Two-stage Visual Cues Enhancement Network for Referring Image Segmentation

Yang Jiao\(^1\)\(^*\) Zequn Jie\(^2\) Weixin Luo\(^2\) Jingjing Chen\(^1\)\(^*\) Yu-Gang Jiang\(^1\) Xiaolin Wei\(^2\) Lin Ma\(^2\)\(^*\)
\(^1\)Shanghai Key Lab of Intelligent Information Processing, School of Computer Science, Fudan University
\(^2\)Meituan

ABSTRACT
Referring Image Segmentation (RIS) aims at segmenting the target object from an image referred by one given natural language expression. The diverse and flexible expressions as well as complex visual contents in the images raise the RIS model with higher demands for investigating fine-grained matching behaviors between words in expressions and objects presented in images. However, such matching behaviors are hard to be learned and captured when the visual cues of referents (i.e. referred objects) are insufficient, as the referents with weak visual cues tend to be easily confused by cluttered background at boundary or even overwhelmed by salient objects in the image. And the insufficient visual cues issue can not be handled by the cross-modal fusion mechanisms as done in previous work. In this paper, we tackle this problem from a novel perspective of enhancing the visual information for the referents by devising a Two-stage Visual cues enhancement Network (TV-Net), where a novel Retrieval and Enrichment Scheme (RES) and an Adaptive Multi-resolution feature Fusion (AMF) module are proposed. Specifically, RES retrieves the most relevant image from an external data pool with regard to both the visual and textual similarities, and then enriches the visual information of the referent with the retrieved image for better multimodal feature learning. AMF further enhances the visual detailed information by incorporating the high-resolution feature maps from lower convolution layers of the image. Through the two-stage enhancement, our proposed TV-Net enjoys better performances in learning fine-grained matching behaviors between the natural language expression and image, especially when the visual information of the referent is inadequate, thus produces better segmentation results. Extensive experiments are conducted to validate the effectiveness of the proposed method on the RIS task, with our proposed TV-Net surpassing the state-of-the-art approaches on four benchmark datasets. Our code is available at: https://github.com/SxJyJay/TV-Net.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ACM Reference Format:
Yang Jiao\(^1\)\(^*\) Zequn Jie\(^2\) Weixin Luo\(^2\) Jingjing Chen\(^1\)\(^*\) and Yu-Gang Jiang\(^1\) Xiaolin Wei\(^2\) Lin Ma\(^2\)\(^*\). 2021. Two-stage Visual Cues Enhancement Network for Referring Image Segmentation. In Proceedings of the 29th ACM International Conference on Multimedia (MM ’21), October 20–24, 2021, Virtual Event, China. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3474085.3475222

1 INTRODUCTION
Referring Image Segmentation (RIS)\(^3\) has emerged as a prominent multi-discipline research problem. Unlike traditional semantic segmentation that segments the objects of predefined classes, RIS aims to accurately segment the object specified by one natural language expression. As shown in Figure 1, given two natural language expressions describing two different tables in the same image, RIS aims to produce the corresponding segmentation results with various natural language forms, such as color, position, category, which will have a wide application prospect in the fields of interactive image editing\(^5\), human-robot interaction\(^32\), etc.

However, RIS is inherently a challenging problem, which needs to not only model the characteristics of natural language expression and image but also exploit the fine-grained interactions between these two modalities. To address such a challenge problem, existing approaches spent efforts in devising various cross-modal learning mechanisms, hoping to understand both the textual and visual semantics and further exploit their matching behaviors between words in expressions and visual information in images\(^11, 19, 21\). Relying on the attention mechanisms, such matching relationships can be captured by dynamically learning a soft association between the image local patches and words and thereby help produce more accurate segmentation results\(^3, 12, 29, 37\).

Despite remarkable progresses have been made with the existing cross-modal learning mechanisms, the segmentation results still remain unsatisfactory, especially when the visual cues of referents (i.e., referred objects) are insufficient. As shown in Figure 1, given the natural language expression “small table next to the chair”, CMPC\(^13\), the state-of-the-art method, is not able to produce a precise segmentation mask, while for “table left bottom”, CMPC
we propose a novel TV-Net for solving the RIS problem by enhancing visual cues for referents to learn the strong correlation between the retrieved images and the referent not being overwhelmed by the salient objects. Specifically, RES makes use of both the textual and visual features to alleviate the ambiguities at the referent boundaries and make the referent not being overwhelmed by the salient objects. Specifically, we propose a novel Two-stage Visual cues enhancement Network (TV-Net), performing Retrieval and Enrichment Scheme (RES) and Adaptive Multi-resolution feature Fusion (AMF).

On the one hand, RES performs the visual cues enhancement by retrieving external relevant images to enrich the visual feature of the referent. To ensure the strong correlation between the retrieved images and the referent, RES makes use of both the textual and visual features to retrieve the most relevant image for the referred object. Afterwards, the enhanced visual features are used for fusing with textual feature to obtain the multimodal features. On the other hand, AMF relies on the obtained multimodal feature to incorporate and fuse the meaningful high-resolution visual features in a dynamic and adaptive manner, which further complements the visual detailed information of the referent. Finally, the enhanced features are utilized to produce the segmentation masks.

We summarize our contributions as follows:

- **Two-stage Visual cues enhancement Network (TV-Net)** is proposed for tackling the RIS task. Comparing to the existing works which focus on devising cross-modal learning mechanisms, TV-Net deals with RIS task from one novel perspective of enhancing visual cues of the referents.
- TV-Net enhances the referent visual cues through a Retrieval and Enrichment Scheme (RES) and an Adaptive Multi-resolution feature Fusion (AMF) module. Specifically, RES incorporates the visual information semantically related to referents from an retrieved image with the original input image. AMF complements the visual details for referents by fusing the high-resolution visual features in a dynamic and adaptive manner.
- Extensive experiments on four benchmark datasets demonstrate that our proposed TV-Net achieves superior performances compared with state-of-the-art methods.

Figure 1: Referring image segmentation (RIS) aims to segment the referent (i.e. referred object), which is specified by the given natural language description. Compared with CMPC [13], a state-of-the-art method for RIS, which fails to generate precise or even correct prediction masks for unsalient referents as shown in (b), our proposed TV-Net can produce segmentation masks as shown in (c), which are highly consistent with the ground-truth in (d).

2 RELATED WORK

Referring image segmentation (RIS) aims to generate a segmentation mask for the specific objects referred in a natural language expression. Different from traditional image segmentation, the key for RIS is to establish explicit correspondence between objects in vision modality and words in language modality.

To address RIS problem, pioneering models [11, 19, 21] achieve basic multimodal interaction via concatenation-convolution scheme, and devise assortment of structures to further explore the relationships between the two modalities. As the first work to address RIS problem, Hu et al. [11] directly concatenate and fuse multimodal features from CNN and LSTM. Liu et al. [21] also follow convolution-operation fashion, and perform the cross-modal interaction in multiple steps to mimic human reasoning process. Similarly, in [19], sequentially fusion manner is also adopted to capture semantics in multi-level visual features.

Recently, richer multimodal fusion mechanism has been explored beyond simple concatenation-convolution scheme. Dynamic filters for each word are proposed in [24] to fully exploit the recursive nature of language. Later, attention mechanism is widely used in [3, 12, 13, 15, 29, 37] as its prevalence in fields of computer vision [10, 14, 36] and natural language processing [34]. In [29], attention mechanism is utilized to extract key words in language expression and model key-word-aware visual context. Further, Ye et al. [37] concatenate every word with multi-level visual features and devise self-attention mechanism to adaptively focus on informative words in the referring expression and important local patches in the input image. Bi-direction attention module is also investigated for establishing cross-modal correlation in [12]. To further explicitly align the vision and language modalities in a co-embedding space, Chen et al. [3] generate the visual-textual co-embedding map in several recurrent steps. As graph neural network [35, 42] presents a new form of mining the relationship between data, Hui et al. [15] and Yang et al. introduce graph structure models to achieve efficient message passing in RIS. Moreover, some works [13, 15] also consider the linguistic roles of each word during multimodal interaction process. Words are classified into four categories, and a progressive comprehension process is proposed under the guidance of different type of words in [13]. Hui et al. introduces dependency parsing tree [2] as prior knowledge to constrain the communication among words, so as to include valid multimodal context and exclude disturbing ones. Different from existing works, we tackle the RIS problem by enhancing visual cues for referents to learn more robust multimodal feature representation.
Figure 2: The proposed TV-Net enhances visual cues for referents in two stages. The first stage adopts a novel Retrieval and Enrichment Scheme (RES), where the most relevant image in data pool is retrieved and used for enriching visual information of referents in the input image. Relying on the enriched visual features, cross-modal interaction is performed under the guidance of text. The second stage aims to further enhance the learned multimodal feature by adaptively incorporate visual details from high-resolution visual feature map in Adaptive Multi-resolution feature Fusion (AMF) module. The solid lines represent information flow during end-to-end training process, while the dotted lines indicate information flow in the offline retrieval. The blue dotted part is a siamese-CNN, and notation "×2 up" represents 2 times bilinear interpolation upsampling.

Figure 3: The illustration of detailed process of Enrichment, which adopts a gated attentional fusion mechanism. A cross-image attention is applied to reorganize feature representations at each spatial location of retrieved image according to their correlation with referent. Then, the reorganized visual feature is selectively fused with original visual representation for enhancement.

3 METHOD
The overall architecture of our proposed Two-stage Visual cues enhancement Network (TV-Net) is illustrated in Figure 2. Given an image and its referring expression, TV-Net first performs visual cues enhancement through a novel Retrieval and Enrichment Scheme (RES), where the image most relevant to the referred object in the data pool is first retrieved with regard to textual and visual similarity. Then a gated attentional fusion mechanism is utilized to control the auxiliary information flow from the retrieved relevant images. Relying on the enriched visual feature from RES and language feature extracted from LSTM, cross-modal fusion is conducted to exploit their relationships and compose them accordingly. Afterwards, an Adaptive Multi-resolution Fusion (AMF) module provides region detail cues in high-resolution visual features complementing to the learned multimodal feature for further enhancement. Finally, the enhanced features after RES and AMF are used for predicting the segmentation masks of referred objects.

3.1 Retrieval and Enrichment Scheme
Retrieval. For a given pair of input image $I$ and language expression $T$, we retrieve an image $I_{sim}$ semantically relevant to the referred object from data pool for enriching visual information of the input image. For an accurate retrieval result, both textual and visual similarities between each candidate sample in the data pool and input are measured. Specifically, an image encoder and a text encoder are first used to generate visual and textual features for input and each candidate sample. For description convenience, we regard input as query, and candidate samples in data pool as keys. Accordingly, the extracted visual feature and text feature of input/candidate are called visual query/key and textual query/key. The visual and textual query pair is denoted as $Q = \{q^v, q^t\}$ and $i$-th visual and textual key pair is denoted as $K_i = \{k^v_i, k^t_i\}$, $i \in \{1,2,...,N\}$, where $N$ is the number of samples in data pool. Next, we match the query $Q$ and all $N$ keys $K_i$ in the data pool to select the most relevant image $I_{sim}$ for the following referring image segmentation process.

Enrichment. For the input image $I$ together with the retrieved similar image $I_{sim}$, a siamese CNN is adopted to simultaneously...
extract their visual feature \(v^\text{in}\) and \(v^\text{sim}\). Here, we define the low-resolution visual feature sets as LR = \(\{V_1, V_2, V_5\}\) containing feature maps from res3, res4 and res5 layers, and high-resolution feature sets as HR = \(\{V_1, V_2\}\) including feature maps from res1 as well as res2 layers of the siamese CNN. Semantic information exists in the low-resolution feature maps from deeper layers of CNN while the high-resolution feature maps mainly contain region details, such as edge or texture as implied in [40]. In order to incorporate visual information semantically relevant to the referred objects, the enrichment process is performed between the low-resolution visual feature sets of input image LR\(\text{sim} = \{v^\text{in}_1, v^\text{in}_4, v^\text{in}_5\}\) and the retrieved similar image LR\(\text{raw} = \{v^\text{sim}_1, V^\text{sim}_4, V^\text{sim}_5\}\).

The details of Enrichment process is depicted in Figure 3. Given visual feature \(V^\text{in}_i \in \mathbb{R}^{h_i \times w_i \times d_i}\) and \(V^\text{sim}_i \in \mathbb{R}^{h_i \times w_i \times d_i}\), where \(h_i, w_i\) and \(d_i\) are the height, width and channel dimensions respectively, we reorganize the feature representations at every spatial location in \(V^\text{sim}_i\) according to their correlation with referent in \(V^\text{in}_i\) via a cross-image attention:

\[
A_i = \text{Softmax}((W_1 v^\text{in}_i)^T (W_2 v^\text{sim}_i))
\]

\[v^\text{sim}_i' = A_i (W_3 v^\text{sim}_i) \tag{1}\]

\(v^\text{in}\) and \(v^\text{sim}\) are the reshaped \(v^\text{in}\) as well as \(v^\text{sim}\). \(W_1\), \(W_2\) and \(W_3\) are learnable parameters. The cross-image attention is calculated in the embedded Gaussian version [36] to get an attention map \(A_i\) which measures the importance of every region in \(v^\text{sim}\) to \(v^\text{in}\). Then \(A_i\) is used for reorganizing the linear transformed \(v^\text{sim}_i\) to generate visual feature \(v^\text{sim}_i' \in \mathbb{R}^{h_i \times w_i \times d_i}\), which will be served for boosting \(v^\text{in}\). Next, to enhance \(v^\text{in}\) with \(v^\text{sim}_i'\), directly adding or concatenating them might introduce noise, especially in the case of inaccurate retrieved results. Thus, we devise a gate conditioning on the original input visual feature \(v^\text{in}\) to selectively incorporate \(v^\text{sim}_i'\) into \(v^\text{in}\), which is formulated as bellow, where we omit the reshape transformation from \(v^\text{in}/v^\text{sim}\) to \(v^\text{in}/v^\text{sim}':

\[v^\prime_i = v^\text{in} + \sigma(\text{Conv}(v^\text{in})) \odot v^\text{sim}_i' \tag{2}\]

\(\sigma(-)\) represents sigmoid function. \(\text{Conv}(-)\) denotes 1×1 convolution operation, and \(\odot\) is the notation of element-wise multiplication.

### 3.2 Cross-modal Fusion

RIS model takes visual and textual features as input, and outputs a predicted segmentation mask. By now the enriched visual feature \(V^\prime_i (i \in \{3, 4, 5\})\) achieved by RES have been well-prepared. As for language feature, we encode each word in sentence \(T\) with a LSTM to generate language feature \(s \in \mathbb{R}^{d_s}\), where \(d_s\) is the dimension of language embedding space. Following prior works [12, 13, 37], an 8-D spatial coordinate feature \(O \in \mathbb{R}^{h_{O} \times w_{O} \times 8}\) is integrated into the enriched visual features \(V^\prime_i\) before cross-modal fusion.

As it is commonly assumed that the language refers to high-level visual concepts while leaving low-level visual processing unaffected [34], it is natural to aggregate language feature with high-level visual features (i.e., the low-resolution feature maps). Moreover, exploring cross-modal interaction is often formulated as a multi-step progressive process [9, 13, 25, 26]. Based on the consideration above, we first adopt Binlar fusion strategy [1] for preliminary multimodal fusion, then refine the raw multimodal feature via integrating features of multiple levels. Such process is formulated as below:

\[M^\text{raw}_i = (W_4 s) \odot (W_5 \odot \text{Conv}([V^\prime_i; O]))\]

\[M^\text{raw}_i = M^\text{raw}_i + \sum_{k \in \{3, 4, 5\} - \{1\}} \text{Gate}(M^\text{raw}_k) \tag{3}\]

where \(W_4 \in \mathbb{R}^{d_s \times d_m}\) and \(W_5 \in \mathbb{R}^{d_m \times d_m}\) are weight matrix for linear transformation, \(\odot\) and \(\text{[\ ]}\) represents the matrix multiplication and concatenation respectively. Gate(-) operation has many choices, such as LSTM-type gate in [37], vision-guided gate in [12] and language-guided gate in [13]. We use the language-guided gate as implemented in [13]. Finally, we apply ConvLSTM [31] to aggregate multimodal features \(M^\text{raw}_3\), \(M^\text{raw}_4\) and \(M^\text{raw}_5\) of different levels to obtain a refined multimodal feature \(M^\text{fine}\).

### 3.3 Adaptive Multi-resolution Feature Fusion

Different from low-resolution visual features, high-resolution visual feature accommodates more region details but are lack of semantics as implied in [40]. In order to adaptively incorporate region detail cues related to referents in high-resolution visual feature for further enhancement, we investigate a novel multi-resolution feature fusion strategy as illustrated in Figure 4. Instead of directly fusing the multimodal feature \(M^\text{fine}\) with the high-resolution visual feature \(V^\text{in}_2\), we suppress the regions irrelevant to referents in \(V^\text{in}_2\) under the guidance of \(M^\text{in}\), and then dynamically reconcile their relative intensities via a gate during aggregation process.

Suppose \(HR^\text{in} = \{V^\text{in}_2, V^\text{in}_1\}\) represents the high-resolution input image \(I\). To reduce computational complexity, we only use \(V^\text{in}_2\) for promoting the visual details. Specifically, we first encode more visual contents into \(CLR\) with convolution and ReLU activation operations, then perform Pixel Shuffle (PS) [30], a kind of upsampling operation, on \(CLR\) to obtain \(CHR\) with the same resolution as \(V_2\). Afterwards, \(CHR\) learns to suppress the regions irrelevant to referents in \(V^\text{in}_2\) during training with concatenation and convolution...
For fusing the background suppressed feature $V_2'$ with upsampled multimodal feature $M^{fine}$, simply adding or concatenating them fails to dynamically and adaptively balance their relative importance at different spatial location. Therefore, we reconcile their intensities using a gate operation, which consists of a sequence of convolution and sigmoid activation operations (i.e., black dotted rectangle parts at the middle of Figure 4):

$$
A_2 = \sigma(Conv([V_2'; M^{fine}]))
$$

$$
M^{fine'} = Conv([Upsample(M^{fine}); (A_2 \circ V_2')])
$$

As shown in Equation (5), an intensity map $A_2 \in \mathbb{R}^{H_2\times W_2\times 1}$ for $V_2'$ is first generated with a gate. Each value in $A_2$ measures the relative intensity of region detail cues in $V_2'$ at current position. Afterwards, the intensity map $A_2$ element-wisely multiplies on $V_2'$ and concatenate with upsampled $M^{fine}$ to obtain the final enhanced feature $M^{fine'}$, which will be used for the segmentation mask prediction.

4 EXPERIMENTS

4.1 Experimental Setup

Datasets. We conduct extensive experiments on four benchmark datasets for RIS: UNC [39], UNC+ [39], G-Ref [23] and ReferIt [16]. Images in UNC, UNC+ and G-Ref are all collected from MS COCO [20], while ReferIt selects images from IAPR TC-12 [7]. UNC contains 19,994 images with 142,209 language expressions for 50,000 segmentation regions. Objects of the same category may appear more than once in each image. UNC+ consists of 141,564 descriptions for 49,856 objects in 19,992 images. No location words appear in UNC+ language descriptions, which means the category and attributes are only hints for segmenting referred objects. G-Ref comprises of 104,560 expressions referring to 54,822 objects in 26,711 images. The length of description sentence is longer as well as diversification. ReferIt includes 19,894 images with 96,654 objects referred by 130,525 language expressions, and Referents in it can be objects or stuff (e.g., ground and sky).

Implementation details. Our network consists of two execution stages. The first stage aims to retrieve a similar image offline and the second stage is an end-to-end training process. Details of experiment settings are listed as follows: In retrieval process, we adopt VGG16 [33] and DistilBERT [28] with weights pretrained on ImageNet [6] and BookCorpus [41] to preprocess the image and text. We simply use cosine similarity as measurement. Note that the matching measurement is not restricted to cosine similarity, any other strategies or well-designed models for calculating the similarity can serve as substitution to obtain more precise retrieval results in the future. Specifically, we first reduce the searching space by matching all text keys $k^T_i$ in data pool with text query $q^T_j$ to select 20 most relevant candidate samples. Further, the most resemble image $I^{sim}$ among these 20 samples is picked according to visual similarity. The whole retrieval process is finished offline, which means that the retrieved similar image for every paired input image and referring expression is prepared for the subsequent end-to-end training.

During the end-to-end training, every input image is resized and zero-padded to 320x320, and fed into DeepLab ResNet-101v2 [4] pretrained on PASCAL-VOC dataset [8] following the previous work [3, 13, 37]. We extract feature maps from DeepLab ResNet-101v2 layers res3, res4, res5 as low-resolution feature set and res1, res2 as high-resolution feature set. The spatial scale of each feature map is 40x40 in low-resolution feature set, and 80x80 in high-resolution feature set. As for language processing, GloVe [27] word embeddings pretrained on Common Crawl 840B tokens are firstly used for constructing word embedding look-up table. Then a LSTM takes GloVe word embeddings as input for capturing context information. Following previous work [11, 37], We set the dimension of hidden state in each LSTM cell as 1000. Adam [17] optimizer is initialized with learning rate of 0.00025 and equipped with polynomial decay with power of 0.9. And the weight decay is set as 0.0005. Considering the GPU memory limits, we set batch size as 1. Ordinary cross-entropy loss is adopted to supervise training process. It is worth noting that we don’t use DenseCRF [18] for further refinement during inference like previous work [12, 13, 37], but still outperforms their models with DenseCRF for prediction mask refinement.

Evaluation metrics. Following previous work [3, 12, 13, 37], we also use two typical metrics for evaluation: Overall Intersection-over-Union (Overall IoU) and Prec@X, where $X \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$. However, previous works don’t discuss the intrinsic characteristics of these 2 metrics. Here we first state the formulas for Overall IoU and Prec@X, then analyze the specific model capability that each of the metric measures.

The formulas of Overall IoU and Prec@X are shown as below:

$$
\text{Overall}_\text{IoU} = \frac{\sum_{i=1}^{M} I_i}{\sum_{i=1}^{M} U_i}
$$

$$
\text{Prec}@X = \frac{\sum_{i=1}^{M} \mathbb{1}[\text{IoU}_i \geq X]}{M}
$$

where $I_i, U_i$ and $\text{IoU}_i$ represent the intersection region, union region and intersection region over union region of i-th data in test set. $M$ is the size of test set. As implied in Equation (6), Overall IoU calculates the total intersection region over union region of all samples. It mainly reflects model performance of pixel-level, which does harm to objects of small size, because small objects with limited pixels are easily dominated by the large objects with amounts of pixels. On the contrary, Prec@X mainly measures the model performance of object level, for it clearly calculates the percentage of objects in test set whose IoU higher than the threshold X.

4.2 Ablation Studies

The proposed network achieves visual boost by RES and AMF. To comprehend the effectiveness of each of them, we conduct ablation experiments on UNC+ and G-Ref dataset via quantitative as well as qualitative comparison. Due to page limitation, only ablation experiment results on UNC+ val set are shown in this paper with Table 1, and the remaining results can refer to the supplementary material. Moreover, in order to especially investigate the effectiveness on
Table 1: Ablation studies of different size of objects on UNC+ val set with Prec@X (X ∈ {0.5,0.6,0.7,0.8,0.9}) and Overall IoU for evaluation.

| row_id | Size  | Method               | Prec@0.5 | Prec@0.6 | Prec@0.7 | Prec@0.8 | Prec@0.9 | Overall IoU |
|--------|-------|----------------------|----------|----------|----------|----------|----------|-------------|
| 1      | All   | CMPC [13]            | 58.47    | 52.28    | 43.43    | 28.94    | 7.24     | 49.56       |
| 2      | Baseline | 55.19            | 48.02    | 38.84    | 25.20    | 6.25     | 48.92     |
| 3      | Baseline+RES | 55.70          | 48.51    | 39.42    | 25.14    | 6.14     | 48.89     |
| 4      | Baseline+RES+Cct | 57.69        | 51.17    | 42.61    | 29.29    | 8.20     | 49.93     |
| 5      | Baseline+RES+AMF | 58.62        | 52.71    | 44.26    | 31.89    | 10.95    | 50.30      |
| 6      | Small  | CMPC [13]            | 44.79    | 38.51    | 29.22    | 14.59    | 1.29     | 30.96       |
| 7      | Baseline | 39.97            | 31.96    | 22.85    | 11.62    | 0.82     | 29.63     |
| 8      | Baseline+RES | 41.09          | 32.83    | 23.45    | 11.19    | 0.73     | 29.86     |
| 9      | Baseline+RES+Cct | 41.74        | 34.04    | 26.03    | 15.15    | 1.51     | 29.03     |
| 10     | Baseline+RES+AMF | 46.47        | 39.50    | 30.77    | 19.45    | 3.27     | 31.54      |

Figure 5: Prediction results by our proposed module. (a) Original image. (b) Results predicted by our baseline. (c) Results predicted by Baseline+RES. (d) Results predicted by Baseline+RES+AMF (Our full model). (e) Ground-Truth.

the referents weak in visual cues, we also test on small objects, a special case of objects with weak visual cues, inside each dataset. Here, we roughly define the concept of "small" as: objects with mask accounting for less than 5% area of the whole image. For the Prec@X measure model performance of object level as stated in Section 4.1, we make analysis mainly based on Prec@X in Table 1. For fair of comparison, we only add RES while keep other settings the same as baseline, which is denoted as Baseline+RES in Table 1. By comparing row 2 and row 3, we found that RES brings improvement on Prec@0.5, Prec@0.6 and Prec@0.7, but decreases Prec@0.8 and Prec@0.9. Such results is reasonable for RES introduces external relevant visual information that makes referents more indicative, thus achieves better performance when threshold of IoU is relatively low. Nevertheless, these introduced visual information may not be distributed to referents in original image consistently at spatial domain, which hinders the model from generating a precise mask. The Prec@0.8 and Prec@0.9, who set a high threshold, only accept generated mask highly consistent with ground-truth as correct prediction, therefore these 2 metrics drop after adding RES.

We further test the effectiveness of RES on small referents in rows 7, 8 of Table 1. By comparing the relative gain on all sizes of referents (row 2 and row 3) and small referents (row 7 and row 8), we notice that improvements brought by RES on small objects (1.12% for Prec@0.5, 0.87% for Prec@0.6, 0.60% for Prec@0.7) are more obvious than all sizes of objects (0.51% for Prec@0.5, 0.49% for Prec@0.6, 0.58% for Prec@0.7), which indicates the external knowledge is more helpful for referents deficient in visual cues (i.e. small objects in our experiment).

AMF. In rows 3, 4, 5 of Table 1, we carefully control variable to verify the effectiveness of AMF. As described in Sec4.2, our baseline equipped with RES (Baseline+RES) only takes visual features from res3, res4, res5 layers of ResNet-101v2 as input, therefore, it is unfair to directly compare our final model (Baseline+RES+AMF) that also uses res2 feature with Baseline+RES model. Based on the above consideration, we add another experiment that directly concatenate the high-resolution visual feature map from res2 with learned low-resolution multimodal feature, and use concatenation followed by a 1×1 convolution layer for fusion, which is denoted as "Baseline+RES+Cct" in Table 1. As illustrated in row 4 of Table 1, although directly concatenate high-resolution feature can achieve considerable improvement, our
Table 2: Comparison with state-of-the-art (SOTA) methods on four benchmark datasets in terms of Overall IoU. "*+DCRF" means utilizing DenseCRF for refinement. Our model don’t use DenseCRF but still outperforms most of SOTA methods.

| Method             | UNC val | UNC testA | UNC testB | UNC+ val | UNC+ testA | UNC+ testB | G-Ref val | ReferIt test |
|--------------------|---------|-----------|-----------|----------|------------|------------|-----------|-------------|
| LSTM-CNN [11]      | -       | -         | -         | -        | -          | -          | 28.14     | 48.03       |
| RMI+DCRF [21]      | 45.18   | 45.69     | 45.57     | 29.86    | 30.48      | 29.50      | 34.52     | 58.73       |
| DMN [24]           | 49.78   | 54.83     | 45.13     | 38.88    | 44.22      | 32.29      | 36.76     | 52.81       |
| KWA [29]           | -       | -         | -         | -        | -          | -          | 36.92     | 59.09       |
| RRN+DCRF [19]      | 55.33   | 57.26     | 53.95     | 39.75    | 42.15      | 36.11      | 36.45     | 63.63       |
| MAttNet [38]       | 58.32   | 60.61     | 55.09     | 43.76    | 47.60      | 37.89      | 39.98     | 63.80       |
| DMN+DCRF [37]      | 55.33   | 57.26     | 53.95     | 39.75    | 42.15      | 36.11      | 36.45     | 63.63       |
| STEP [3]           | 60.04   | 63.46     | 57.97     | 48.19    | 52.33      | 40.41      | 46.40     | 64.13       |
| BRINet [12]        | 60.98   | 62.99     | 59.21     | 48.17    | 52.32      | 42.11      | 47.57     | 63.11       |
| BRINet+DCRF [12]   | 61.35   | 63.37     | 59.57     | 48.57    | 52.87      | 42.13      | 48.04     | 63.46       |
| CMPC+DCRF [13]     | 61.36   | 64.53     | 59.64     | 49.56    | 53.44      | 43.23      | 49.05     | 65.53       |
| LSCM+DCRF [15]     | 61.47   | 64.99     | 59.55     | 49.34    | 53.12      | 43.50      | 48.05     | **66.57**   |
| Ours               | 61.87   | 65.61     | 60.10     | 50.30    | 54.43      | 43.52      | **65.38** |             |

Table 3: Comparison with SOTA methods on small referents in each of the four benchmark datasets using Overall IoU as metric. Our model improves the performance on small referents significantly.

| Dataset | Setname | Methods |
|---------|---------|---------|
| CMPC    | [13]    | 43.40   |
| LSCM    | [15]    | 39.95   |
| Ours    |         | **44.69**|
| val     |         | 42.47   |
| testA   |         | 44.55   |
| testB   |         | 37.65   |
| UNC     | testA   | 30.96   |
|         | testB   | 29.30   |
| UNCP    | testA   | 34.90   |
|         | testB   | 23.74   |
| G-Ref   | val     | 29.67   |
|         | test    | 29.83   |
| ReferIt | test    | **31.14**|
|         |         | 30.35   |
|         |         | **30.77**|

4.3 Comparisons with State-of-the-art Approaches

We compare the proposed approach with several State-Of-The-Art (SOTA) method [12, 13, 37], including 2 strongest models (i.e. CMPC [13] and LSCM [15]). The comparison results are shown in Table 2. From the result, We observe that our pure model even achieve better results than SOTA methods using DenseCRF [18] for further refinement on most benchmark datasets. Remarkably, particularly obvious gain is obtained on G-Ref by our model, while the performance improvements are not as high on UNC and UNC+. One reason is that the UNC is easier as indicated by the metrics (Nearly 16% higher Overall IoU than G-Ref), and relatively early work (e.g. CSMA [37], STEP [3]) already perform well. Another reason is that small objects appear more frequently in G-Ref than in UNC+ as shown in 7, therefore, the G-Ref is better able to benefit from our model. In addition, two methods (STEP [3] and LSCM [15]) adopting high-resolution visual features should be especially noticed. STEP [3] iteratively fuses 5 levels of visual features for 25 times and LSCM [15] sequentially aggregates 4 levels of visual features through bottom-to-up and top-to-down style as in [22]. Instead of repeatedly utilizing high-resolution visual feature, we directly feed high-resolution into proposed AMF block, but still yields 3.68% and 2.03% overall IoU boost against STEP and LSCM+DCRF on G-Ref val set, which indicates the effectiveness of our design.

To further investigate the effectiveness of proposed TV-Net on the cases weak in visual cues, we compare our model with two strong SOTA models (CMPC and LSCM) on small referents within each dataset in Table 3. Besides, in order to evaluate pure model
Figure 6: Comparison of our results and two state-of-art methods (LSCM and CMPC). (a) Original image. (b) Results predicted by LSCM. (c) Results predicted by CMPC. (d) Our prediction results. (e) Ground-Truth.

Figure 7: The distribution of objects size in G-Ref val and UNC+ val, testA, testB sets. The proportion of small objects in the Gref val set is higher than that in UNC+ val, testA and testB sets. The definition for “small” is the same as in Section 4.2, and “medium” is defined as object whose mask takes 5% to 10% area of the image. The rest of cases is regarded as “big” objects.

ability, we don’t use DenseCRF [18] as further refinement in LSCM as well as CMPC. As illustrated in Table 3, our method surpasses both CMPC and LSCM with a large margin on most of the benchmarks. In general, our model achieves significant superior results on cases weak in visual cues while also gets better performance on all objects than SOTA methods.

The qualitative results of LSCM, CMPC and our model are depicted in Figure 6. From the first example in the left part of row one, we observe that CMPC unables to find the correct referred banana. Compared with CMPC, although LSCM can correctly locate target banana, it generates a prediction mask far from complete. Our prediction result is highly consistent with the ground-truth. In the example right part of row one, LSCM and CMPC ignore the lower part of the surfboard, while our model can accurately segment the whole surfboard, which proves that our model can well handle objects deficient in visual cues in RIS.

4.4 Failure Cases

We also demonstrate some interesting failure cases in Figure 8. We argue that such failure cases are mainly due to following two aspects.

First, the inaccurate language expressions mislead segmentation prediction. For example, the ground-truth mask for expression “laptop bottom” only includes the screen part of referred laptop, while our model segments the complete laptop. Second, extreme ambiguity at boundaries (e.g., splash of water around bird wings in the second row of Figure 8) deteriorates segmentation results.

5 CONCLUSION

In this paper, we dedicate on the insufficient visual cues problem and propose a Two-stage Visual cues enhancement Network (TV-Net) that first utilizes external data to enrich visual information through a novel Retrieval and Enrichment Scheme (RES), and further enhances visual details for referred objects by dynamically incorporating region detail cues in high-resolution visual features with an Adaptive Multi-resolution feature Fusion (AMF) module. Our model achieves improvements on four benchmark datasets compared with the state-of-the-art models. And such improvements are more evident in small objects, a special case of objects with weak visual cues, which thus demonstrates effectiveness of our model.

6 ACKNOWLEDGEMENT

This work was supported by National Natural Science Foundation of Project (62072116) and Shanghai Pujiang Program (20PJ1401900).
