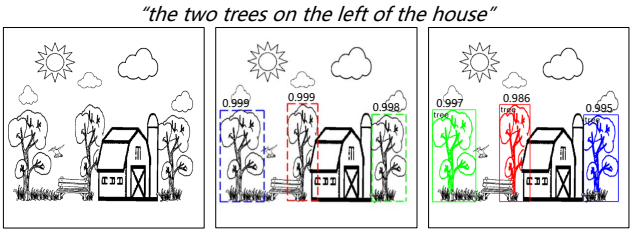

Figure 3: *Example of incorrect segmentation without global reference mechanism (GRM). The numbers above the boxes indicate the probability of being a positive proposal (b) or the predicted object category (c). We omit other proposals in (b) for brevity.*

In the *RoI Fusion Step*, each RoI is injected with the language features to determine whether it is matched to the expression. However, the RoIs that primarily store regional features have a local view and lack global information, *i.e.*, the relative arrangement of all objects and their own positions in the image. Consequently, they are not able to respond to the location information in the expression and the finer filtering may fail. For example, in Fig. 3, after the rough filtering in RPN, the rightmost tree is still kept as a high-probability positive proposal. In the absence of reference information from a global perspective, this RoI is finally classified into a tree object and output as a segmented instance, which indicates that the incorrect RoIs cannot be filtered out as non-objects according to the expression. To this end, we introduce a *global reference mechanism (GRM)* to offer each RoI a global view of its spatial relationship with other objects. As shown in Fig. 2, the global reference information for each RoI is then fused with its own visual features and the language features, which enables a joint modeling of the location information from both expression and each RoI to measure their alignment. This mechanism facilitates the finer filtering of mismatched RoIs and improves the segmentation performance.

The key idea of the global reference mechanism is to assign references to each RoI. To this end, we use the proposals output from the RPN (after the Non-Maximum Suppression (NMS) filtering operation) as the input to the GRM module, and select the references from the proposals. Each of them is with a probability of being a positive proposal, and thus we choose top-K positive ones as the references. As illustrated in Fig. 2-(c), when the rightmost tree serves as a candidate instance, top-K positive proposals ($K = 3$ proposals are visualized in this example for clarity) including the house and the other two trees are utilized as the references to establish its spatial relationship and identify its position. Different from MAttNet [YLS*18] which uses a number of K surrounding objects with the shortest distance as references (*distance-based*), we choose the references according to the aforementioned probability that indicates the degree of consistency (*consistency-based*) to the input expression, and thus make it possible to access necessary instances far away. For example, when determining whether the rightmost tree in Fig. 2-(c) matches the input language expression, all the trees should be considered jointly. Thus, this candidate has a close relationship with the tree on the far left, and our mechanism allows the latter with a high consistency probability to be accessed by the candidate regardless of the long distance. This is why we call it a *global* reference.

Specifically, for each candidate RoI C_i, we first find its top-K

references $R_j(j=1,...,K)$. Then, we calculate the offsets and area ratio between the candidate and the references as the reference information (similar to MAttNet [YLS*18]), which is formulated as:

$$\delta_{ij}^{(5)} = [\frac{[\Delta x_{tl}]_{ij}}{w_i}, \frac{[\Delta y_{tl}]_{ij}}{h_i}, \frac{[\Delta x_{br}]_{ij}}{w_i}, \frac{[\Delta y_{br}]_{ij}}{h_i}, \frac{w_j h_j}{w_i h_i}], \quad (1)$$

where $(\Delta x_{tl}, \Delta y_{tl})$ and $(\Delta x_{br}, \Delta y_{br})$ denote the offsets of the top-left and bottom-right positions, respectively. w and h are the width and height of the bounding box of the RoIs. All the 5-d vectors $\delta_{ij}^{(5)}$ of $R_j(j=1,...,K)$ are tiled to spatial feature maps and concatenated with the language features as well as the visual features of the corresponding RoI C_i. To incorporate more positional information into the candidate RoIs, we also exploit the absolute position information $[\frac{x_{tl}}{W}, \frac{y_{tl}}{H}, \frac{x_{br}}{W}, \frac{y_{br}}{H}, \frac{hw}{HW}]$ of each individual RoI and concatenate it with the hybrid features above. W and H denote the width and height of the entire scene sketch image, respectively.

3. Experiments

3.1. Datasets and Implementation Details

Datasets. We evaluate our approach on a scene sketch dataset called SketchyScene-Matching [ZMG*19] that has a wide variety of subsets of target instances, *e.g.*, a single tree, two of the trees, or all the trees in Fig. 3. This helps to better evaluate the performance of models on handling the expression diversity in real-life scenarios. Specifically, the SketchyScene-Matching dataset contains 1,695 scene sketches from the SketchyScene dataset [ZYD*18] and 38,557 pairs of object instance segmentation mask and language expression. These data are split into three sets for training, validation and testing, respectively. 24 object categories are included in this dataset.

Implementation Details. In the fusion of the language features and the segmentation pipeline (Fig. 2-(a)), we first tile the language feature vector to a spatial feature map, and concatenate it with the image features from the ResNet backbone and the RoI features after the RoIAlign operation. In the global reference mechanism (GRM) module, the K in top-K is set to 10. During training, the regression loss in the RPN stage is calculated with positive anchors, and the regression loss and mask loss for detected RoIs are applied to positive ones. We train for 270k iterations with a batch size of 4. AdamW [LH17] is adopted as the optimizer, with an initial learning rate 0.0002. In terms of the language model, we employ bi-directional LSTM with a maximum time step of 15 to process the input expression.

Quantitative Evaluation Metrics. Our approach outputs a category label, a bounding-box and a segmentation mask for each predicted instance, so we follow Zou *et al.* [ZMG*19] and Mask R-CNN [HGDG17], and use the standard COCO metrics [LMB*14] including mask AP (Average Precision), AP_{50} and AP_{75} to quantitatively evaluate the performance.

3.2. Comparison with Existing Approaches

3.2.1. Evaluation Settings

We mainly compare with the only approach designed for multi-instance referring image segmentation of scene sketches proposed by Zou *et al.* [ZMG*19]. The evaluation is conducted on the test set of the SketchyScene-Matching dataset [ZMG*19], and we directly report the quantitative results of Zou *et al.* from their paper.

Table 1: *Quantitative comparisons with the baseline methods.*

Task	Method	AP	AP_{50}	AP_{75}
Standard	DMN [MTPBA18]	4.01	9.02	1.15
Standard	CMSA [YRLW19]	8.77	28.42	3.53
Standard	CMPC [HHL*20]	10.20	28.08	5.99
Standard	CEFNet [FHZL21]	26.72	55.28	23.07
Multi-instance	Zou *et al.* [ZMG*19]	45.97	70.79	53.90
Multi-instance	GRM-Net (ours)	**59.39**	**71.37**	**62.07**

For sufficient experimental validation, we also adapt the dataset above for conventional methods developed for the standard task of referring expression segmentation, *i.e.*, binary instance segmentation in natural image domain, given the lack of baseline methods tailored for multi-instance segmentation scenario. Specifically, when there are multiple instances in a ground-truth segmentation, we merge them into a binary segmentation mask, which fits in with the conventional methods. We compare with four representative conventional methods, DMN [MTPBA18], CMSA-Net [YRLW19], CMPC [HHL*20] and CEFNet [FHZL21], and use the same hyperparameters in their official implementations. During evaluation, we treat the output mask of these approaches as the segmentation mask for a single instance, and calculate the mask AP likewise as the quantitative evaluation.

3.2.2. Results

From Table 1, we can see that the conventional methods designed for the standard task perform poorly. This is because when applied to the examples with multiple targets, they output a unified mask for all targets, which tends to lead to poor instance segmentation measurement (AP). As for the comparison with the only approach (two-stage) closest to our framework (one-stage), namely Zou *et al.* [ZMG*19], ours outperforms by a large margin, as shown in Table 1. From the qualitative results in Fig. 4, we can see that Zou *et al.* [ZMG*19] fail in some cases where the expression specifies all the instances of the same category (a). In some cases where a single target is specified (b), it tends to output undesired instances. Moreover, when it is able to locate the correct instances, some segmentation masks are not able to cover the whole object completely (c). In contrast, our model works better on these cases and produces visually more complete masks. The quantitative and qualitative results both imply that our one-stage framework is more effective because of the joint training in an end-to-end manner. More results are shown in the supplemental materials.

3.3. Effectiveness of Global Reference Mechanism

3.3.1. Ablation Study

In the global reference mechanism module, we choose top-K proposals in terms of consistency probability for each RoI as references, and fuse the 5-d reference information (shown in Eq. (1)) and the absolute positional information with the RoI features. We

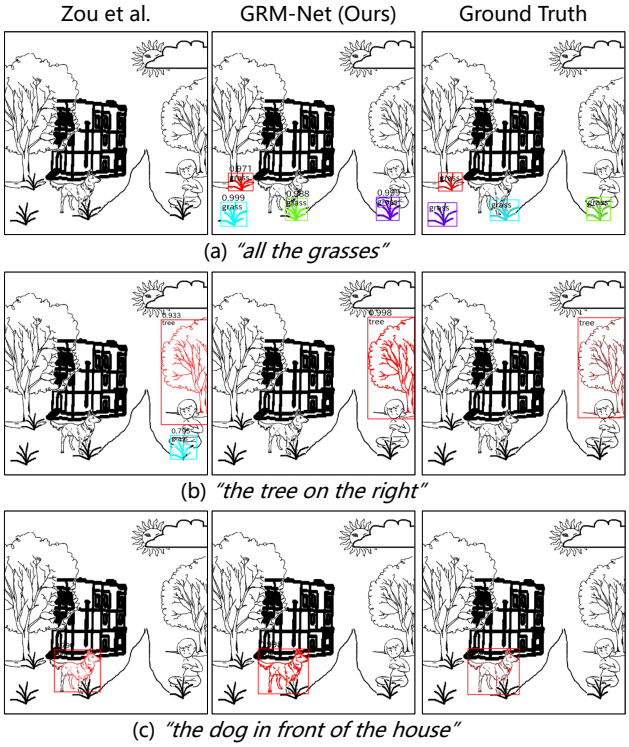

Figure 4: *Comparisons with the baseline method.*

evaluate the following design choices: (a) Different manners of selecting the references, *i.e.*, our consistency-based method or the distance-based one from the MAttNet [YLS*18]; (b) Incorporation of the absolute positional information; (c) The number of K and the performance of model without the global reference mechanism ($K = 0$); (d) Different kinds of reference information, *i.e.*, our 5-d vector or a 3-d one $\delta_{ij}^{(3)} = [\frac{[\Delta x_{center}]_{ij}}{w_i}, \frac{[\Delta y_{center}]_{ij}}{h_i}, \frac{w_j h_j}{w_i h_i}]$, where the offsets of center position $(\Delta x_{center}, \Delta y_{center})$ are used instead of the top-left and bottom-right ones.

Quantitative results in Table 2 show that the distance-based selection manner works worse than our consistency-based approach (#1 *vs.* #7), probably due to its lack of ability in accessing necessary references far away. Incorporating the absolute positional information brings slight improvement (#2 *vs.* #3), but lacking reference information from the GRM hinders further improvement (#3 *vs.* #4-#7). When employing GRM with $K = 10$, the performance improves by a large margin, but the model suffers performance degradation (#4/#5 *vs.* #7) with $K = 5$ or 20. Our model with 5-d reference information is superior to the model with 3-d one (#6 *vs.* #7).

3.3.2. Visualization and Analysis

In this section, we show how the proposed global reference mechanism (GRM) works by visualizing the intermediate results, as shown in Fig. 5. After the rough filtering in the RPN stage (b, d), model with or without GRM has positive proposals covering the correct and incorrect instances. Without the GRM (c), the model predicts the positive proposals after the RPN stage as object-class instances (*e.g.*, the rightmost tree in the top-row example), leading to incorrectly identified outputs. With the reference assignment and the global view offered by the GRM (e), the model learns to classify the incorrectly identified proposals (*e.g.*, the house and the rightmost tree in dashed boxes in the top-row example) as non-objects which have a low probability of being their real object class. Hence, these non-object instances that are mismatched with the expression are removed from the final outputs.

Table 2: *Ablation studies of configurations of the global reference mechanism (GRM). The last row shows the configurations in our model. "Manner" denotes the selection criteria of the references. "PosInfo." indicates whether absolute positional information is used. "Dim." means the dimension of reference information.*

#	Manner	PosInfo.	Top-K	Dim.	AP	AP_{50}	AP_{75}
1	**Distance**	✓	10	5	53.35	64.61	59.01
2	Consistency	×	0	-	53.29	67.71	56.52
3	Consistency	✓	**0**	-	53.55	68.35	56.96
4	Consistency	✓	**5**	5	59.03	70.41	61.38
5	Consistency	✓	**20**	5	57.32	70.23	59.79
6	Consistency	✓	10	**3**	57.46	69.97	59.71
7	Consistency	✓	10	5	**59.39**	**71.37**	**62.07**

4. Conclusion and Limitations

We present a one-stage framework for multi-instance referring image segmentation of scene sketches, which is demonstrated through our experiments to be superior to the two-stage approaches with a pre-training process. Our proposed global reference mechanism (GRM) provides a global view for the model to identify the position of each detected candidate, in order to improve the segmentation accuracy. Finally, we also show how the GRM works by visualizing the intermediate results.

Limitations. Since we use the training and test sets in the SketchyScene-Matching dataset [ZMG*19], where textual expressions are created automatically based on templates. Therefore, some of them seem unnatural and may contain grammar mistakes, for example in the bottom case in Fig. 5. When we revise these expressions into natural ones for our approach, the framework works with some but fails on the others. This is probably because our framework is built on an RNN text encoder that projects words into different feature vectors and has difficulty in recognizing the correlation between two different expressions with similar semantic contents. This also indicates the limitation of our framework in processing general expressions beyond the dataset. This issue might be overcome by the large-scale pre-training models on texts, e.g., BERT [DCLT18] and CLIP [RKH*21], which produces similar language features for two prompts with slight differences in expression. Therefore, our model trained with unnatural prompts could be tested on natural ones with similar contents while maintaining a comparable performance.

In our GRM module, we employ a simple process of information combination, and thus the framework tends to fail on some com-

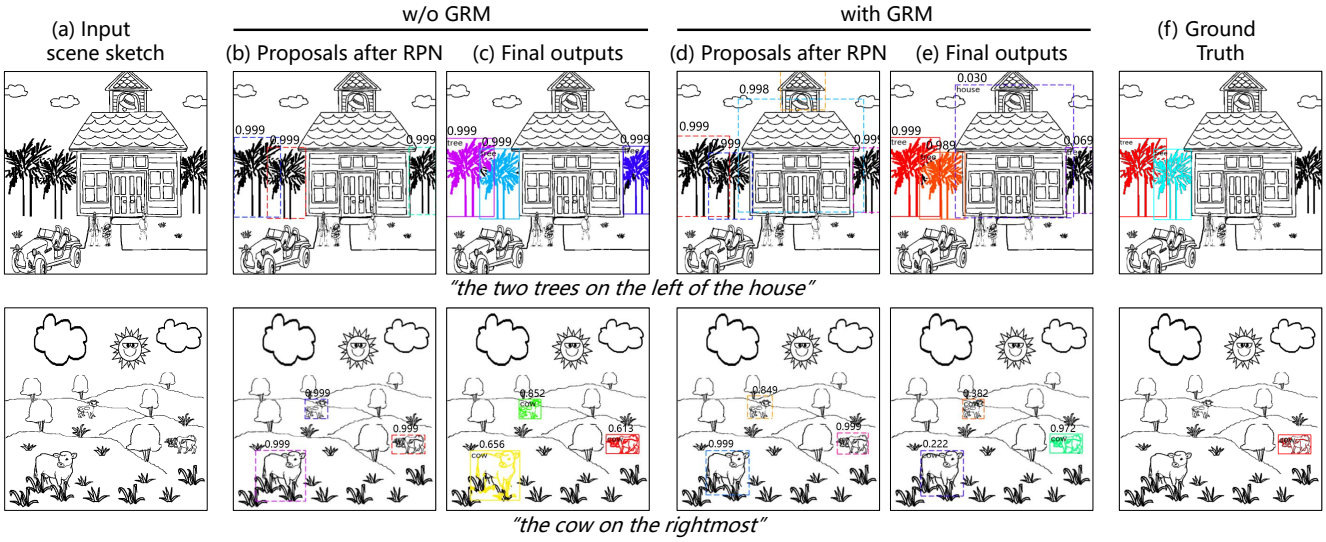

Figure 5: *Effectiveness of global reference mechanism (GRM). Boxes in dashed line in (b) and (d) are the proposals after the RPN (other proposals are omitted for brevity). Boxes in solid line in (c) and (e) are the final segmented instances. In (e), the boxes in dashed line are assigned non-object labels and thus are* ***not*** *the output instances; we visualize the predicted probability of their real object class in this case.*

plicated scenes with numerous instances, as shown in the supplemental file. Given the graph-type representation of the references as shown in Fig. 2-(c), graph neural networks (GNN), which have powerful ability of modeling the relationships in a group, could be adopted in the GRM model to strengthen reference information.

Acknowledgments

This work was supported by the Natural Science Foundation of Guangdong Province, China (Grant No. 2022A1515011425, 2019A1515011075) and the National Natural Science Foundation of China (Grant No.U1811262).

References

[DCLT18] DEVLIN J., CHANG M.-W., LEE K., TOUTANOVA K.: Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018). 5

[DLWJ21] DING H., LIU C., WANG S., JIANG X.: Vision-language transformer and query generation for referring segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 16321–16330. 1

[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, 4

[HFS*20] HU Z., FENG G., SUN J., ZHANG L., LU H.: Bi-directional relationship inferring network for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 4424–4433. 1

[HGDG17] HE K., GKIOXARI G., DOLLÁR P., GIRSHICK R.: Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision* (2017), pp. 2961–2969. 2, 4

[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. 4

[LH17] LOSHCHILOV I., HUTTER F.: Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017). 4

[LLS*17] LIU C., LIN Z., SHEN X., YANG J., LU X., YUILLE A.: Recurrent multimodal interaction for referring image segmentation. In *Proceedings of the IEEE International Conference on Computer Vision* (2017), pp. 1271–1280. 1

[LMB*14] LIN T.-Y., MAIRE M., BELONGIE S., HAYS J., PERONA P., RAMANAN D., DOLLÁR P., ZITNICK C. L.: Microsoft coco: Common objects in context. In *European conference on computer vision* (2014). 4

[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. 4

[RKH*21] RADFORD A., KIM J. W., HALLACY C., RAMESH A., GOH G., AGARWAL S., SASTRY G., ASKELL A., MISHKIN P., CLARK J., ET AL.: Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning* (2021), PMLR, pp. 8748–8763. 5

[YLS*18] YU L., LIN Z., SHEN X., YANG J., LU X., BANSAL M., BERG T. L.: Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 1307–1315. 1, 3, 4, 5

[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, 4

[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. 2, 4, 5

[ZYD*18] ZOU C., YU Q., DU R., MO H., SONG Y.-Z., XIANG T., GAO C., CHEN B., ZHANG H.: Sketchyscene: Richly-annotated scene sketches. In *Proceedings of the european conference on computer vision (ECCV)* (2018), pp. 421–436. 4