RefDetector: A Simple Yet Effective Matching-based Method for Referring Expression Comprehension

Yabing Wang[1], Zhuotao Tian[2], Zheng Qin[1], Sanping Zhou[1], Le Wang[1]*

[1] Xi'an Jiaotong University
[2] Harbin Institute of Technology, Shenzhen

Abstract

Despite the rapid and substantial advancements in object detection, it continues to face limitations imposed by pre-defined category sets. Current methods for visual grounding primarily focus on how to better leverage the visual backbone to generate text-tailored visual features, which may require adjusting the parameters of the entire model. Besides, some early methods, *i.e.,* matching-based method, build upon and extend the functionality of existing object detectors by enabling them to localize an object based on free-form linguistic expressions, which have good application potential. However, the untapped potential of the matching-based approach has not been fully realized due to inadequate exploration. In this paper, we first analyze the limitations of the current matching-based method (*i.e.,* , mismatch problem and complicated fusion mechanisms), then present a simple yet effective matching-based method, namely RefDetector. To tackle the above issues, we devise a simple heuristic rule to generate proposals with improved referent recall. Additionally, we introduce a straightforward vision-language interaction module that eliminates the need for intricate manually-designed mechanisms. Moreover, we have explored the visual grounding based on the modern detector DETR, and achieved significant performance improvement. Extensive experiments on three REC benchmark datasets, *i.e.,* RefCOCO, RefCOCO+, and RefCOCOg validate the effectiveness of the proposed method.

Introduction

Referring expression comprehension (REC) is a multi-modal task that aims to localize the object in an image conditioning on a free-form linguistic expression. Unlike object detection, which focuses on predefined categories, REC identifies a specific object mentioned in a sentence among multiple instances. This task holds significant potential for various applications, such as cross-modal retrieval (Wang et al. 2022, 2023, 2024b,a; Dong et al. 2022; Liu et al. 2022a)

Recent advancements (Su et al. 2023; Deng et al. 2021; Li and Sigal 2021; Zhu et al. 2022) predominantly adopt the regression-based method, with a focus on generating text-tailored visual features to accurately predict the bounding box of the target object, as depicted in Figure 1 (a). Despite making great progress, these methods still encounter significant challenges. One of the main challenges is the lack of prior information about the object, necessitating the model to learn how to predict the location of the target object. Additionally, these methods often require adjusting the parameters of the entire model, leading to increased computational demands and resource requirements.

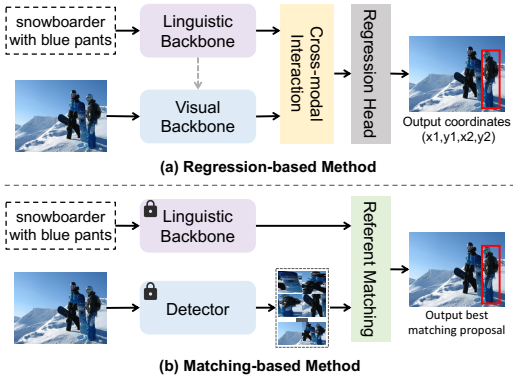


Figure 1: The comparison of REC pipelines. (a) Regression-based pipeline, which focuses on cross-modal interaction or extracts visual features guided by expression. (b) Matching-based pipeline, which utilizes proposal-expression matching to identify the most suitable proposal generated by the object detector.

Moreover, considering the remarkable progress of modern object detectors, it is a natural solution to expand their capabilities, enabling them to accurately localize the target object based on a given linguistic expression. In line with this, early research (Yu et al. 2018; Chen et al. 2021; Hong et al. 2019; Liu et al. 2019) endeavors to transform the localization task into an expression-region matching task, expanding the capabilities of the detectors, as shown in Fig.1(b). These methods effectively utilize the prior object information provided by detectors, reducing the complexity of model learning. Unfortunately, the matching-based method has not received significant attention in recent years, and its full potential remains untapped. Alternatively, in this study, we take a different approach by revisiting the matching-based pipeline, and we identify two key issues which are outlined as follows:

1) *Mismatch problem*: These methods assume that all pro-

*Corresponding author

posals contain the object and often filter out candidates by setting a high confidence threshold. This approach would potentially result in a mismatch between the proposals and the ground truth, as proposals belonging to the true target may be erroneously filtered out due to improper thresholds.

2) *Complicated fusion modules*: the fusion modules in these methods rely on specific predefined structures for language queries or image scenes and strong assumptions, involving complex manually-designed mechanisms that limit model adaptability, as shown in Figure 2 (a).

To mitigate the potential adverse effects mentioned above, we propose a simple and effective matching-based REC framework, namely RefDetector. This framework exhibits high generalization ability by leveraging a simple interaction between visual and text semantic cues, while maintaining a high referent recall [1]. Specifically, our method does not depend on the strong assumption about proposal quality. Instead, we devise a simple heuristic rule that enables more potential proposals can be included, thus significantly improving referent recall and effectively alleviating the issue of mismatches. Moreover, unlike the previous work involving manually-designed interaction mechanisms, we introduce a transformer-based vision-language interaction module (as shown in Fig. 2 (b)). As pointed out by (Deng et al. 2021), the structured fusion modules can be replaced by a simple stack of transformer encoder layers. This module naturally incorporates both visual and semantic interaction components, effectively fusing the context and expression-related semantic information. Furthermore, we also introduce the three expression-proposal alignment objectives (*i.e.,* expression-proposal contrastive learning, expression-proposal matching, and hard proposal learning) to encourage the model to adaptively identify the best-matched proposals.

Additionally, we investigate the application of the widely adopted object detector, DETR (Carion et al. 2020), for the REC task. Our empirical findings demonstrate that the learnable queries in DETR, which encapsulate detailed region content and positional information, contribute to the generation of more distinctive proposal features (see Tabel 5). Furthermore, DETR can produce a limited number of proposals and achieve higher referent recall than other two-stage detectors (*e.g.*, the referent recall can reach up to 99.96% with top-100 proposals in Fig. 4 (a)).

Extensive experiments conducted on three popular REC benchmarks (Yu et al. 2016; Mao et al. 2016) demonstrate the superior performance of our proposed method. Its practical merits are further demonstrated by successful integration with classic two-stage detectors Faster RCNN (Ren et al. 2015) and Mask RCNN (He et al. 2017), as well as query-based DETR. Notably, RefDetector achieves remarkable results with just 18 training epochs on a single V100 GPU, outperforming the previous state-of-the-art method VG-LAW (Su et al. 2023) requiring 90 training epochs on A100 GPUs. It is essential to emphasize that the core idea of this work is to propose a simple yet general matching-based framework. While designing intricate modules may potentially improve performance, it is not the primary focus of this study.

[1]"referent recall" refers to the accuracy of the proposals generated by the detector that involve the referent object (IoU>0.5).

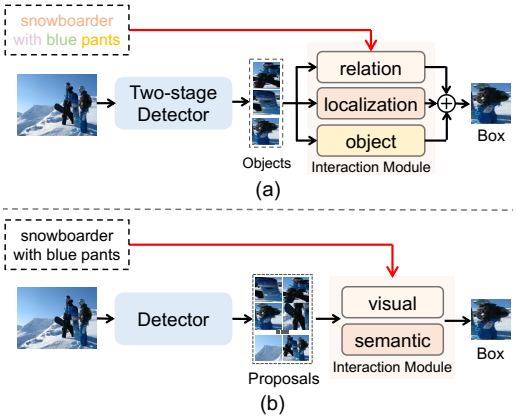


Figure 2: Comparison of matching-based frameworks: (a) Existing methods require parsing the query into predefined semantic structures and designing interaction modules accordingly. Moreover, they heavily rely on the assumption that all proposal candidates contain the object. (b) Our method introduces a simpler yet more general transformer-based visual-semantic interaction module. It is not constrained by predefined semantic structures and does not rely on the aforementioned strong assumption.

Related Work

Referring Expression Comprehension (REC)

REC aims to generate a bounding box in an image that corresponds to a given referring expression, which holds significant potential for multi-modal applications (Yang et al. 2023; Lai et al. 2023; Yang et al. 2024a,b; Zheng et al. 2023). Current research can be divided into two groups, *i.e.,* regression-based and matching-based methods.

Regression-based methods (Deng et al. 2021; Su et al. 2023; Zhu et al. 2022; Zhao, Zhou, and Ong 2022; Yang et al. 2020) aim to densely fuse linguistic and visual features, directly predicting a single or a set of coordinates for the target box. For instance, (Deng et al. 2021) establishes the multi-modal correspondence by leveraging multi-modal transformer and finally outputs 4-dim coordinates.

Matching-based methods (Rohrbach et al. 2016; Yu et al. 2017, 2018; Liu et al. 2019; Hong et al. 2019; Chen et al. 2021) typically formulate the localization task as the process of selecting the most appropriate region from a collection of proposals generated by a two-stage object detector. Previous studies (Mao et al. 2016; Wang, Li, and Lazebnik 2016; Rohrbach et al. 2016; Yu et al. 2017) employ the joint embedding model or CNN-LSTM structure to look for the target region maximizing the probability. While recent studies (Hu et al. 2017; Yu et al. 2018; Hong et al. 2019; Liu et al. 2019; Chen et al. 2021) aim to enhance the correspondence between words (or phrases) and visual regions by structural modeling. For example, MAttNet (Yu et al. 2018) introduces the modular design that decomposes expressions into three embeddings (subject, location, and relation). Although they have made remarkable progress, the complex manual-designed

mechanisms restrict the adaptability of the model, and these mechanisms require to be built on a high-quality object set. In this work, we focus on the matching-based method and propose a simple yet effective network in REC task based on the modern object detector.

Object Detection

Object detection (Girshick 2015; Ren et al. 2015; Carion et al. 2020; Dai et al. 2021; Zhang et al. 2022; Liu et al. 2022b) aims to identify and locate objects in images with their corresponding category labels. It has made remarkable progress in recent years. Early methods (Girshick 2015; Ren et al. 2015) mainly rely on hand-crafted anchors or reference points. Recently, (Carion et al. 2020) proposed to build an end-to-end transformer-based framework, which directly predicts a set of bounding boxes without post-processing. It achieved a great breakthrough in the field of object detection and inspired a series of DETR-based works (Zhang et al. 2022; Chen et al. 2023). In this work, we present a straightforward REC framework that builds upon and extends the functionality of existing object detectors by enabling them to localize an object based on free-form linguistic expressions.

Preliminary

Before formally introducing our method, we will briefly revisit the existing matching-based methods. As shown in Figure 2 (a), the matching-based methods typically have two steps: 1) Proposal Generation: use the object detector to generate a set of proposals, and 2) Referent Matching: select the best one based on the proposal-expression matching results. Details are as follows.

Proposal Generation. Existing matching-based methods (Hu et al. 2017; Yu et al. 2018; Hong et al. 2019; Liu et al. 2019; Chen et al. 2021) typically employ a pre-trained two-stage object detector (e.g., Faster RCNN) to generate a set of region proposals, which remains frozen during training. These proposals are then subjected to Non-Maximum Suppression (NMS) to eliminate redundant object regions. Subsequently, the proposals are further refined by filtering them based on their detection confidences. However, this approach may miss some referents if the detection confidence threshold is set too high to reduce overlap. Although (Chen et al. 2021) proposes a solution by introducing an additional expression-related score as the NMS criterion, this approach still requires a fixed confidence threshold.

Referent Matching. Once the region proposals $P = \{p_1, p_2, ..., p_n\}$ are obtained, they are fed into the interaction module to interact with the expressions, and select the best-matched proposal. Existing methods typically parse the expressions into predefined semantic roles, interact with region features in specific semantic modules, and then select the best-matched region based on the scores of each module. This can be formulated as the following ranking problem:

$$\begin{aligned} p^* &= \arg\max_{i \in [1,n]} \mathcal{S}(p_i, L) \\ \mathcal{S}(p_i, L) &= S(p_i|l^{subj}) + S(p_i|l^{loc}) + S(p_i|l^{rel}) \end{aligned} \tag{1}$$

where p^*, l^{subj}, l^{loc} and l^{rel} denote the best matching proposal and corresponding semantic phrase embeddings regarding subject, location, and relationship, respectively. $S(\cdot)$ is a referent matching function that evaluates the relevance between phrase embeddings and proposals.

Method

Fig. 3 shows an overview of our approach. In what follows, we will introduce our method developed with DETR, including proposal feature generation, interaction module, and referent matching, respectively. We will also demonstrate the generalization ability of our method by applying it to the classic two-stage detectors.

Proposal Feature Generation

In the following, we first introduce our proposed heuristic rule to select the proposals, and then describe the details of proposal feature generation.

Heuristic Rule. We use a top-k strategy to select proposals instead of a fixed threshold, as not all samples fit the same criteria. For DETR, we directly use the top-k proposals due to their small number. For two-stage detectors (*e.g.*, Faster RCNN), each proposal p predicts N_{class} confidences and bounding boxes in the first stage, resulting in a large number of proposal candidates. To achieve the balance between the number of proposals and referent recall (as shown in Fig. 4), we first apply the *softmax* function on the confidences along the class dimension, and select the highest confidence and the corresponding box as the final prediction for each proposal. Then, we perform NMS on these proposals to remove duplicate boxes, and finally apply a flexible top-k strategy.

Proposal Feature. Following previous matching-based methods, we extract region features from the predicted bounding boxes $B = \{\mathbf{b_1}, \mathbf{b_2}, ..., \mathbf{b_k}\}$ using RoIAlign. Furthermore, as the learnable query features $Q = \{\mathbf{q_1}, \mathbf{q_2}, ..., \mathbf{q_k}\}$ generated by DETR can converge to different object regions and capture their content and position information, we also incorporate them with the RoI features to obtain more distinctive region proposal features $F = \{\mathbf{f_1}, \mathbf{f_2}, ..., \mathbf{f_k}\}$:

$$\mathbf{a_i} = \psi_a(RoIAlign(\mathbf{C}, \mathbf{b_i})), i \in [1, k] \tag{2}$$

$$\mathbf{f_i} = \psi_f(Cat(\mathbf{a_i}, \psi_q(\mathbf{q_i}))), i \in [1, k] \tag{3}$$

where $\psi(\cdot)$ denotes the linear layer, and C represents the feature map output by the ResNet-C2 (last convolutional output of 2th-stage). Besides, for two-stage detectors that do not utilize learnable queries, we directly utilize $\{\mathbf{a_1}, \mathbf{a_2}, \ldots, \mathbf{a_k}\}$ as the proposal features.

Interaction Module

The interaction modules in previous methods are typically built under the assumption that all generated proposals consist of valid object regions. In contrast, our approach does not rely on this assumption and allows for the inclusion of a large number of potentially invalid regions, such as background or incomplete objects. Following, we introduce a straightforward interaction module that exploits the interaction between visual and text semantic cues to enhance the context and discriminative information in proposal features.

Visual Interaction. For referring expression grounding, both context information and the relationships among proposals

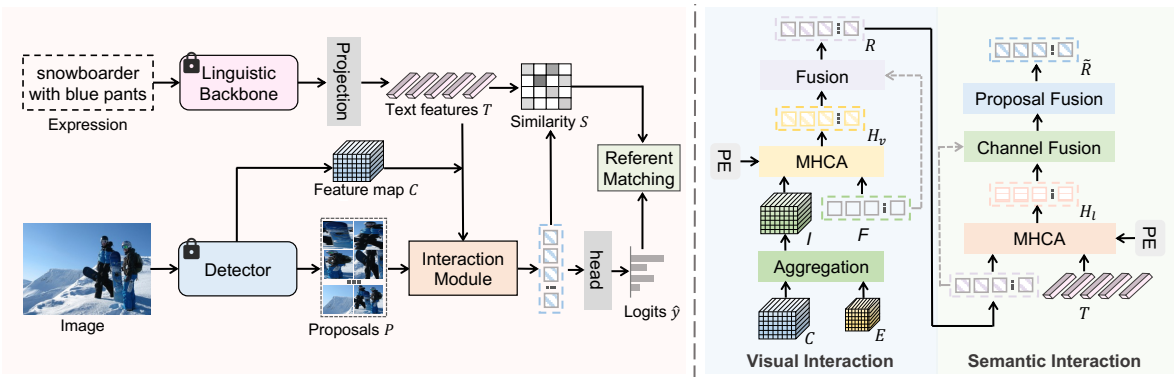


Figure 3: Illustration of the RefDetector Framework. The left part illustrates the pipeline of our method, which includes proposal generation, interaction module, and referent matching. On the right, an example of the interaction module is illustrated, which includes visual interaction and semantic interaction components. This module effectively fuses contextual and expression-related semantic information.

play a crucial role. However, the proposal features obtained from the convolutional network do not fully capture this information. This limitation arises from the limited receptive field of the convolutional network, which primarily focuses on local context information. To address this problem, we introduce a visual interaction component. Unlike existing methods that rely on self-interaction mechanisms to capture proposal relationships, we directly extract context features from the image feature map. Hence, we can avoid the adverse impacts of the noisy proposal features and leverage the rich spatial information within the image feature map, enabling proposals to capture more informative contextual features.

Specifically, we begin by integrating the output features E from the transformer encoder of DETR, which captures the global relations within the image, with the larger solution feature map C. This integration enables the model to produce an image feature map that encompasses abundant contextual information. The process can be formulated as:

$$\mathbf{I} = \psi_c(\mathbf{C} + \psi_e(\delta(\mathbf{E}))) \tag{4}$$

where $\delta(\cdot)$ denotes the interploation operation, $\psi_e(\cdot)$ and $\psi_c(\cdot)$ are the linear layer. If we want to apply it to the two-stage detector, we need to add a transformer encoder, similar to that of DETR.

Then, we treat the proposal features $\mathbf{F}$ as the query, and the image feature map $\mathbf{I}$ as the key and value. By employing the cross-attention mechanisms (MHCA), we extract context-related features for each proposal and fuse them with the proposal features. To incorporate positional awareness, we add positional encoding to the cross-attention process:

$$\mathbf{H_v} = MHCA(\mathbf{F} + PE(B), \mathbf{I}, \mathbf{I}) \tag{5}$$

$$\mathbf{R} = LN(\phi(\psi_v(\mathbf{H_v})) + \mathbf{F}) \tag{6}$$

where $\phi(\cdot)$, $\psi_v(\cdot)$ and $LN(\cdot)$ denote the gelu activation function, linear layer, and layer norm, respectively. $PE(\cdot)$ is the position encoding function, which represents the position encoding function, which maps float coordinates to a vector. Besides, there are two feasible options for position encoding:

(1) sinusoidal embedding: we first convert the float coordinate values (x, y, w, h) into sinusoidal embeddings, and then utilize the multilayer perceptron (MLP) layer to generate the position embedding:

$$PE(B) = MLP(Cat(sine(x, y, w, h))) \tag{7}$$

(2) location embedding: we directly use an MLP layer to project the 5-d vector $(x1, y1, x2, y2, w \cdot h)$ to a location embedding:

$$PE(B) = MLP([x1, y1, x2, y2, w \cdot h]) \tag{8}$$

Semantic Interaction. To enhance the discriminability of the proposal features, we interact the proposal features with the text features. Unlike previous approaches that involve parsing the sentence into predefined semantic roles and employing complicated fusion modules (as shown in Fig. 2 (a)), we introduce a simple, lightweight, and effective interaction module. Specifically, we start by extracting text features by inputting the expression L with m words into the frozen linguistic backbone $\mathcal{F}_l(\cdot)$. Next, we apply the trainable projection layer $f(\cdot)$ to obtain the transformed text embedding $\mathbf{T}$:

$$\mathbf{T} = f(\mathcal{F}_l(L)) \in \mathbb{R}^{m \times d} \tag{9}$$

After that, we treat $\mathbf{R}$ as the query and $\mathbf{T}$ as the key and value of the MHCA layer to aggregate the query-relevant semantic information, such that the aggregated semantic representations $\mathbf{H_l}$ are derived as:

$$\mathbf{H_l} = MHCA(\mathbf{R} + PE(B), \mathbf{T}, \mathbf{T}) \tag{10}$$

Finally, we integrate the proposal features with the aggregated semantic features along with the channel and proposal dimensions, respectively:

$$\hat{\mathbf{R}} = LN(\psi_l(\mathbf{H_l})) + \mathbf{R} \tag{11}$$

$$\widetilde{\mathbf{R}} = LN(\psi_r(\hat{\mathbf{R}})) + \hat{\mathbf{R}} \tag{12}$$

where $\psi_l(\cdot)$ and $\psi_r(\cdot)$ represent the linear layer.

Referent Matching

We introduce three objectives: expression-proposal contrastive learning (EPCL), expression-proposal matching (EPM), and hard proposal learning (HPL) to enhance the

alignment between expressions and their corresponding proposals.

Expression-proposal Contrastive Learning. Given the expressions and the corresponding proposal sets, our goal is to employ a contrastive learning approach to learn an optimal scoring function that ensures the scores of matched expression-proposal pairs are higher than those of unmatched pairs. Following, we regard the proposals that have high overlap (i.e., IoU $>$ 0.5) with the ground-truth box as the positives (denoted by $\widetilde{\mathbf{R}}^{+}$), and the rest as negatives:

$$\mathcal{L}_{EPCL} = -\frac{1}{N}\sum_{i=1}^{N}\frac{exp(\mathcal{S}(\bar{\mathbf{T}}_\mathbf{i},\widetilde{\mathbf{R}}_i^{+})/\tau)}{\sum_{j=1}^{k} exp(\mathcal{S}(\bar{\mathbf{T}}_\mathbf{i},\widetilde{\mathbf{R}}_\mathbf{j})/\tau)} \tag{13}$$

where $\mathcal{S}(,)$ represents cosine similarity function, $\bar{T}$ denotes the $[eos]$ token in the text representations, N and τ denote the mini-batch size and temperature parameter, respectively.

Expression-proposal Mathcing. In addition to the EPCL, we introduce a matching head to predict the matching logits $\hat{Y}$ based on the proposal features $\widetilde{R}$. We use the cross-entropy loss (CE) as the training objective:

$$\hat{Y} = \sigma(MLP(\widetilde{\mathbf{R}})) \tag{14}$$

$$\mathcal{L}_{EPM} = -\frac{1}{N}\sum_{i=1}^{N} Y_i log(\hat{Y}_i) \tag{15}$$

where $\sigma(\cdot)$ denotes the sigmoid function, and $Y \in (0, 1)$ is the binary labels, if the proposal has high overlap with the ground-truth is 1, otherwise 0.

Hard Proposal Learning. Considering the presence of numerous overlapping regions within the proposal features, we introduce a hard proposal learning objective to assist the model in effectively distinguishing between matched and hard negative proposals. It can be formulated as follows:

$$\hat{S} = Norm(S + \hat{Y}) \tag{16}$$

$$\mathcal{L}_{HPL} = -\frac{1}{N}\sum_{i=1}^{N} max(0, \hat{S}_i^{+} - \hat{S}_i^{-} + \alpha) \tag{17}$$

where α indicates a margin constant, $\hat{S}^{+}$ and $\hat{S}^{-}$ denote the mean score of the positive proposals and the hardest negative proposal, respectively. Among them, $S = \mathcal{S}(\bar{\mathbf{T}}, \widetilde{\mathbf{R}})$ denotes similarities between expression and corresponding proposals.

Finally, our objectiveness can be formulated as the combination of the above three losses:

$$\mathcal{L} = \mathcal{L}_{EPCL} + \lambda_1\mathcal{L}_{EPM} + \lambda_2\mathcal{L}_{HPL} \tag{18}$$

where λ denotes the weight of each objective.

Inference. During inference, we utilize the weighted sum of the similarities S of expression-proposals and the logits $\hat{Y}$ output by the matching head to select the best-matched proposal p^*, which can be formulated as:

$$\bar{S} = \beta \cdot S + (1 - \beta) \cdot \hat{Y} \tag{19}$$

$$p^* = \underset{i\in[1,k]}{\arg\max}\ \bar{S}_i \tag{20}$$

where β is used to adjust the weight of each matching score.

Application on Two-stage Detector

Our approach is a simple and general method that can be applied to both end-to-end object detectors and two-stage detectors. Because of the structural differences between these two detectors, we will introduce some extra designs for two-stage detectors. Specifically, since the two-stage detector does not use the learnable query and encoder-decoder transformer, we obtain proposal features by RoIAlign operation (Eq. 2). Additionally, like DETR, we also integrate a transformer encoder to generate the image representation E from the image feature map (ResNet C4) as described in Eq. 4.

Discussion

Note that our work focuses on exploring a simple and general interaction module, although more sophisticated structures might yield additional performance improvements, this is not the primary focus of our research. We leave further exploration in this direction to future studies. *Additionally, like existing matching-based methods, our work does not aim to design an object detector, but rather to explore how to apply object detectors (which are frozen during training) to REC tasks effectively.*

Experiments

Datasets and Evaluation Metric

RefCOCO/RefCOCO+/RefCOCOg. RefCOCO (Yu et al. 2016) comprises 19,994 images with 50,000 referred objects, divided into train, val, testA, and testB sets. Similarly, RefCOCO+ (Yu et al. 2016) contains 19,992 images with 49,856 referred objects. RefCOCOg (Mao et al. 2016) has 25,799 images with 49,856 referred objects and expressions, and split to train, val, and test sets.

Evaluation Metric. We use Prec@0.5 evaluation protocol to evaluate the accuracy. Given a referring expression, a predicted region is considered correct if its intersection-overunion (IoU) with the ground-truth bounding box is greater than 0.5.

Implementation Details

Following previous work (Yu et al. 2018), we initialize the object detectors with pre-trained weights, which remain fixed during training. We employ the frozen CLIP model (Radford et al. 2021) as the text encoder. For the hyperparameters, we set $\beta = 0.4$, $\lambda_1 = 1$, and $\lambda_2 = 0.5$ for DETR and $\beta = 0.4$, $\lambda_1 = 10$, and $\lambda_2 = 1$ for the two-stage detectors. The model is end-to-end optimized by AdamW (Loshchilov and Hutter 2017) with a batch size of 128, and the learning rate is set to 1e-3 for DETR and 1e-4 for the two-stage detectors.

Comparisons with State-of-the-art Methods

To evaluate the effectiveness of our proposed method, we compare it with other state-of-the-art methods on three REC benchmarks, namely RefCOCO, RefCOCO+, and RefCOCOg, in Tab. 1. Our method consistently outperforms the matching-based methods on both detectors. Notably, we achieve superior results with a smaller backbone, RN50, compared to the Ref-NMS (Chen et al. 2021). Furthermore, when

Methods	Venue	Visual Backbone	RefCOCO			RefCOCO+			RefCOCOg	
			val	testA	testB	val	testA	testB	val	test
Regression-based:										
ReSC-Large (Yang et al. 2020)	ECCV20	DN53	77.63	80.45	72.30	63.59	68.36	56.81	67.30	67.20
MCN (Luo et al. 2020)	CVPR20	DN53	80.08	82.29	74.98	67.16	72.86	57.31	66.46	66.01
RealGIN (Zhou et al. 2021)	TNNLS21	DN53	77.25	78.80	72.10	62.78	67.17	54.21	62.75	62.33
TransVG (Deng et al. 2021)	ICCV21	RN101	81.02	82.72	78.35	64.82	70.70	56.94	68.67	67.73
RefTR (Li and Sigal 2021)	NIPS21	RN101	82.23	85.59	76.57	71.58	75.96	62.16	69.41	69.40
PFOS (Sun et al. 2022)	TMM22	RN101	78.44	81.94	73.61	65.86	72.43	55.26	64.53	67.89
PLV-FPN (Liao et al. 2022)	TIP22	RN101	81.93	84.99	76.25	71.20	77.40	61.08	70.45	71.08
SeqTR (Zhu et al. 2022)	ECCV22	DN53	81.23	85.59	76.08	68.82	75.37	58.78	71.35	71.58
Word2Pix (Zhao, Zhou, and Ong 2022)	TNNLS22	RN101	81.20	84.39	78.12	69.74	76.11	61.24	70.81	71.34
QRNeT (Ye et al. 2022)	CVPR22	Swin-S	84.01	85.85	82.34	72.94	76.17	63.81	71.89	73.03
MRLN (Hua et al. 2023)	TMM23	DN53	83.46	86.46	78.59	70.67	74.80	61.93	71.59	70.00
VG-LAW (Su et al. 2023)	CVPR23	ViT-B	86.06	88.56	82.87	75.74	80.32	66.69	75.31	75.95
LGR-NET (Lu et al. 2024)	TCSVT24	ResNet-101	83.69	86.42	79.25	73.50	78.36	65.02	71.38	74.14
Matching-based:										
+FasterRCNN:										
CMN (Hu et al. 2017)	CVPR17	VGG16	-	71.03	65.77	-	54.32	47.76	-	-
VC (Zhang, Niu, and Chang 2018)	CVPR18	VGG16	-	73.33	67.44	-	58.40	53.18	-	-
ParalAttn (Zhuang et al. 2018)	CVPR18	VGG16	-	75.31	65.52	-	61.34	50.86	-	-
RvG-Tree (Hong et al. 2019)	TPAMI19	RN101	75.06	75.06	69.85	63.51	67.45	56.66	66.95	66.95
CM-A-E (Liu et al. 2019)	CVPR19	RN101	78.35	83.14	71.32	68.09	73.65	58.03	67.99	68.67
RefDetector	-	RN50	**80.65**	**84.23**	**76.54**	**68.40**	**73.81**	**60.22**	**71.39**	**70.98**
+MaskRCNN:										
MAttNet (Yu et al. 2018)	CVPR18	RN101	76.65	81.14	69.99	69.99	71.62	56.02	66.58	67.27
Ref-NMS (Chen et al. 2021)	AAAI21	RN101	80.70	84.00	76.04	68.25	73.68	59.42	70.55	70.62
RefDetector	-	RN50	**81.82**	**85.01**	**76.98**	**69.27**	**74.21**	**61.07**	**72.11**	**71.73**
+DETR:										
RefDetector	-	RN50	85.10	86.57	82.74	73.61	77.98	66.32	75.70	75.29
RefDetector	-	RN101	**87.21**	**89.02**	**83.51**	**75.87**	**80.69**	**67.89**	**77.00**	**76.43**

Table 1: Comparisons with the state-of-the-art approaches on three REC benchmarks, *i.e.,* RefCOCO (Yu et al. 2016), RefCOCO+ (Yu et al. 2016), RefCOCOg (Mao et al. 2016). We report the results of our method with various detectors. RN101, DN53, Swin-S, and ViT-B are shorthand for the ResNet101, DarkNet53, Swin-Transformer Small, and ViT-Base, respectively.

Method	Visual	Semantic	val	test
+FasterRCNN	✓		68.35	68.03
+FasterRCNN		✓	69.22	68.81
+FasterRCNN	✓	✓	**71.39**	**70.98**
+DETR	✓		72.77	71.92
+DETR		✓	72.00	71.77
+DETR	✓	✓	**77.00**	**76.43**

Table 2: Ablation study of interaction components on RefCOCOg. "Visual" and "Semantic" indicate the visual and semantic interaction components, respectively.

Method	Sine	Location	test	val
+FasterRCNN			69.95	69.37
+FasterRCNN	✓		70.22	69.83
+FasterRCNN		✓	**71.39**	**70.98**
+DETR			76.01	75.19
+DETR	✓		**77.00**	**76.43**
+DETR		✓	76.32	75.61

Table 3: Ablation study of position encoding function on RefCOCOg. "Sine" and "Location" indicate the sinusoidal embedding and location embedding, respectively.

applying our method to DETR, we observe significant performance improvements. Notably, despite VG-LAW (Su et al. 2023) adjusting the visual backbone with language-adaptive weights, our method, which freezes both the detector and linguistic backbone, surpasses VG-LAW by +1.34%, +0.52%, +1.98% on RefCOCO, +0.17%, +0.46%, +1.80% on RefCOCO+, and +2.24%, +0.43% on RefCOCOg.

Ablation Studies

In this section, we conduct ablation studies on RefCOCOg to investigate the effectiveness of our proposed method. For DETR, we use the ResNet101 version.

Analysis of the Interaction Module In Tab. 2, we observed a significant performance decline when any interaction component (visual or semantic) was removed. The results demonstrate the importance of both components in improving the accuracy of matching expressions to proposals.

Analysis of the position encoding. In Tab. 3, incorporating position encoding clearly improves performance, demonstrating its importance. Additionally, we observe that the sinusoidal embedding performs best with DETR, while location embedding is most effective with Faster RCNN. This may be because DETR uses sinusoidal positional encoding, whereas Faster RCNN, which lacks positional encoding, benefits more from the simpler location embedding.

Analysis of the referent matching. In Tab. 4, the performance is further improved when we incorporate the matching head to predict the semantic matching logits. This demonstrates that our EPM objective helps aggregate visually relevant semantic information. Furthermore, we introduce HPL to enhance the ability of the model to distinguish hard negative samples, resulting in significant performance gains.

Method	EPCL	EPM	HPL	val	test
+FasterRCNN	✓			68.27	67.90
+FasterRCNN		✓		67.03	66.74
+FasterRCNN	✓		✓	69.41	68.73
+FasterRCNN		✓	✓	67.85	67.70
+FasterRCNN	✓	✓		70.98	70.28
+FasterRCNN	✓	✓	✓	**71.39**	**70.98**
+DETR	✓			74.20	73.85
+DETR		✓		71.53	71.29
+DETR	✓		✓	75.06	74.88
+DETR		✓	✓	73.92	73.76
+DETR	✓	✓		76.11	75.42
+DETR	✓	✓	✓	**77.00**	**76.43**

Table 4: Ablation study of referent matching on RefCOCOg. "EPCL", "EPM", and "HPL" indicate expression-proposal contrastive learning, expression-proposal matching, and hard proposal learning, respectively.

Row	C1	C2	C3	C4	Query	val	test
1	✓					75.86	75.49
2		✓				76.02	75.91
3			✓			74.02	73.36
4				✓		70.59	70.83
5	✓				✓	76.78	76.11
6		✓			✓	**77.00**	**76.43**
7			✓		✓	76.28	75.62
8				✓	✓	74.85	73.90
9			✓	✓	✓	75.17	75.00
10	✓	✓			✓	76.39	76.27
11	✓	✓	✓		✓	76.33	75.67

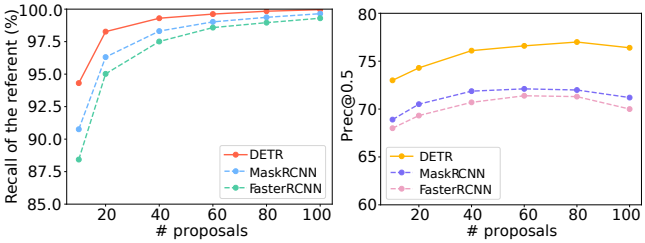
Table 5: Ablation study of the proposal features from DETR on RefCOCOg. "C_i" represents the RoI features derived from the ResNet-C_i stage, and "Query" represents the learnable query features in DERT.

Method	MHSA	MHCA	val	test
+FasterRCNN	✓		69.75	69.16
+FasterRCNN		✓	**71.39**	**70.98**
+DETR	✓		75.32	74.59
+DETR		✓	**77.00**	**76.43**

Table 6: Ablation study of visual interaction on RefCOCOg. "MHSA" represents the self-attention operation on the proposal features, "MHCA" represents the cross-attention operation on the proposal features and image feature map.

Analysis of the proposal features from DETR. Previous matching-based methods typically applied RoIAlign to the C4 stage. In Tab. 5, we conduct a detailed study on proposal feature generation. We found that using low-level features yielded better results, likely due to higher resolution and richer fine-grained information in shallow networks. Additionally, incorporating learnable queries significantly improved performance, with the best results achieved by combining learnable queries with C2 RoI features.

Analysis of the visual interaction component. In Tab. 6, performance drops significantly with MHSA due to invalid regions in the proposals, which may result in uninformative features and hinder feature interaction. However, our proposed method interacts with the image context feature map directly, thus avoiding the above issues, and being more suitable for practical application scenarios.



(a) Referent recall (b) Performance

Figure 4: The referent recall and performance of the top-k proposals, indicating that both the performance and recall metrics are low when the number of proposals is limited.

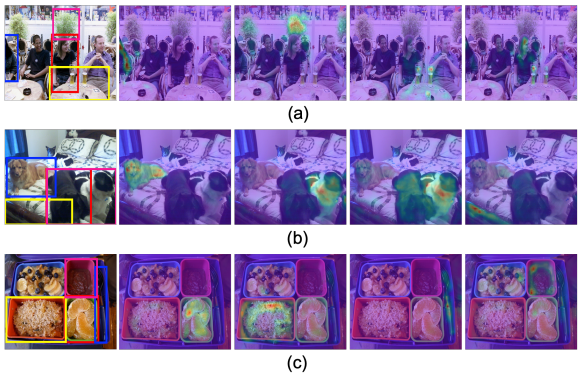


Figure 5: Visualization of the attention score between proposal features R and image feature map I in visual interaction from DETR. The four examples on the right represent the visualized results of the region proposals in the first image.

Qualitative Results

In Fig. 5, the region proposal features effectively focus on corresponding objects in the image. For instance, in case (a), the third example assigns high attention weights to objects related to the yellow box (e.g., the table and goblet). Additionally, in case (b), the irrelevant proposal focuses only on its corresponding region, preventing interference with other proposal features.

Conclusion

In this paper, we revisit the matching-based pipeline and introduce RefDetector, a simple yet effective framework addressing the mismatch problem and complex fusion in existing methods. Extensive experiments on three benchmarks highlight the superior performance and practical advantages of our method. **Limitations:** We have only applied our method to the REC task. In the future, we will explore extending our approach to open-vocabulary object detection and referring image segmentation tasks.

Acknowledgments

This work was supported in part by National Science and Technology Major Project under Grant 2023ZD0121300, National Natural Science Foundation of China under Grants 62088102, 12326608 and 62106192, Natural Science Foundation of Shaanxi Province under Grant 2022JC-41, and Fundamental Research Funds for the Central Universities under Grant XTR042021005.

References

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, 213–229. Springer.

Chen, L.; Ma, W.; Xiao, J.; Zhang, H.; and Chang, S.-F. 2021. Ref-nms: Breaking proposal bottlenecks in two-stage referring expression grounding. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 1036–1044.

Chen, Q.; Chen, X.; Wang, J.; Zhang, S.; Yao, K.; Feng, H.; Han, J.; Ding, E.; Zeng, G.; and Wang, J. 2023. Group detr: Fast detr training with group-wise one-to-many assignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6633–6642.

Dai, Z.; Cai, B.; Lin, Y.; and Chen, J. 2021. Up-detr: Unsupervised pre-training for object detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1601–1610.

Deng, J.; Yang, Z.; Chen, T.; Zhou, W.; and Li, H. 2021. Transvg: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1769–1779.

Dong, J.; Wang, Y.; Chen, X.; Qu, X.; Li, X.; He, Y.; and Wang, X. 2022. Reading-strategy Inspired Visual Representation Learning for Text-to-Video Retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*.

Girshick, R. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 1440–1448.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

Hong, R.; Liu, D.; Mo, X.; He, X.; and Zhang, H. 2019. Learning to compose and reason with language tree structures for visual grounding. *IEEE transactions on pattern analysis and machine intelligence*, 44(2): 684–696.

Hu, R.; Rohrbach, M.; Andreas, J.; Darrell, T.; and Saenko, K. 2017. Modeling relationships in referential expressions with compositional modular networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1115–1124.

Hua, G.; Liao, M.; Tian, S.; Zhang, Y.; and Zou, W. 2023. Multiple relational learning network for joint referring expression comprehension and segmentation. *IEEE Transactions on Multimedia*, 25: 8805–8816.

Lai, X.; Tian, Z.; Chen, Y.; Li, Y.; Yuan, Y.; Liu, S.; and Jia, J. 2023. LISA: Reasoning Segmentation via Large Language Model. *arXiv preprint arXiv:2308.00692*.

Li, M.; and Sigal, L. 2021. Referring transformer: A one-step approach to multi-task visual grounding. *Advances in neural information processing systems*, 34: 19652–19664.

Liao, Y.; Zhang, A.; Chen, Z.; Hui, T.; and Liu, S. 2022. Progressive language-customized visual feature learning for one-stage visual grounding. *IEEE Transactions on Image Processing*, 31: 4266–4277.

Liu, B.; Zheng, Q.; Wang, Y.; Zhang, M.; Dong, J.; and Wang, X. 2022a. FeatInter: exploring fine-grained object features for video-text retrieval. *Neurocomputing*, 496: 178–191.

Liu, S.; Li, F.; Zhang, H.; Yang, X.; Qi, X.; Su, H.; Zhu, J.; and Zhang, L. 2022b. Dab-detr: Dynamic anchor boxes are better queries for detr. *arXiv preprint arXiv:2201.12329*.

Liu, X.; Wang, Z.; Shao, J.; Wang, X.; and Li, H. 2019. Improving referring expression grounding with cross-modal attention-guided erasing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1950–1959.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Lu, M.; Li, R.; Feng, F.; Ma, Z.; and Wang, X. 2024. LGR-NET: Language Guided Reasoning Network for Referring Expression Comprehension. *IEEE Transactions on Circuits and Systems for Video Technology*.

Luo, G.; Zhou, Y.; Sun, X.; Cao, L.; Wu, C.; Deng, C.; and Ji, R. 2020. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 10034–10043.

Mao, J.; Huang, J.; Toshev, A.; Camburu, O.; Yuille, A. L.; and Murphy, K. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 11–20.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Rohrbach, A.; Rohrbach, M.; Hu, R.; Darrell, T.; and Schiele, B. 2016. Grounding of textual phrases in images by reconstruction. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, 817–834. Springer.

Su, W.; Miao, P.; Dou, H.; Wang, G.; Qiao, L.; Li, Z.; and Li, X. 2023. Language adaptive weight generation for multi-task visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10857–10866.

Sun, M.; Suo, W.; Wang, P.; Zhang, Y.; and Wu, Q. 2022. A proposal-free one-stage framework for referring expression comprehension and generation via dense cross-attention. *IEEE Transactions on Multimedia*, 25: 2446–2458.

Wang, L.; Li, Y.; and Lazebnik, S. 2016. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5005–5013.

Wang, Y.; Dong, J.; Liang, T.; Zhang, M.; Cai, R.; and Wang, X. 2022. Cross-Lingual Cross-Modal Retrieval with Noise-Robust Learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, 422–433.

Wang, Y.; Wang, F.; Dong, J.; and Luo, H. 2023. CL2CM: Improving Cross-Lingual Cross-Modal Retrieval via Cross-Lingual Knowledge Transfer. *arXiv preprint arXiv:2312.08984*.

Wang, Y.; Wang, L.; Zhou, Q.; Wang, Z.; Li, H.; Hua, G.; and Tang, W. 2024a. Multimodal LLM Enhanced Cross-lingual Cross-modal Retrieval. arXiv:2409.19961.

Wang, Y.; Wang, S.; Luo, H.; Dong, J.; Wang, F.; Han, M.; Wang, X.; and Wang, M. 2024b. Dual-view Curricular Optimal Transport for Cross-lingual Cross-modal Retrieval. *IEEE Transactions on Image Processing*, 33: 1522–1533.

Yang, S.; Chen, Y.; Tian, Z.; Wang, C.; Li, J.; Yu, B.; and Jia, J. 2024a. VisionZip: Longer is Better but Not Necessary in Vision Language Models. *arXiv preprint arXiv:2412.04467*.

Yang, S.; Qu, T.; Lai, X.; Tian, Z.; Peng, B.; Liu, S.; and Jia, J. 2023. An Improved Baseline for Reasoning Segmentation with Large Language Model. *arXiv preprint arXiv:2312.17240*.

Yang, S.; Tian, Z.; Jiang, L.; and Jia, J. 2024b. Unified Language-driven Zero-shot Domain Adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23407–23415.

Yang, Z.; Chen, T.; Wang, L.; and Luo, J. 2020. Improving one-stage visual grounding by recursive sub-query construction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, 387–404. Springer.

Ye, J.; Tian, J.; Yan, M.; Yang, X.; Wang, X.; Zhang, J.; He, L.; and Lin, X. 2022. Shifting more attention to visual backbone: Query-modulated refinement networks for end-to-end visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15502–15512.

Yu, L.; Lin, Z.; Shen, X.; Yang, J.; Lu, X.; Bansal, M.; and Berg, T. L. 2018. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1307–1315.

Yu, L.; Poirson, P.; Yang, S.; Berg, A. C.; and Berg, T. L. 2016. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, 69–85. Springer.

Yu, L.; Tan, H.; Bansal, M.; and Berg, T. L. 2017. A joint speaker-listener-reinforcer model for referring expressions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7282–7290.

Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L. M.; and Shum, H.-Y. 2022. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*.

Zhang, H.; Niu, Y.; and Chang, S.-F. 2018. Grounding referring expressions in images by variational context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4158–4166.

Zhao, H.; Zhou, J. T.; and Ong, Y.-S. 2022. Word2pix: Word to pixel cross-attention transformer in visual grounding. *IEEE Transactions on Neural Networks and Learning Systems*.

Zheng, Q.; Dong, J.; Qu, X.; Yang, X.; Wang, Y.; Zhou, P.; Liu, B.; and Wang, X. 2023. Progressive localization networks for language-based moment localization. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(2): 1–21.

Zhou, Y.; Ji, R.; Luo, G.; Sun, X.; Su, J.; Ding, X.; Lin, C.-W.; and Tian, Q. 2021. A real-time global inference network for one-stage referring expression comprehension. *IEEE Transactions on Neural Networks and Learning Systems*.

Zhu, C.; Zhou, Y.; Shen, Y.; Luo, G.; Pan, X.; Lin, M.; Chen, C.; Cao, L.; Sun, X.; and Ji, R. 2022. Seqtr: A simple yet universal network for visual grounding. In *European Conference on Computer Vision*, 598–615. Springer.

Zhuang, B.; Wu, Q.; Shen, C.; Reid, I.; and Van Den Hengel, A. 2018. Parallel attention: A unified framework for visual object discovery through dialogs and queries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4252–4261.