A Survivor in the Era of Large-Scale Pretraining: An Empirical Study of One-Stage Referring Expression Comprehension

Gen Luo, Member, IEEE, Yiyi Zhou, Member, IEEE, Jianmu Sun, Member, IEEE, Xiaoshuai Sun, Member, IEEE, and Rongrong Ji, Senior Member, IEEE

Abstract—One-stage Referring Expression Comprehension (REC) is a task that requires accurate alignment between text descriptions and visual content. In recent years, numerous efforts have been devoted to cross-modal learning for REC, while the influence of other factors in this task still lacks a systematic study. To fill this gap, we conduct an empirical study in this article. Concretely, we ablate 42 candidate designs/settings based on a common REC framework, and these candidates cover the entire process of one-stage REC from network design to model training. Afterwards, we conduct over 100 experimental trials on three REC benchmark datasets. The extensive experimental results reveal the key factors that affect REC performance in addition to multi-modal fusion, e.g., multi-scale features and data augmentation. Based on these findings, we further propose a simple yet strong model called SimREC, which achieves new state-of-the-art performance on these benchmarks. In addition to these progresses, we also find that with much less training overhead and parameters, SimREC can achieve better performance than a set of large-scale pre-trained models, e.g., UNITER and VILLA, portraying the special role of REC in existing V&L research.

Index Terms—Computer vision, object recognition.

I. INTRODUCTION

Referring expression comprehension [1], [2], [3], [4], [5], also known as visual grounding [6], [7], [8], [9], is a task of locating the target instance in an image based on the natural language expression. As an vision and language (V&L) task, REC is not limited to a fixed set of object categories and can theoretically perform any open-ended detection according to text descriptions [1]. These appealing properties enable REC to garner widespread attention from the communities of both computer vision (CV) and vision and language (V&L) [1], [2], [3], [4].

Due to the obvious advantage in efficiency, one-stage modeling has recently become the main research focus in REC [1], [2], [7], [8], [9]. Compared with the two-stage methods [3], [5], [6], [11], [12], one-stage models can omit the generation of region proposals and directly output the bounding box of the referent without complex image-text ranking, thereby improving the inference speed by an order of magnitude, as shown in Fig. 1. However, one-stage models often have worse performance than the two-stage ones [1], [2], [7]. Practitioners mainly attribute this shortcoming to the lack of enough multi-modal reasoning ability [1], [8], [9], [13]. To this end, recent endeavors put numerous efforts into the design of the multi-modal networks for one-stage REC, which have achieved notable success.

Despite these advances, the existing literature still lacks a systematic study to gain insight into one-stage REC. Differing
We build a comprehensive and easy-to-use codebase and additional is a task to locate the referent based on a natural often follow a two-stage pipeline. Concretely, two-stage and RefCOCOg We present the first systematic study for one-stage REC, We propose a strong and simple model called SimREC, which outperforms a set of REC methods and large-scale pre-trained models, e.g., VILLA-Large [25], on all benchmark datasets. We believe this result can be very enlightening for the existing V&L research, where the most tasks are dominated by the expensive large-scale models.

Overall, the contributions of this article are three-fold:

- We present the first systematic study for one-stage REC, yielding several key factors in addition to multi-modal fusion.
- We propose a strong and simple model called SimREC, which outperforms a set of REC methods and large-scale pre-trained models in both performance and efficiency.
- We build a comprehensive and easy-to-use codebase based on the content of this article, which can greatly promote the development of one-stage REC.

II. RELATED WORK

Referring Expression Comprehension (REC) [1], [2], [3], [5], [8], [22], [27], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38] is a task to locate the referent based on a natural language expression. Early works [3], [5], [27], [29], [30], [31], [33], [39] often follow a two-stage pipeline. Concretely, two-stage models first detect the salient regions of an image, and then

![Model framework and candidate designs/settings. The solid lines depict the basic structure. The dashed parts list candidate designs for the empirical study, which are used to examine the effects of different components in one-stage REC. The visual backbone and word embeddings are frozen during training.](image-url)
regard REC task as region-expression ranking problem. Despite the great success, these two-stage methods have obvious defects in model efficiency and generalization [7, 40].

To this end, one-stage REC has recently become a popular research direction [1], [2], [4], [7], [8], [13], [40], [41], [42], [43], [44]. By omitting the steps of region detection and image-text matching in multi-stage modeling, one-stage models greatly reduce the inference time, e.g., 28.6 images per second, as reported in [1]. However, one-stage models often have worse REC performance compared with the two-stage methods, like MatchNet [3], mainly due to the limited reasoning ability. In this case, recent advances are dedicated to improving the reasoning ability of one-stage REC and propose various novel multi-modal networks for one-stage REC [1], [8], [9], [13], [37], [39]. In particular, ReSC [8] proposes a multi-round reasoning module to handle long and complex expressions. LBLYL [13] proposes a novel landmark feature conversation to better align the vision and language features. TransVG [9] improves multi-modal fusion via stacking multi-layer transformer blocks [45]. More recently, researchers have formulated REC as a sequence prediction task and proposed a set of novel sequence-to-sequence frameworks [41], [46]. For example, SeqTR [41] represents the bounding box of the referent with a sequence of discrete coordinate tokens, which are predicted via a Transformer architecture. In addition to the design of the network architecture, numerous efforts are devoted into multi-task learning [2], [41], [46], [47], [48], [49], and greatly improve the upper bound of REC models. Despite the effectiveness, except the well studied multi-modal fusion, there still lacks a systematic study to examine the key components in one-stage REC.

Driven by the progresses of large-scale V&L pre-training, the new SOTA performance of REC is achieved by recent large-scale pre-training models [23], [24], [25], [26], [50]. Specifically, these models are pre-trained with millions of image-text examples and excessive parameters, thereby dominating various downstream VL tasks, e.g., VQA [20] and multi-modal retrieval [51]. In particular, common pre-training objectives of early VL pre-training are masked language modeling [24], [52] and image-text matching [23], [24]. When transferring to REC task, most of these models follow the two-stage paradigm and regard this task as a text and image region matching problem [23], [24], [25], [26]. Nevertheless, these BERT-style pre-trained models are often defective in their expensive training costs and slow inference speed. Meanwhile, their pre-training objectives are also inefficient for the adaptation to REC task. In this article, we are keen to find out whether a simple and efficient one-stage model can outperform these expensive pre-trained models.

III. THE FRAMEWORK AND ABLATION DESIGNS

To accomplish the empirical study, we first build a common one-stage REC framework, as depicted in Fig. 2. Then, we prepare a set of ablation designs and settings to examine the impact of different components.

A. Framework

Specifically, given an image \( I \) and an expression \( E \), an one-stage REC model often uses a visual backbone and a language encoder to extract the visual and text features, denoted as \( F_v \in \mathbb{R}^{h \times w \times d} \) and \( F_t \in \mathbb{R}^{l \times d} \), respectively. Here, \( h \times w \) denotes the resolution of visual feature maps, and \( l \) is the length of the expression. Here, the text features \( F_t \) are attentively pooled [53] to a global vector \( f_t \in \mathbb{R}^{d} \). Afterwards, a simple multi-modal fusion module is deployed to obtain the joint representations of two modalities, denoted as \( F_m \in \mathbb{R}^{h \times w \times d} \), which is obtained by:

\[
f_m = \sigma (f_v W_v) \odot \sigma (f_t W_t).
\]

(1)

Here, \( f_m \) and \( f_t \) are the feature vectors of \( F_m \) and \( F_t \), respectively, and \( \sigma \) denotes the ReLU [54] activation function. To keep a certain reasoning ability, we also apply an attention unit called GARAN [1] after the multi-modal fusion.

Lastly, a regression layer is deployed to predict the bounding box of the referent in the image. Specifically, for each grid \((x_g, y_g)\), SimREC predicts raw coordinates of the bounding box \( \{t_x, t_y, t_w, t_h\} \) and its corresponding confidence score \( c \). Afterwards, the final bounding box \( b = \{b_x, b_y, b_w, b_h\} \) is calculated based on the pre-defined anchor box \( (p_w, p_h) \) by

\[
\begin{align*}
    b_x &= \text{sigmoid}(t_x) + g_x, \\
    b_y &= \text{sigmoid}(t_y) + g_y, \\
    b_w &= p_w^c W_w, \\
    b_h &= p_h^c W_h.
\end{align*}
\]

(2)

Given the ground-truth box \( b' = \{b'_x, b'_y, b'_w, b'_h\} \) and the ground-truth confidence \( c' \), the loss function of SimREC is defined as:

\[
l(b_i, b'_i, c, c') = \sum_{i=1}^{h \times w \times n} l_{box}(b'_i, b_i) + l_{conf}(c'_i, c_i)
\]

(3)

Here, \( n \) is the number of pre-defined anchor boxes. \( l_{box} \) and \( l_{conf} \) denote the regression loss and the binary prediction loss of the confidence score, respectively. During the test stage, SimREC will select the bounding box with the highest confidence score as the final prediction.

Although the framework is simple, it includes the main design patterns of most existing one-stage REC models, i.e., image encoder, language encoder, fusion branch and detection head. Meanwhile, a simple framework can be used to better examine other key factors for one-stage REC in addition to multi-modal fusion. With this simple framework, we can better ablate the key factors in one-stage REC.

B. Ablation Designs

Visual Backbone: We use DarkNet [55] as the default visual backbone. VGG [56] and the recently proposed CSPDarkNet [57] are the candidate choices. Meanwhile, we also test the effects of different image resolutions.

Language Encoder: The default language model is an LSTM network [58] with GLOVE embeddings [59]. During experiments, we will enhance the ability of language encoder by adding advance designs like self-attention [45] and BERT embedding [52].

Multi-scale Features and Detection: We also experiment the model with multi-scale features and multi-scale detection setting. In terms of multi-scale features, we use the last \( k \) scale
feature maps of the visual backbone as our visual features, denoted as \( \{ \mathbf{F}_{vi} \}_{i=1}^{k} \). The features of each scale are fused with the text ones by the multi-scale fusion scheme proposed in [2], denoted as \( \{ \mathbf{F}_{mi} \}_{i=1}^{k} \). Only the last scale feature maps are used, i.e., \( \mathbf{F}_{mk} \). In terms of multi-scale detection, each scale feature in \( \{ \mathbf{F}_{mi} \}_{i=1}^{k} \) will use a corresponding detection head to predict bounding boxes.

**Detection Paradigms:** In addition to the anchor-based detection head [55], we also experiment the recently popular anchor-free detection [60]. In anchor-free detection, the model directly predict the width and height of bounding boxes for each grid. The objectives we test include IoU loss [61], GIoU loss [61], Smooth-L1 loss [62] and Focal loss [63].

**Data Augmentation:** We adopt additional data, i.e., Visual Genome [19], and image augmentation methods [15], [16], [17], [18], [64] for data augmentation. In terms of image augmentation, we try ElasticTransform, CutMix [17], Random Resize, RandAugment [16] and MixUp [64]. Their detailed modification for REC are described in our appendix.

**Optimizations:** To regularize and accelerate the model training, we also examine some training techniques, e.g., label smoothing [65], exponential moving average (EMA) [66] and cosine decay learning schedule [65].

### C. SimREC

After the extensive ablations, we select the most effective designs for each REC component to form our strong base model, denoted as SimREC. The detailed configurations are listed in Table I.

### IV. Experiments

#### A. Datasets and Metric

RefCOCO & RefCOCO+ [10] contain 142 k referring expressions for 50 k bounding boxes in 19 k images from MS-COCO [51]. There are four splits in RefCOCO and RefCOCO+, i.e., train, val, testA and testB, with a number of 12 k, 10 k, 5 k, 5 k images, respectively. Test A contains more people instances, while Test B has more object ones. The expressions of RefCOCO are mainly about absolute spatial relationships, while the ones of RefCOCO+ are about attributes and relative relations. RefCOCOCg [14], [73] has 104 k expressions for 54 k objects from 26 k images. It has two different partitions, i.e., Google split [73] and UMD split [14]. Google split only contains training set and validation set, and images of which are overlapped. Instead, UMD split avoids overlapping partitions, and it contains three splits, i.e., train, validation and test. In particular, the expressions of RefCOCOCg are longer and more complex than those of RefCOCO and RefCOCO+. Visual Genome [19] contains 5.6 M referring expressions for 101 k Images. Compared to RefCOCO, RefCOCOCg+ and RefCOCOCg, the expressions of Visual Genome are more diverse but their annotations are relatively noisy.

**Intersection-over-Union (IoU)** is the metric used in REC, which measures the overlap degree between the prediction and the ground-truth. Following previous works [1], [2], [7], [8], we use IoU@0.5 to measure prediction accuracy.

#### B. Experimental Settings

**Network Configurations:** The visual backbone and the language encoder of SimREC are a CSPDarknet [57] and an LSTM [58] with a self-attention layer [45], respectively. For SimREC, we use three-scale visual features from the visual backbone outputs of stride = 8, stride = 16 and stride = 32, of which dimensions are 256, 512, 1024, respectively. For language encoder, its dimension is set to 512 and the word embedding is a 300-d matrix initialized by pre-trained GLOVE vectors [59]. By default, the input image is resized to 416 × 416, and the maximum length of input expressions is set to 15 for RefCOCO and RefCOCO+, and 20 for RefCOCOCg. For detection head and loss function, we apply the anchor-free head [60] with IoU loss [61] and BCE loss. For data augmentations, random resize is used in the training. To improve the training efficiency, we further apply the exponential moving average (EMA) [66] and cosine decay learning schedule.

**Training settings:** We use Adam [74] to train our models for 40 epochs, which can be reduced to 25 epochs when EMA is applied. The learning rate is initialized with 1e-4 and will be decayed by 10 at the 35-th, 37-th and 39-th epoch, respectively. We also try the cosine decay as the learning rate schedule. All visual backbones are pre-trained on MS-COCO [51], while the images appeared in the val and test sets of three REC datasets are removed. We also apply the region annotations in the Visual Genome [75] to pre-train SimREC. Particularly, pre-trained stage takes 15 epochs with a batch size of 256. The learning rate and optimization are kept the default settings.

#### C. Experimental Results

1) **Ablation Study:** We first conduct extensive experiments on benchmark datasets to ablate the candidate designs/settings described in Section IV-C1, of which results are shown in Table I.

**What matters in one-stage REC?** Form Table I, we first observe that the visual components are critical for one-stage REC. The use of a better backbone or detection head can lead to significant performance gains. In particular, Swin-Transformer [67] provides the most significant gains, i.e., +9.6% on RefCOCO+. Nevertheless, such a large visual backbone also slows down the inference speed of the one-stage REC model. In contrast, anchor-free head not only boosts the detection performance but also maintains the inference efficiency. Meanwhile, larger image resolutions are also more beneficial for performance. These results are within the scope of existing CV and V&L studies [55], [62], [76], [77], [78].

Multi-scale visual features are the factor that can boost performance. On RefCOCO+, its performance gain can reach up to +14.3%, while the ones on the other two datasets are also about +6%. However, when combining it with multi-scale detection, the performance is not further improved. This result suggests that multi-scale features are basically used to enhance the descriptive power of visual backbones rather than multi-scale detection,
which is obviously different to the use in object detection [55], [62].

Data augmentation is another factor that greatly affects performance. Using Visual Genome (VG) as the additional training data can greatly boost performance on all benchmarks. Meanwhile, a simple image augmentation like Resizing also obtain obvious performance gains. These results suggest that one-stage REC has a great demand in training data, especially compared with other V&L tasks [20], [21]. For example, VG is often used as additional data in VQA, while the improvement is often marginal [53], [79]. From Table I, we can also summarize a prerequisite for data augmentation, which is that the semantics of image-text pairs should not be damaged. The operations like Horizontal Flip and Random Crop will reduce the completeness or the correctness of REC examples, thus leading to counterproductive results. Meanwhile, strong augmentation methods are inferior than the simple ones, e.g., RandomErasing and CutMix. When combing all positive augmentation methods, the REC performance encounters a significant decline, as shown in Table II, which is opposite to that in object detection [78]. These results greatly differ REC from object detection in terms of data augmentation.

In Table III, we conduct cumulative ablations to validate the combination of SimREC settings. From the results, we observe

### Table I
Ablation Studies of Different Factors on Three REC Datasets

| Factors | Choices | RefCOCO val | RefCOCO+ val | RefCOCOg val | Inference Speed | Training Time |
|---------|---------|-------------|--------------|--------------|----------------|---------------|
| Visual Backbone | DN531 | 70.63 | +0.00 | 50.39 | +1.00 | 58.78 | +0.00 | 59.2 | -0.00 | 6.6 h | +0.0h |
| | VGG16 | 70.12 | -0.51 | 49.81 | -0.58 | 58.22 | -0.56 | 46.9 | -12.3 | 9.9 h | +3.3h |
| | CDN531 | 71.90 | +1.17 | 57.24 | +6.85 | 57.58 | -1.20 | 74.1 | +14.9 | 5.4 h | -1.2h |
| | ViT-B | 72.60 | 1.97 | 58.12 | +7.73 | 60.19 | -4.41 | 30.6 | -28.6 | 10.1h | +3.5h |
| | Swin-B | 73.10 | 6.47 | 59.96 | +9.57 | 62.03 | +3.25 | 36.2 | -23.0 | 9.7 h | +3.1h |
| Language Encoder | LSTM | 70.63 | +0.00 | 50.39 | +0.00 | 58.78 | +0.00 | 59.2 | -0.00 | 6.6 h | +0.0h |
| | LSTM+GLOVE | 70.60 | -0.57 | 49.54 | -0.85 | 57.78 | -1.00 | 59.2 | 0.0 | 6.6 h | +0.0h |
| | LSTM+GLOVE+SA(1) | 71.56 | +0.93 | 51.26 | +0.87 | 58.11 | -0.67 | 58.1 | -1.1 | 6.7 h | +0.1h |
| | LSTM+GLOVE+SA(2) | 71.17 | +0.54 | 50.65 | +0.26 | 58.42 | -0.36 | 57.1 | -2.1 | 6.7 h | +0.1h |
| | LSTM+GLOVE+SA(3) | 71.43 | +0.80 | 48.73 | -1.66 | 59.15 | +0.37 | 55.9 | -3.3 | 6.7 h | +0.1h |
| | LSTM+BERT | 70.66 | +0.03 | 50.27 | -0.12 | 60.31 | +1.53 | 42.0 | -17.2 | 7.8 h | +1.2h |
| | LSTM+RoBERTa | 70.84 | +0.21 | 50.44 | -0.05 | 60.56 | +1.78 | 41.3 | -17.9 | 7.9 h | +1.3h |
| | LSTM+ALBERT | 71.44 | +0.81 | 50.76 | +0.37 | 60.44 | +1.66 | 51.1 | -8.1 | 7.4 h | +0.6h |
| Multi-scale Features | One-scale Feature† | 70.63 | +0.00 | 50.39 | +0.00 | 58.78 | +0.00 | 59.2 | +0.0 | 6.6 h | +0.0h |
| | Two-scale Features | 76.79 | +6.16 | 63.71 | +13.32 | 62.62 | +3.84 | 54.1 | -5.1 | 7.9 h | +1.3h |
| | Three-scale Features† | 77.41 | +6.78 | 64.72 | +14.33 | 65.51 | +6.73 | 49.0 | -10.2 | 10.8 h | +4.2h |
| Multi-scale Detection | One-scale Feature+Head×1 †† | 70.63 | +0.00 | 50.39 | +0.00 | 58.78 | +0.00 | 59.2 | -0.0 | 6.6 h | +0.0h |
| | Three-scale Features+Head×1 †† | 77.40 | +6.77 | 64.72 | +14.33 | 65.51 | +6.73 | 49.0 | -10.2 | 10.8 h | +4.2h |
| Detection Head | Anchor-Based† †† | 70.63 | +0.00 | 50.39 | +0.00 | 58.78 | +0.00 | 59.2 | +0.0 | 6.6 h | +0.0h |
| | Anchor-Free† †† | 76.63 | +3.02 | 53.49 | +3.10 | 59.46 | +0.68 | 55.9 | -3.3 | 9.1 h | +2.5h |
| Loss Function | BCE+MSE† | 70.63 | +0.00 | 50.39 | +0.00 | 58.78 | +0.00 | 59.2 | +0.0 | 6.6 h | +0.0h |
| | BCE+SmoothL1 | 71.12 | +0.49 | 50.29 | -0.10 | 59.23 | +0.45 | 59.2 | +0.0 | 6.6 h | +0.1h |
| | BCE+IoU loss† | 71.39 | +0.76 | 51.68 | +1.29 | 59.54 | +0.76 | 59.2 | +0.0 | 6.6 h | +0.1h |
| | BCE+GIoU loss | 70.76 | +0.13 | 50.34 | -0.05 | 59.42 | +0.64 | 59.2 | +0.0 | 6.6 h | +0.1h |
| Reward loss† † † | 69.77 | -0.86 | 50.89 | +0.60 | 60.10 | +1.32 | 15.6 | -43.6 | 17.2 h | +10.5h |
| Input | 416×116† †† | 70.63 | +0.00 | 50.39 | +0.00 | 58.78 | +0.00 | 59.2 | +0.0 | 6.6 h | +0.0h |
| | 256×256 | 64.57 | -6.06 | 47.25 | -3.14 | 52.63 | -6.15 | 69.0 | +9.8 | 3.6 h | -3.0h |
| Resolution | 320×320 | 65.83 | +1.20 | 49.13 | -1.26 | 56.86 | -1.92 | 67.1 | -7.9 | 5.0 h | -1.6h |
| | 512×512 | 70.97 | +0.34 | 51.28 | +0.89 | 59.97 | +1.19 | 46.9 | -12.3 | 9.6 h | +3.0h |
| | 608×608 | 71.15 | +0.52 | 51.59 | +1.20 | 59.38 | -0.30 | 41.0 | -18.2 | 12.6 h | +6.0h |

† denotes the settings of the basic structure and † † denotes the settings of SimREC. "DN531", "CDN531", "Swin-B" and "ViT-B" refer to the visual backbone of DarkNet-53 [55], CSPDarkNet-53 [60], Swin-Transformer (Base) [67] and Vision Transformer (Base) [68], respectively. For all the results, we average three experimental results of different random seeds.

1. Reward loss is borrowed from Iter-Shrinking [4], which requires multiple forwards during inference.
2. Inference speed is tested on the 1080Ti (1GB).
that each factor consistently improves performance when combined with other factors. Particularly, multi-scale features and VG data augmentations still play the most critical role on performance, providing up to +12.94% gains on RefCOCOg test set. Meanwhile, the combination of other factors also brings notable performance gains. For example, language encoder, visual backbone, detection head, optimizations and random resize provide the improvements of +1.75%, +1.65%, +2.61%, +0.27% and +1.76% on RefCOCOg test set, respectively. And their combination finally improves the performance by +8.04% on RefCOCOg test set. After combining all factors, +20.98% performance gains can be observed on RefCOCOg test set. These results further confirm the effectiveness of SimREC.

What are less important in one-stage REC? Compared with the above mentioned factors, the other components have less impact on one-stage REC. Above all, the role of language encoder is not as important as that in other V&L tasks. By deploying the advanced designs like BERT [52] and ALBERT [80], the performance gains are not obvious. Even on RefCOCOg, of which expressions are long and complex, the improvement is only +1.78%. One assumption about this result is that the expressions of REC is not so difficult to understand. But more importantly, it may also suggest that REC is less affected by language priors, which is often serious in other V&L tasks [20], [21], [81]. In this case, a simple design can meet the requirement of expression comprehension.

In Table I, we evaluate the effects of different loss functions and training tricks. The loss functions are deployed on the anchor-based detector [55], while their performance gains are not obvious, especially compared with object detection [61]. This results may suggest that pre-defined anchors are not suitable for one-stage REC. We also validate the effect of the reinforcement learning approach [4], which reduces the performance and the efficiency. In terms of model optimization, all candidate methods pose positive impacts on performance although not so obvious. Particularly, we find that EMA [66] can accelerate training convergence.

How do these factors affect REC? Based on the above findings, we further explore how these key factors affect one-stage REC. Specifically, we investigate their effects on expressions of different content and lengths in Fig. 3, and the quality of predicted bounding boxes in Fig. 4.

From Fig. 3, we can first observe that multi-scale features and data augmentations can improve the model performance on all expressions, especially the ones with attribute descriptions (words). In this regard, we think that multi-scale features are to enhance the visual representations to facilitate fine-grained recognition, while data augmentation can significantly strengthen the learning of language-vision alignments. These two aspects are critical for REC. In addition, we also notice that the improvement of the expressions with spatial information (words) is small. We attribute it to the intrinsic merit of one-stage REC modeling in spatial relation modeling [1]. Meanwhile, we can also find that with the improvement in visual modeling, the model also has a better ability to process the long and complex expressions, especially with the help of VG. This result also provides a useful hint for the study of multi-modal reasoning.

Compared with these two factors, the benefit of detection head is more balanced. For this end, we understand that detection head mainly affects the model’s detection ability. Besides, Fig. 3 also reflects the inference of language encoder, which is relatively small. Since most of the sentences in benchmark

![Fig. 3. Relative performance gains of five settings in SimREC on attribute descriptions (row-1), spatial descriptions (row-2), the expression length (row-3). All results are calculated on the val set.](image-url)
datasets are short, the help of language encoder is not obvious. In contrast, it also affects its performance on long sentences. This observation is consistent with the experimental results in Table I.

Fig. 4 depicts the IoU score distributions of the model’s predictions about the ground-truth bounding boxes. These results reflect the impact of the factors on the quality of model predictions. From this figure, we observe that the anchor-free detector obviously improves the detection ability. On the high-quality detection (IoU 0.9–1), the number is almost doubled by this detector. Meanwhile, we notice that the use of VG

Fig. 4. Distribution of the predicted bounding boxes on different IoU metrics. All results are calculated on the val set.

TABLE IV

| Models        | Dataset | Visual Features | Pretrain Images | #Params | mAP | mAP 0.5 | mAP 0.75 | mAP 0.9 | mAP 0.95 | mAP 0.99 | mAP 0.999 | mAP 0.9999 |
|---------------|---------|----------------|----------------|---------|-----|---------|---------|--------|---------|---------|----------|----------|
| Ours (ours)   | DSN5    | -               | -              | -       | 80.5| 71.9    | 65.8    | 58.3   | 51.2    | 48.7    | 46.3     | 44.5     |
| Ours (ours)   | DSN5    | -               | -              | -       | 80.5| 71.9    | 65.8    | 58.3   | 51.2    | 48.7    | 46.3     | 44.5     |

Additional Data Sets: MDETR [50], UniTAB [69], SeqTr [41], SimRec [9], and SimRec [9].

1. Reference speed is tested on the 1088TI (16GB).
2. The calculation of network parameters excludes the visual backbone.
helps the model improve the detection accuracy (IoU > 0.5), but its performance on high-quality detection declines obviously. This is mainly attributed to the less accurate bounding box annotations in VG [19].

2) Comparison With the State-of-the-Art Methods: After making trade-offs between performance and efficiency, we further combine some findings from Table I to strengthen our baseline network SimREC. The cumulative ablation results are given in Table III. Then, we compare SimREC to a set of state-of-the-art (SOTA) methods in REC.

Comparison to one-stage and two-stage SOTAs: We first compare SimREC with SOTA methods specifically designed for REC, of which results are given in Tables IV and VI. For a fair comparison, we remove VG augmentations when compared to these models. Other training details, such as data augmentation, are also remain comparable with newly proposed REC models [41, 44]. From this table, we can see that SimREC outperforms all these methods, e.g., +1.51%, +2.47% and +4.01% on three datasets, respectively. Besides, SimREC also has superior efficiency. With much better performance, SimREC further enhances the speed advantage against two-stage methods, e.g., +21.4 times. Compared with one-stage SOTAs, SimREC is also much more lightweight and efficient, i.e., +35.5 fps.

We also compare SimREC with existing pre-trained REC models. Under the similar setups, SimREC achieves better performance than MDETR [50] and UniTAB [69], e.g., +4.02% on RefCOCO. Compared to multi-task approaches [41, 46, 47], which requires additional labeled data, SimREC also demonstrates comparable results and significant efficiency. For instance, SimREC achieves similar performance to PolyFormer-L on RefCOCO with 22 times faster inference speed. These results greatly confirm that SimREC achieves the better trade-offs between performance and efficiency than existing methods.

Comparison with large-scale BERT-style pre-trained models: We compare SimREC with a set of large-scale BERT-style pre-trained models in Table V, which includes ViLBERT [23], ERNIE-ViL [26], UNITER [24] and VILLA [25]. These methods all apply the expensive BERT-style pre-training with millions of vision-language examples [19, 51, 83, 84], and they are also fine-tuned on REC datasets. From Table V, we surprisingly observe that our simple yet lightweight model can outperform these large models on all benchmark datasets. Compared to VILLA-L [25], the performance gains of SimREC can be up to +11.22% on RefCOCO testB, and the inference time is also reduced by 23 times. Even compared to the recently proposed multi-task model, i.e., OFA [48], SimREC also has comparable performance and better efficiency. More importantly, the parameter size and training overhead of SimREC are all far less than these methods.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
Since these large models have dominated various V&L tasks [21], [51], these results actually confirm that one-stage REC can be a “survivor” in the era dominated by large BERT-style pre-trained models. From the experiments, we find that one-stage REC is an efficient way for learning vision-and-language alignment. Thus, SimREC can obtain superior performance than the pre-trained models with much fewer training examples. Meanwhile, the objective of most BERT-style pre-trained models is based on region-text matching and their structures are also modular. In this case, their performance upper-bounds are prone to be limited by the independent visual backbones like Faster RCNN [62].

3) Generalization to Transformer-Based REC Model: To validate the generalization ability of SimREC framework, we apply our findings to Transformer-based models, i.e., ViLT [82] and TransVG [9], of which results are given in Table VII. From these results, the first observation is that our findings can significantly improve the performance of ViLT [82] on REC, e.g., +8.94% on RefCOCO+ testB. Besides, we also find that the performance gains are more obvious on complex expressions and descriptions about objects. For example, the performances gains are +3.09% on RefCOCO testA, which are improved to +7.7% on RefCOCOg test. However, SimREC still outperforms these models on three datasets, greatly confirming the simple and lightweight structure of SimREC. Overall, these results greatly validate the generalization ability of our findings on Transformer-based model.

4) Qualitative Analysis: In Fig. 5, we visualize the predictions by SimREC with different settings, and also give some typical failure cases.

Fig. 5. Visualizations of predictions by SimREC with different settings and failure cases. The box of green color is the ground-truth, and the one of red is the prediction.
confirm the effectiveness of our findings, and also demonstrate the upper bound of SimREC.

V. CONCLUSION

In this article, we present an empirical study for one-stage REC, which ablates 42 candidate designs/settings via over 100 experimental trials. This study not only yields the key factors for one-stage REC in addition to multi-modal fusion, but also reflects some findings against common impressions about REC. By combing the empirical findings, we also improve the simple REC network (SimREC) by a large margin, which greatly outperforms existing REC models in both accuracy and efficiency. More importantly, SimREC have better performance than a set of large-scale pre-trained V&L models with much less training overhead and parameters. We believe that the findings of this article can provide useful references for the development of REC, and also give some inspirations for the V&L research.

REFERENCES

[1] Y. Zhou et al., “A real-time global inference network for one-stage referring expression comprehension,” IEEE Trans. Neural Netw. Learn. Syst., vol. 34, no. 1, pp. 134–143, Jan. 2023.
[2] G. Luo et al., “Multi-task collaborative network for joint referring expression comprehension and segmentation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 10034–10043.
[3] L. Yu et al., “MATTNet: Modular attention network for referring expression comprehension,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 1307–1315.
[4] M. Sun, J. Xiao, and E. G. Lim, “Iterative shrinking for referring expression grounding using deep reinforcement learning,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2021, pp. 14060–14069.
[5] P. Wang et al., “Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 1960–1968.
[6] D. Liu, H. Zhang, F. Wu, and Z.-J. Zha, “Learning to assemble neural module tree networks for visual grounding,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 4673–4682.
[7] Z. Yang et al., “A fast and accurate one-stage approach to visual grounding,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 4683–4693.
[8] Z. Yang, T. Chen, L. Wang, and J. Luo, “Improving one-stage visual grounding by recursive sub-query construction,” in Proc. 16th Eur. Conf. Comput. Vis., 2020, pp. 387–404.
[9] J. Deng, Z. Yang, T. Chen, W. Zhou, and H. Li, “TransVG: End-to-end visual grounding with transformers,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 1769–1779.
[10] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg, “Modeling context in referring expressions,” in Proc. 14th Eur. Conf. Comput. Vis., 2016, pp. 69–85.
[11] M. Bajaj, L. Wang, and L. Sigal, “GraphGround: Graph-based language grounding,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 4281–4290.
[12] X. Liu, Z. Wang, J. Shao, X. Wang, and H. Li, “Improving referring expression grounding with cross-modal attention-guided erasing,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 1950–1959.
[13] B. Huang, D. Lian, W. Luo, and S. Gao, “Look before you leap: Learning landmark features for one-stage visual grounding,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2021, pp. 16888–16897.
[14] V. K. Nagaraja, V. I. Morariu, and L. S. Davis, “Modeling context between objects from referring expression understanding,” in Proc. 14th Eur. Conf. Comput. Vis., 2016, pp. 792–807.
[15] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, “Random erasing data augmentation,” in Proc. AAAI Conf. Artif. Intell., 2020, pp. 13001–13008.
[16] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, “Randaugment: Practical automated data augmentation with a reduced search space,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops, 2020, pp. 702–705.
H. Zhao, J. T. Zhou, and Y.-S. Ong, “Word2Pix: Word to pixel cross-attention transformer in visual grounding,” IEEE Trans. Neural Netw. Learn. Syst., early access, Jun. 24, 2022, doi: 10.1109/TNNLS.2022.3183827.

J. Ye et al., “Shifting more attention to visual backbone: Query-modulated refinement networks for end-to-end visual grounding,” in Proc. IEEE/CVF Comput. Vis. Pattern Recognit., 2022, pp. 15502–15512.

A. Vaswani et al., “Attention is all you need,” Adv. Neural Inf. Process. Syst., vol. 30, pp. 5998–6008, 2017.

J. Liu et al., “PolyFormer: Referring image segmentation as sequential polygon generations,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2023, pp. 18653–18663.

B. Yan et al., “Universal instance perception as object discovery and retrieval,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2023, pp. 15325–15336.

P. Wang et al., “OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework,” in Proc. Int. Conf. Mach. Learn., 2022, pp. 23318–23340.

Z. Zhang, Z. Wei, Z. Huang, R. Niu, and P. Wang, “One for all: One-stage referring expression comprehension with dynamic reasoning,” Neurocomputing, vol. 518, pp. 523–532, 2023.

A. Kamath et al., “MDETR-modulated detection for end-to-end multi-modal understanding,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 1760–1769.

T. Lin et al., “Microsoft COCO: Common objects in context,” in Proc. 13th Eur. Conf. Comput. Vis., 2014, pp. 740–755.

J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proc. North Amer. Chapter Assoc. Comput. Linguistics-Hum. Lang. Technol., 2018, pp. 4171–4186.

Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, “Deep modular co-attention networks for visual question answering,” in Proc. IEEE/CVF Comput. Vis. Pattern Recognit., 2019, pp. 6281–6290.

V. Nair and G. E. Hinton, “Rectified linear units improve restricted Boltzmann machines,” in Proc. 27th Int. Conf. Mach. Learn., 2010, pp. 807–814.

J. Redmon and A. Farhadi, “YOLOv3: An incremental improvement,” 2018, arXiv:1804.02767.

K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in Proc. Int. Conf. Mach. Learn. Representations, 2015.

C.-Y. Wang et al., “CSPNet: A new backbone that can enhance learning capability of cnn,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops, 2020, pp. 390–391.

S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

J. Pennington, R. Socher, and C. Manning, “Glove: Global vectors for word representation,” in Proc. Conf. Empirical Methods Natural Lang. Process., 2014, pp. 1532–1543.

Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, “YOLOX: Exceeding yolo series in 2021,” 2021, arXiv:2107.08430.

H. Rezatofighi et al., “Generalized intersection over union: A metric and a loss for bounding box regression,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 658–666.

S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 2980–2988.

H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, “Mixup: Beyond empirical risk minimization,” in Proc. Int. Conf. Learn. Representations, 2018.

T. He et al., “Bag of tricks for image classification with convolutional neural networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 558–567.

A. Tarvainen and H. Valpola, “Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results,” in Proc. 31st Int. Conf. Neural Inf. Process. Syst., 2017, pp. 1195–1204.

Z. Liu et al., “Swin transformer: Hierarchical vision transformer using shifted windows,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 10012–10022.

A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” in Proc. Int. Conf. Learn. Representations, 2020.

Z. Yang et al., “UniTAB: Unifying text and box outputs for grounded vision-language modeling,” in Proc. Eur. Conf. Comput. Vis., 2022, pp. 521–539.

L. Wang, Y. Li, J. Huang, and S. Lazebnik, “Learning two-branch neural networks for image-text matching tasks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 2, pp. 394–407, Feb. 2019.

Z. Yu et al., “Rethinking diversified and discriminative proposal generation for visual grounding,” in Proc. 27th Int. Joint Conf. Artif. Intell., 2018, pp. 1114–1120.

A. Sadhu, K. Chen, and R. Nevatia, “Zero-shot grounding of objects from natural language queries,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 4694–4703.

J. Mao et al., “Generation and comprehension of unambiguous object descriptions,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 11–20.

J. C. Duchi, E. Hazan, and Y. Singer, “Adaptive subgradient methods for online learning and stochastic optimization,” J. Mach. Learn. Res., vol. 12, pp. 2121–2159, 2011.

R. Krishna et al., “Visual genome: Connecting language and vision using crowdsourced dense image annotations,” Int. J. Comput. Vis., vol. 123, no. 1, pp. 32–73, 2017.

H. Jiang, I. Misra, M. Rohrbach, E. Learned-Miller, and X. Chen, “In defense of grid features for visual question answering,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 10267–10276.

G. Luo et al., “Cascade grouped attention network for referring expression segmentation,” in Proc. 28th ACM Int. Conf. Multimedia, 2020, pp. 1274–1282.

Z. Zhang et al., “Bag of freebies for training object detection neural networks,” 2019, arXiv:1902.04103.

P. Anderson et al., “Bottom-up and top-down attention for image captioning and visual question answering,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 6077–6086.

Z. Lan et al., “ALBERT: A lite BERT for self-supervised learning of language representations,” in Proc. Int. Conf. Learn. Representations, 2019.

R. Kiros, R. Salakhutdinov, and R. S. Zemel, “Multimodal neural language models,” in Proc. Int. Conf. Mach. Learn., 2014, pp. 595–603.

W. Kim, B. Son, and I. Kim, “ViLT: Vision-and-language transformer without out-of-vocabulary or region supervision,” in Proc. Int. Conf. Mach. Learn., 2021, pp. 5583–5594.

P. Sharma, N. Ding, S. Goodman, and R. Soricut, “Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning,” in Proc. 56th Ann. Meeting Assoc. Comput. Linguistics (Volume 1: Long Papers), 2018, pp. 2556–2565.

V. Ordonez, G. Kulkarni, and T. Berg, “Im2Text: Describing images using 1 million captioned photographs,” in Proc. Annu. Conf. Neural Inf. Process. Symp., 2011, pp. 1143–1151.

Gen Luo is currently working toward the Ph.D. degree from Xiamen University, Xiamen, China. His research interests include vision and language learning.
Jiamu Sun received the bachelor’s degree from the Hebei University of Technology, Tianjin, China, in 2021. He is currently working toward the postgraduation degree with Media Analytics and Computing Lab, Xiamen University, Xiamen, China, supervised by Prof. Rongrong Ji.

Xiaoshuai Sun (Member, IEEE) received the B.S. degree in computer science from Harbin Engineering University, Harbin, China, in 2007, and the M.S. and Ph.D. degrees in computer science and technology from the Harbin Institute of Technology, Harbin, in 2009 and 2015, respectively. From 2015 to 2016, he was a Postdoctoral Research Fellow with The University of Queensland, St Lucia, QLD, Australia. From 2016 to 2018, he was a Lecturer with the Harbin Institute of Technology. He is currently an Associate Professor with Xiamen University, Xiamen, China. He was the recipient of the Microsoft Research Asia Fellowship in 2011.

Rongrong Ji (Senior Member, IEEE) is currently a Nanqiang Distinguished Professor with Xiamen University, Xiamen, China, the Deputy Director of the Office of Science and Technology, Xiamen University, and the Director of Media Analytics and Computing Lab. He has authored or coauthored more than 50 papers in ACM/IEEE Transactions, including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and International Journal of Computer Vision, and more than 100 full papers on top-tier conferences, such as Computer Vision and Pattern Recognition and Conference on Neural Information Processing Systems. His publications have got more than 10K citations in Google Scholar. His research interests include computer vision, multimedia analysis, and machine learning. He was the recipient of the National Science Foundation for Excellent Young Scholars in 2014, National Ten Thousand Plan for Young Top Talents in 2017, National Science Fundation for Distinguished Young Scholars in 2020, and Best Paper Award of ACM Multimedia 2011. He has served as Area Chairs in top-tier conferences, such as Computer Vision and Pattern Recognition and ACM Multimedia. He is also an Advisory Member for Artificial Intelligence Construction in the Electronic Information Education Committee of the National Ministry of Education.