# Weakly Supervised Referring Image Segmentation with Intra-Chunk and Inter-Chunk Consistency

Jungbeom Lee[1] Sungjin Lee[1] Jinseok Nam[1] Seunghak Yu[2] Jaeyoung Do[1] Tara Taghavi[1]
[1]Amazon [2]NAVER Search US
jbeom.lee93@gmail.com, {sungjinl, jinseo}@amazon.com, seunghak.yu@navercorp.com,
{domjae, taghavit}@amazon.com

## Abstract

*Referring image segmentation aims to localize the object in an image referred by a natural language expression. Most previous studies learn referring image segmentation with a large-scale dataset containing segmentation labels, but they are costly. We present a weakly supervised learning method for referring image segmentation that only uses readily available image-text pairs. We first train a visual-linguistic model for image-text matching and extract a visual saliency map through Grad-CAM to identify the image regions corresponding to each word. However, we found two major problems with Grad-CAM. First, it lacks consideration of critical semantic relationships between words. We tackle this problem by modeling the relationship between words through intra-chunk and inter-chunk consistency. Second, Grad-CAM identifies only small regions of the referred object, leading to low recall. Therefore, we refine the localization maps with self-attention in Transformer and unsupervised object shape prior. On three popular benchmarks (RefCOCO, RefCOCO+, G-Ref), our method significantly outperforms recent comparable techniques. We also show that our method is applicable to various levels of supervision and obtains better performance than recent methods.*

## 1. Introduction

Referring image segmentation aims to obtain a pixel-level segmentation mask of the object in an image referred by a natural language expression. It has a wide range of practical applications in the real world such as human-robot interaction [38, 46] and visual navigation [37]. To learn referring image segmentation, a neural network should not only comprehend the semantics of image and text respectively, but also be able to capture the semantic alignment between the two modalities. A general approach to achieving this objective is to leverage the dataset with fully supervised labels. It necessitates numerous pairs of images and texts, together with a pixel-level segmentation mask of the referred object.

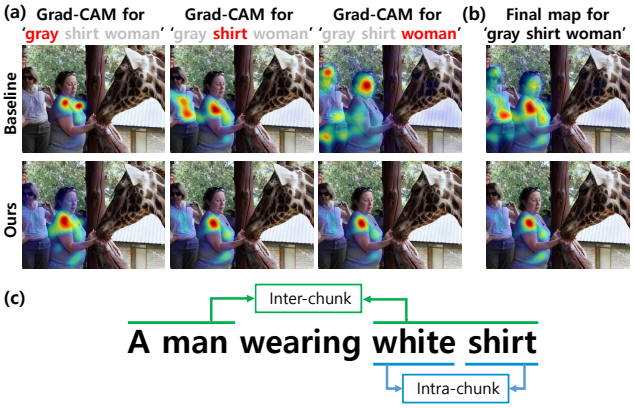


Figure 1: (a) Examples of Grad-CAMs of each word corresponding to **red** word given "gray shirt woman". (b) Final resulting maps of baseline and ours for "gray shirt woman". (c) Illustration of intra-chunk and inter-chunk relationship.

By using these explicit connections between image and text, recent studies [19, 47, 50] have successfully performed referring image segmentation.

However, constructing a dataset equipped with pixel-level segmentation labels is extremely laborious and expensive. For instance, annotating a segmentation label for a single image featuring a complex scene (*e.g.,* CityScapes [8]) requires more than 90 minutes. As such, our objective is to mitigate this quandary with weakly supervised learning, which trains a neural network by using only readily available image-text pairs, without expensive segmentation labels.

There have been some weakly supervised approaches [10, 35, 43] to localize the object referred by the given text, in the form of a bounding box instead of a segmentation mask. Nevertheless, these approaches are hindered by two key drawbacks: 1) they do not provide pixel-level localization of the referred object, and 2) most of them heavily depend on pre-trained object detector, which actually requires explicit object localization labels (*i.e.,* box [21]). To the best of our knowledge, there is only one recent work [42] that learns referring image segmentation using only image-text pairs,

but their performance lags behind fully supervised methods.

To accomplish referring image segmentation utilizing solely image-text pairs, we first train a visual-linguistic Transformer [27] via an image-text matching (ITM) objective, where the model is trained to determine whether a given text describes the corresponding image or not. To realize ITM, the model should learn the joint semantics of image and text. We then extract the knowledge of the trained model using Grad-CAM [40]. It computes the rationale for why the model thinks the given image-text pair is matched. As a result, Grad-CAM provides a visual saliency map representing which image regions are correlated with each word in a given sentence.

Ideally, the Grad-CAM of each word in a given sentence should be generated by considering the semantics of its neighboring words, where the relationship between words can be modeled by self-attention in Transformer [44]. For instance, in the sentence "gray shirt woman" as depicted in Figure 1(a), the Grad-CAM of 'woman' should identify only the woman wearing a gray shirt, not other women in the image, by taking into account the semantics of 'gray' and 'shirt'. However, our baseline model, which was trained only with ITM, struggles to capture the compositional consistency between words, resulting in inconsistent localization of the Grad-CAMs of each word in the sentence. For example, in Figure 1(a), the baseline's Grad-CAM for 'woman' identifies all the women in the image without considering the neighboring words. As a result, the final localization, obtained by averaging Grad-CAMs over the words in the sentence, fails to exclusively locate the woman wearing a gray shirt, as shown in Figure 1(b). This means that the baseline lacks consideration of the relationships between words, which is problematic for referring image segmentation where only the referred object should be identified. Therefore, in this work, we propose a novel regularization technique to incorporate the intra-chunk and inter-chunk relationships so that the model considers the relationships between words in the given expression.

Specifically, sentences consist of several noun-chunks. A noun-chunk is a group of words, which consists of a head noun and its modifying (dependent) words. In the training of ITM, we add regularization terms to produce consistent localization maps *between words* in a single chunk (intra-chunk), and *between chunks* (inter-chunk) (Figure 1(c)). Although tree-based recursive linguistic structure [6] is popular for capturing such relationships, they usually contain too much of unrelated details, thereby making it hard to extract only necessary information. Thus, we simplify this with noun-chunk-level representations.

The Grad-CAM obtained with intra-chunk and inter-chunk consistency provides a more accurate localization of the referred object, but it still has two drawbacks. First, it identifies only small regions of the referred object because all the regions of the target object are not necessary for ITM. Second, due to the absence of object shape information in the image-text pair, Grad-CAM does not represent the exact boundary of the object. Therefore, we propose two refinement techniques to obtain a more complete segmentation of the referred object, using patch affinities obtained from visual Transformers and unsupervised object shape prior.

Our main contributions include (1) intra-chunk and inter-chunk consistency to improve Grad-CAMs by considering the relationship between words in a given text; (2) two refinement techniques for a more accurate segmentation of the referred object; (3) significantly better performance on the three popular benchmarks than existing methods under the same level of supervision; and (4) versatility of our approach that allows integration with various levels of supervision.

## 2. Related Works

### 2.1. Referring Image Segmentation

Most of referring image segmentation methods depend on fully supervised labels. Early methods [14, 28, 31] obtain visual features from convolutional neural network and text features from recurrent neural network separately, and merge the two types of information with a simple fusion technique such as concatenation or summation. With the development of Transformers and attention modules, many recent works [19, 30, 47, 50] have focused on deep interactions between image and text through multi-modal cross-attention. Tremendous progress on referring image segmentation has been made with advancements in Transformers and attention modules. However, exact segmentation labels are essential to such methods, which are costly.

### 2.2. Localizing Referred Object from Weak Labels

To alleviate the heavy dependency on fully supervised labels, many researchers have tried to localize the object referred by a natural language expression from weak labels. However, most are limited to localizing the referred object in the form of bounding boxes, also known as weakly supervised visual grounding [10, 12, 35, 43, 45, 48, 52]. They frequently depend on a pre-trained object detector, but the large-scale dataset with bounding box labels is required to pre-train the object detector. Therefore, it is difficult to consider that these methods use only image-text pairs to perform referred object localization. Some detector-free methods have also been proposed [1, 3, 17, 53]. However, since they have difficulties in capturing the whole regions of a target object, they evaluate their localization performance using pointing game accuracy [53] that only considers the most confident region in the predicted localization.

Weakly supervised semantic segmentation provides segmentation masks from the class names [18, 22–25] or bounding boxes [26], but limited to pre-defined and fixed set of

classes. Zero-shot segmentation and open-world segmentation from the text are also recent active topics. Most works [32, 39, 54] still require segmentation labels for a set of specific classes (base classes) to segment objects of novel classes, but these segmentation labels are still expensive. Recent works [29, 49] achieve zero-shot segmentation without any segmentation labels, by using only image-text pairs. However, they are only applicable when the given expression is a class name, which is short and simple, and they also require delicate prompt engineering. Therefore, those methods are not suitable for referring image segmentation, which requires the ability to process long and complex textual expressions.

A few works learn referring image segmentation using only image-text pairs. TSEG [42] learns the local matching between image regions and text through the multiple instance learning technique, but its performance is far from that of fully supervised ones. Peekaboo [5] leverages diffusion models [7, 13, 41]. Another work, Feng *et al.* [11], proposes a method of utilizing bounding boxes as supervision for referring image segmentation. However, the box labels are pretty expensive compared to image-text pairs.

## 3. Proposed Method

We first obtain the matching between image regions and each word in a given sentence from the knowledge of the visual-linguistic Transformer in Section 3.1. To obtain more accurate region-word matching, we introduce intra-chunk and inter-chunk consistency in Section 3.2. We then propose refinement techniques to obtain complete segmentation of the referred object in Section 3.3. Finally, we propose a method of utilizing additional supervision in Section 3.4.

### 3.1. Localization of Referred Object

To localize the referred object in an image using only image-text pairs, a powerful joint representation between image and text is required. One way to obtain such a representation is to leverage a visual-linguistic multi-modal Transformer equipped with cross-attention layers. To this end, we choose ALBEF [27] as our base model. It contains a visual encoder $E_v$ and a text encoder $E_t$. Image tokens $x_v \in \mathbb{R}^{(1+N_p)\times d}$ can be obtained by dividing a given image into $N_p$ non-overlapping patches and encoding each patch into $d-$dimension vector. Word tokens $x_t \in \mathbb{R}^{(1+N_w)\times d}$ can be obtained by tokenizing and encoding each word into $d-$dimension vector, where $N_w$ is the number of words in the given expression. Note that a learnable [CLS] token is attached to each modality's token.

We now obtain visual features $E_v(x_v) \in \mathbb{R}^{(1+N_p)\times d}$ from $x_v$ and word features $E_t(x_t) \in \mathbb{R}^{(1+N_w)\times d}$ from $x_t$. These uni-modal features are merged by multi-modal encoder $E_m$ that contains several cross-attention layers, resulting in multi-modal features. The [CLS] token of the multi-modal feature contains the joint representation of image and text, so we obtain image-text matching (ITM) score $s^{\text{ITM}}$ by appending a fully connected layer to the [CLS] token. The ITM score $s^{\text{ITM}}$ is trained to produce positive scores for matched image-text pairs, and negative scores for mismatched ones. As a result of the training, the model obtains the ability to model the joint semantics of image and text, which in turn allows us to extract the relationship between image regions and words using model interpretation techniques such as Grad-CAM [40].

Grad-CAM [40] is a popular technique to interpret the output of the neural network. It computes the contribution of intermediate features of the network to its output from gradient flows. We obtain Grad-CAM for ALBEF [27] as follows: We first obtain the cross-attention map[1] $A_{\text{multi}} \in \mathbb{R}^{(1+N_p)\times(1+N_w)}$ from $E_m$, and compute the contribution of $A_{\text{multi}}$ to the $s^{\text{ITM}}$ as $\partial s^{\text{ITM}}/\partial A_{\text{multi}}$. The Grad-CAM $M \in \mathbb{R}^{(1+N_p)\times(1+N_w)}$ can be expressed as follows:

$$M = \text{ReLU}(A_{\text{multi}} \times \frac{\partial s^{\text{ITM}}}{\partial A_{\text{multi}}}). \tag{1}$$

Each element in $M \in \mathbb{R}^{(1+N_p)\times(1+N_w)}$ measures how much of a contribution each word-patch pair makes to $s^{\text{ITM}}$. Therefore, we can obtain the visual saliency map for each word $w$, *i.e.,* $M_w \in \mathbb{R}^{\sqrt{N_p}\times\sqrt{N_p}}$, by collecting all of the contribution scores with respect to $w$. More specifically, we collect the Grad-CAM scores corresponding to word $w$ (*i.e.,* $M[:, \text{id}_w] \in \mathbb{R}^{(1+N_p)}$ where $\text{id}_w$ is the index corresponding to $w$), remove [CLS] token in image patch tokens, and reshape it to 2-D matrix. The examples of Grad-CAM of each word are shown in Figure 1(a).

Ideally, the Grad-CAM of each word should be generated by considering the semantics of its neighboring words, which can be realized by self-attention between words in Transformer [44]. However, the model trained only with ITM struggles to capture the compositional consistency between words, resulting in inconsistent localization of the Grad-CAMs of each word in the sentence. For example, in Figure 1(a), the baseline's Grad-CAM for 'woman' identifies all the women in the image without considering the neighboring words such as 'gray' and 'shirt'. This is problematic for referring image segmentation where only the referred object should be identified. Therefore, in this work, we propose a novel regularization technique to incorporate the intra-chunk and inter-chunk relationships so that the model considers the relationships between words in the given expression.

### 3.2. Chunk-Level Representation Learning

Although tree-based recursive linguistic structure [6] is popular for capturing the compositional relationship, they usually contain too much of unrelated details, essentially

[1] For simplicity, the multi-head attentions are averaged.

making it hard to extract only necessary information. Thus, we employ a chunk-level representation through noun chunking. Specifically, we assume a given text contains a set of noun-chunks $\mathcal{C} = \{c\}$. A noun-chunk is a group of words $c = \{w\}$, which consists of a head noun and its modifying (dependent) words. We model the relationship *between the words* in a noun-chunk (Intra-chunk relationship) and the relationship *between chunks* (Inter-chunk relationship) to consider both local and global relationships (Figure 1(c)). In order to incorporate these consistency requirements, we train our model with the loss $L$:

$$L = L_{\text{ALBEF}} + \lambda_1 L_{\text{intra}} + \lambda_2 L_{\text{inter}}, \tag{2}$$

where $L_{\text{ALBEF}}$ is the one used in ALBEF [27] and $L_{\text{intra}}$ and $L_{\text{inter}}$ are intra-chunk and inter-chunk consistency losses respectively, described in detail in the following sections.

#### 3.2.1 Intra-Chunk Consistency

Since a single noun-chunk consists of a head noun and its dependent words describing the head noun, all the words in a noun-chunk should indicate the same object in an image. However, Grad-CAMs of each word in a noun-chunk tend to identify different objects, as shown in Figure 1(a). Therefore, we introduce a regularization such that the words in a single noun-chunk have similar Grad-CAMs. For each pair of words $(w_i, w_j)$ in a chunk $c$, we introduce $L_{\text{intra}}$ to reduce the difference between the two Grad-CAMs of $w_i$ and $w_j$ ($M_{w_i}$ and $M_{w_j}$). Specifically, $L_{\text{intra}}$ is defined as follows:

$$L_{\text{intra}} = \sum_{c \in \mathcal{C}} \sum_{(w_i, w_j) \in c} \texttt{cos}(\texttt{vec}(M_{w_i}), \texttt{vec}(M_{w_j})), \tag{3}$$

where $\texttt{vec}$ is to vectorize the 2-D matrix into 1-D vector, and $\texttt{cos}$ is the cosine distance between two maps. We choose cosine distance, which inherently normalizes its inputs, because $M_{w_i}$ and $M_{w_j}$ may have different scales.

#### 3.2.2 Inter-Chunk Consistency

Since all the words in a single noun-chunk indicate the same object in an image, we were able to make the Grad-CAMs of words in a chunk similar to each other in Eq. 3. However, incorporating the relationship *between chunks* is not straightforward because two chunks may indicate the same object or different objects depending on the semantic meaning of the predicate that connects the two chunks. For example, for the sentence "man wearing white shirt", the chunk 'man' and the chunk 'white shirt' correspond to the same object. Whereas, for the sentence "man holding a donut", the chunk 'man' and the chunk 'a donut' should not be mapped to the same object. Thus, without explicit localization cues, it is difficult to model the relationships between chunks.


Figure 2: Example of inter-chunk consistency given "Man in blue shirt". (*left*) Image. (*middle*) Localization of 'man' and resulting isolated regions. (*right*) Localization of 'blue shirt'. Of the three isolated regions, the yellow one is closest to the 'blue shirt'. We thus enhance the scores in the yellow region, and suppress the scores of the blue and red regions.

Therefore, we propose a closeness prior: The objects corresponded to adjacent chunks must be spatially close to each other in the image. In the example of "man holding a donut", among many men in an image, a man near a donut is likely to be the target object. Our goal is to make the regions identified by two related chunks spatially close to each other in the pixel-space. However, the definition of *closeness* is ambiguous because the pixel distance criterion of 'close' and 'far' depends on the object sizes.

We thus define *closeness* using the isolated regions of Grad-CAM. We first compute the localization map of a chunk $c$ by averaging the Grad-CAMs of words in the chunk: $M_c = \frac{1}{|c|} \sum_{w \in c} M_w$. Without loss of generality, for adjacent chunks $c_i$ and $c_j$ ($i \neq j$), we make the region indicated by $M_{c_i}$ spatially close to that indicated by $M_{c_j}$. In the example of Figure 2, $c_i$ is 'man' and $c_j$ is 'blue shirt'. If $M_{c_i}$ identifies some wrong objects, it often has multiple isolated regions, as shown in Figure 2. Following the closeness prior, the isolated region close to $M_{c_j}$ is likely to be the referred object, so the isolated regions far from $M_{c_j}$ should be suppressed. To realize this, we obtain a set of isolated regions $R$ from $M_{c_i}$, which can be readily obtained by using the $\texttt{cv2}$ library. Note that each $r \in R$ is a set of pixel locations within each isolated region. We then select the closest isolated region $r^* \in R$ to $M_{c_j}$ as follows:

$$r^* = \arg\min_{r \in R} d_{\text{H}}(r, p(M_{c_j}, \tau)), \tag{4}$$

where $d_{\text{H}}$ is the distance between two sets of locations, and $p(X, \tau)$ is the set of point locations whose localization scores in $X$ are larger than $\tau$. Motivated by Hausdorff distance [16], we define $d_{\text{H}}(X, Y)$ as the distance between two closest pair of points in $X$ and $Y$, as follows:

$$d_{\text{H}}(X, Y) = \min_{x \in X} \min_{y \in Y} ||x - y||^2. \tag{5}$$

We now design $L_{\text{inter}}$ to enhance the scores of $M_{c_i}$ corresponding to locations of $r^*$, which is the closest region to

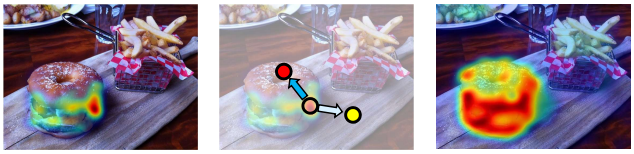

Figure 3: Illustration of the refinement with affinity. The score of each pixel is propagated to the semantically similar regions (orange-red circles), not to the dissimilar regions (orange-yellow circles). This can be realized by the patch affinity obtained from self-attention in visual Transformer.

$M_{c_j}$, while suppressing the scores of $M_{c_i}$ corresponding to locations of $R \setminus r^*$, as follows:

$$L_{\text{inter}} = -\sum_{k \in r^*} M_{c_i}^k + \sum_{k \in R \setminus r^*} M_{c_i}^k, \tag{6}$$

where $k$ denotes the location index and $M^k$ is the score of $M$ at the location $k$. In practice, we compute Eq 6 for all the neighboring chunks and average them.

### 3.3. Refinement Technique

During the inference process, we consider the average of Grad-CAMs over words in the given sentence as the final localization map of the referred object, $M_{\text{final}}$. The $M_{\text{final}}$ provides an accurate indication of the referred object. However, because only a small part of the target object can provide strong signals for the ITM, the Grad-CAMs tend to cover only small regions of the target object, as shown in Figure 1(a). In addition, because image-text pairs do not provide any object shape prior, the resulting Grad-CAMs do not depict the exact boundary of the object. Therefore, we propose two refinement techniques to obtain a more accurate segmentation of the referred object.

#### 3.3.1 Patch Affinity in Vision Transformer

Similar to Vision Transformer (ViT) [9], our base model ALBEF [27] has a visual encoder consisting of self-attention layers. The self-attention layer captures the semantic relationship between image patches, and this relationship can be considered the affinity between patches. We thus propagate the localization score of each pixel to its semantically relevant neighbors based on the affinities, as shown in Figure 3. More specifically, we have the self-attention map $A_v \in \mathbb{R}^{N_p \times N_p}$ obtained from the $l^{\text{th}}$ layer in the visual encoder[2]. The refined map $M^r$ can be obtained by propagating the scores in $M_{\text{final}}$ using $A_v$, as follows:

$$M^r = \texttt{reshape}_{2d}(A_v^T \texttt{vec}(M_{\text{final}})), \tag{7}$$

where $\texttt{reshape}_{2d}$ is to reshape 1-D vector back to 2-D matrix. The self-attention map $A_v$ is automatically obtained during the forward process of ALBEF, so that the additional process for this refinement is only a single matrix multiplication in Eq. 7, which incurs negligible additional computation.

[2]we choose $l$ =9, and the multi-head attentions are simply averaged.

#### 3.3.2 Unsupervised Shape Prior

The refinement with affinity in Section 3.3.1 enables us to find a more complete region of the referred object. However, due to the absence of the object shape prior, the resulting localization maps cannot represent the exact boundary of the referred object. Therefore, we further refine our localization maps using unsupervised object shape prior. We utilize multi-scale combinatorial grouping (MCG) [2] to refine our localization maps. MCG operates on the low-level information of images in an unsupervised manner, so that it does not violate the basic requirements of weakly supervised learning. MCG generates multiple mask proposals $\{m\}$ for a single image. Among these proposals, we choose the proposal $m^*$ that overlaps the most with $M^r$, in terms of intersection-over-union (IoU). We determine $m^*$ as the final localization.

### 3.4. Utilization of Additional Supervision

Our method can operate with various levels of supervision, which shows the generality and practicality of our method. We consider two settings: 1) weakly supervised setting with bounding box labels and 2) semi-supervised setting. In the setting 1), we assume we have bounding box labels of the referred objects for the training images. We obtain the pixel-level localization in the given box using the BBAM [26] technique, which is the mask generator from a box. We consider the resulting localization map as a pseudo ground truth segmentation $Y_{\text{box}}$. We then add a box loss $L_{\text{box}}$ to Eq. 2 so that the produced Grad-CAM $M_{\text{final}}$ is similar to $Y_{\text{box}}$: $L_{\text{box}} = \texttt{cos}(M_{\text{final}}, Y_{\text{box}})$. In the setting 2), we assume we have fully supervised pixel-level labels $Y_{\text{full}}$ only for a small number of training images. For these images, we compute $L_{\text{semi}} = \texttt{cos}(M_{\text{final}}, Y_{\text{full}})$ and optimize the network together with Eq. 2. Note that we do not apply any refinement techniques in these two settings because the provided explicit localization cues can inherently address the drawbacks of Grad-CAMs mentioned in Section 3.3.

## 4. Experiments

### 4.1. Experimental Setup

**Dataset:** We conducted experiments on the three popular benchmarks: RefCOCO [51], RefCOCO+ [51], and G-Ref with Google split [33]. RefCOCO, RefCOCO+, and G-Ref datasets contain respectively 142,209, 141,564, and 104,560 expressions. It is known that RefCOCO+ and G-Ref are more challenging datasets than RefCOCO, becuase RefCOCO+ prohibits object location information in the expression, and G-Ref has a longer average expression length.

Table 1: Comparison with referring image segmentation methods. All the results are obtained before applying CRF [20]. $\mathcal{F}$–full supervision, $\mathcal{W}$–weak supervision (image-text pairs).

| Method | Sup. | RefCOCO *val* | RefCOCO *testA* | RefCOCO *testB* | RefCOCO+ *val* | RefCOCO+ *testA* | RefCOCO+ *testB* | GRef *val* |
|---|---|---|---|---|---|---|---|---|
| RMI ICCV '17 [31] | $\mathcal{F}$ | 44.33 | 44.74 | 44.63 | 29.91 | 30.37 | 29.43 | 33.11 |
| DMN ECCV '18 [34] | $\mathcal{F}$ | 49.78 | 54.83 | 45.13 | 38.88 | 44.22 | 32.29 | 34.52 |
| Hu *et al.* CVPR '20 [15] | $\mathcal{F}$ | 60.98 | 62.99 | 59.21 | 48.17 | 52.32 | 42.11 | 47.57 |
| Kim *et al.* CVPR '22 [19] | $\mathcal{F}$ | 67.22 | 69.30 | 64.45 | 55.78 | 60.44 | 48.27 | 54.48 |
| GroupViT CVPR '22 [49] | $\mathcal{W}$ | 12.97 | 14.98 | 12.02 | 13.21 | 15.08 | 12.41 | 16.84 |
| TSEG arXiv '22 [42] | $\mathcal{W}$ | 25.44 | - | - | 22.01 | - | - | 22.05 |
| ALBEF NeurIPS '21 [27] | $\mathcal{W}$ | 23.11 | 22.79 | 23.42 | 22.44 | 22.07 | 22.51 | 24.18 |
| Ours | $\mathcal{W}$ | **31.06** | **32.30** | **30.11** | **31.28** | **32.11** | **30.13** | **32.88** |

Table 2: Precision @30, 50, 70 of weakly supervised methods on the RefCOCO+ validation set.

| Method | Pr@30 | Pr@50 | Pr@70 |
|---|---|---|---|
| GroupViT CVPR '22 [49] | 16.13 | 4.47 | 0.55 |
| ALBEF NeurIPS '21 [27] | 26.06 | 3.49 | 0.22 |
| Ours w/o. MCG | 45.32 | 15.09 | 1.83 |
| Ours w/. MCG | **46.81** | **25.07** | **9.72** |

Table 3: oIoU values of Feng *et al.* [11] and our method under bounding box supervision.

| Method | *val* | *testA* | *testB* |
|---|---|---|---|
| Dataset: RefCOCO | | | |
| Feng *et al.* TNNLS '22 [11] | 58.01 | 60.52 | **55.48** |
| Ours | **58.12** | **61.23** | 55.47 |
| Dataset: RefCOCO+ | | | |
| Feng *et al.* TNNLS '22 [11] | 47.12 | 50.86 | 40.26 |
| Ours | **48.19** | **53.01** | **42.83** |
| Dataset: G-Ref | | | |
| Feng *et al.* TNNLS '22 [11] | 46.03 | - | - |
| Ours | **49.64** | - | - |

**Evaluation metric:** We evaluate our results by calculating mean intersection-over-union (mIoU) values for validation and test images. We also present the values of Precision @$\tau$ ($\tau = 30, 50, 70$), which represent the percentage of test samples that has higher IoU than the threshold $\tau$.

**Reproducibility:** We set ($\lambda_1$, $\lambda_2$) to (0.5, 3.0), (1.0, 3.0), and (1.0, 5.0) for the RefCOCO, RefCOCO+, and G-Ref datasets, respectively. We set $\tau$ to 0.3. We obtain noun-chunks from the expression sentence using the `spacy` python library. More details are presented in the Appendix.

### 4.2. Experimental Results

**Comparison with State-of-the-Arts:** Table 1 compares our method with various state-of-the-art methods. Our method outperforms all strong baselines under the same level of supervision, including GroupViT [49] and TSEG [42]. Specifically, our method obtains a 41.7% relative gain compared to our baseline ALBEF [27] on the RefCOCO+ testA set. We also note that our method obtains better performance than the fully supervised method RMI [31] on the RefCOCO+ dataset, by using only weakly supervised image-text pairs. Examples of generated localization maps by GroupViT [49], ALBEF [27], and our method are shown in Figure 4. Since our baseline ALBEF [27] has difficulty in capturing the relationship between words, it also localizes another pizza in the first example and another truck in the third example. With intra-chunk and inter-chunk consistency, the referred objects are correctly identified.

**Precision at $\tau$:** We compare precision values at various IoU thresholds in Table 2. Our method can obtain consistently better results on all thresholds than other strong baselines. All the methods using only image-text pairs produce poor precision @70, due to the absence of object boundary information. By using unsupervised shape prior provided by MCG [2], our method significantly improves precision @70. Note that GroupViT [49] produces low IoU (Table 1) but higher precision @50 and 70 than ALBEF [27]. This is because zero-shot segmentation methods like GroupViT have strong segmentation ability, but struggle to identify the actual referred object. They are specialized for the expressions of a class name, which are short and simple. Therefore, when given long and complex sentences, they tend to identify all objects of the same category as referred objects, as shown in Figure 4. The results in Table 1 also support this: GroupViT [49] works better on G-Ref than RefCOCO and RefCOCO+, because the G-Ref's average number of objects of the same category in an image is less than other datasets.

**Weakly Supervised Learning with Boxes:** To our knowledge, there is only one comparable work using bounding box labels: Feng *et al.* [11], which is very recently published. In this experiment, we choose overall IoU (oIoU) instead of mIoU as the evaluation metric, for a fair comparison with Feng *et al.* [11]. oIoU is computed as the total intersection over the total union over all the test images. Table 3 compares our method with Feng *et al.* [11] on three datasets.

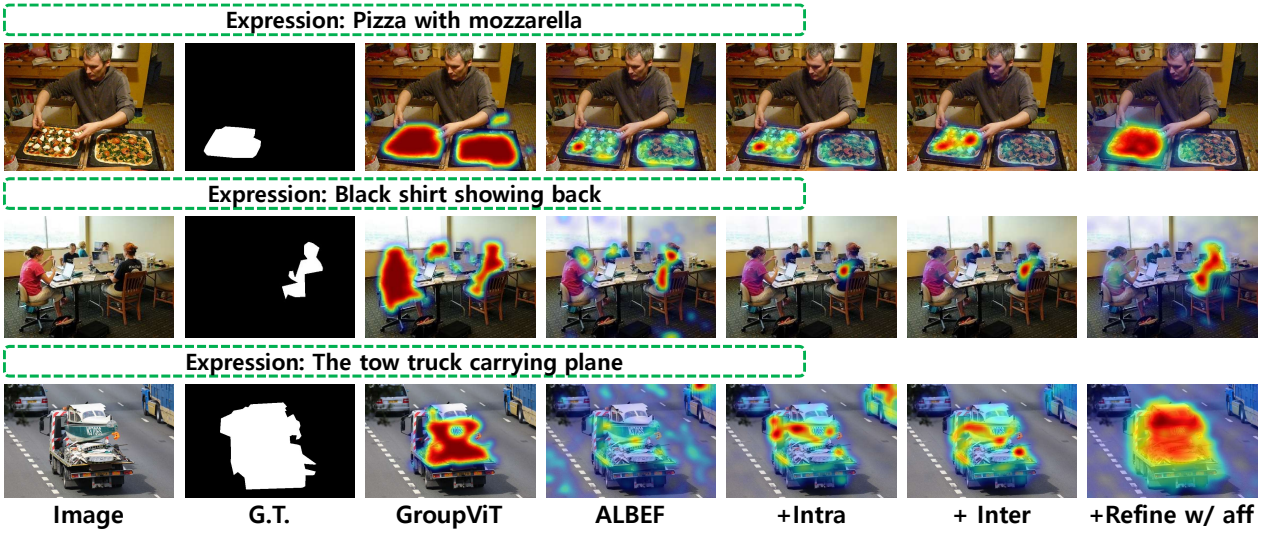


Figure 4: Examples of localization maps obtained by GroupViT [49], our baseline ALBEF [27], and ours by sequentially adding intra-chunk consistency, inter-chunk consistency, and refinement with affinity.

Our method obtains significantly better results than Feng *et al.* [11] on RefCOCO+ and G-Ref (e.g., 3.61%p on G-Ref) and comparable performance on RefCOCO, showing that our method is more beneficial for the challenging datasets. In addition, by using only 13% of the label cost[3], our method outperforms Hu *et al.* [15], a fully supervised method published in CVPR 2020 (Table 1), on the RefCOCO+ and G-Ref datasets.

**Semi-Supervised Learning:** We also discuss the performance under the semi-supervised setting. We choose RefCOCO+ [51] for the experiments. We vary the amount of fully supervised labels from 1% to 50%, and compare the results using 100% fully supervised labels. Figure 5 shows that competitive performance can be obtained with very few labels. Using only 1% of fully supervised labels, we obtain 68.1% relative gain over the weakly supervised setting (*i.e.,* 0% of fully supervised labels) on the testA set. Furthermore, with only 10% of fully supervised labels, our method achieved 90% of the performance of the fully supervised equivalent. The above analyses show that our method can be effectively combined with various types of supervision.

**Comparison to Fully Supervised Counterparts:** Our achievable upper-bound performance, by using 100% fully supervised labels, can be found in Figure 5. For instance, it exhibits 47.6 mIoU and 44.4 oIoU on RefCOCO+ *testB*. This means that our method achieves 63.3% and 96.5% of fully supervised performance by using image-text pairs and boxes respectively. We expect a better performance by using more powerful architecture (*e.g.,* advanced attention modules like LAVT [50] and larger-scale model), but the search for this architecture is out of our current priorities.

[3]According to Bellver *et al.* [4], boxes can be obtained 6× more efficiently than segmentation labels.

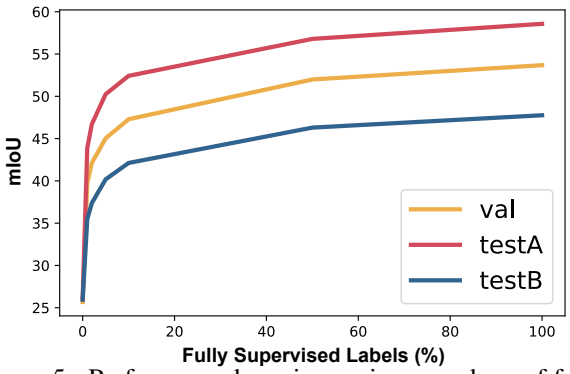


Figure 5: Performance by using various numbers of fully supervised labels for semi-supervised setting.

### 4.3. Analysis

**Ablation Studies:** We analyze the effectiveness of each proposed technique through ablation studies. Table 4 presents the mIoU values by sequentially adding each proposed module to the baseline ALBEF [27]. We can see consistent improvements with each proposed technique. Inter-chunk and intra-chunk consistency are beneficial for obtaining accurate localization of the referred object, resulting in improved precision @30. However, we do not observe significant improvements in precision @50 and 70 by intra-chunk and inter-chunk consistency. Even if the target object is accurately identified, it is difficult for Grad-CAM to make a segmentation of IoU 50 or higher due to the inherent problems of Grad-CAM mentioned in Section 3.3. The proposed refinement methods can expand the identified regions and obtain a more accurate boundary, significantly improving the values of precision @50 and 70. Analysis of hyper-parameter values is discussed in the Appendix.

Table 4: Effectiveness of each proposed technique on the *val* set. Starting from our baseline, we report mIoU and precision at 30, 50, and 70 by sequentially adding inter- and intra-chunk consistency, refinement with affinity, and refinement with MCG.

| | RefCOCO | | | | RefCOCO+ | | | | G-Ref | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | mIoU | Pr@30 | Pr@50 | Pr@70 | mIoU | Pr@30 | Pr@50 | Pr@70 | mIoU | Pr@30 | Pr@50 | Pr@70 |
| Baseline | 23.11 | 28.50 | 4.08 | 0.30 | 22.44 | 26.58 | 4.02 | 0.20 | 24.18 | 32.14 | 6.07 | 0.48 |
| +Inter-chunk | 25.48 | 34.72 | 5.03 | 0.19 | 24.35 | 31.59 | 4.89 | 0.20 | 26.61 | 38.16 | 8.04 | 0.68 |
| +Intra-chunk | 26.52 | 37.15 | 6.28 | 0.46 | 25.68 | 35.03 | 5.99 | 0.34 | 27.14 | 39.15 | 8.86 | 0.84 |
| +Refine w/ Aff | 29.00 | 42.51 | 14.50 | 1.86 | 29.58 | 45.32 | 15.09 | 1.83 | 29.76 | 46.47 | 16.63 | 2.51 |
| +Refine w/ MCG | **31.06** | **46.12** | **23.88** | **9.02** | **31.28** | **46.81** | **25.07** | **9.72** | **32.88** | **48.96** | **26.46** | **11.08** |

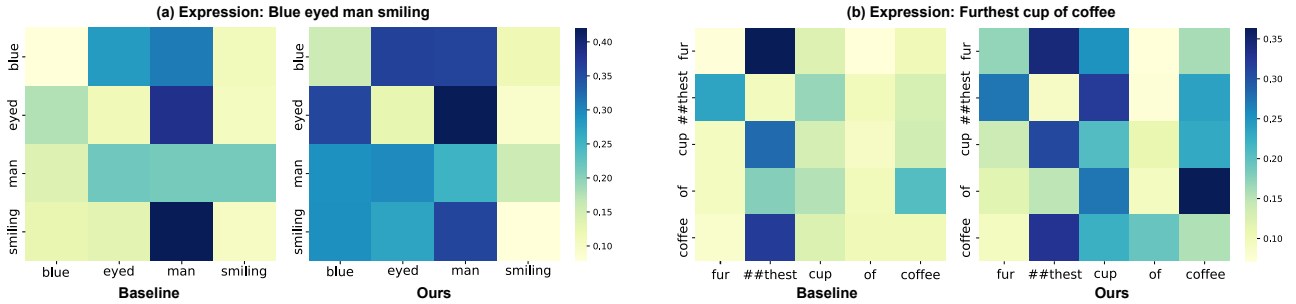


Figure 6: Visualization of self-attention scores between word tokens of baseline and ours for (a) "Blue eyed man smiling" and (b) "Furthest cup of coffee".

Figure 4 presents qualitative comparison of each ablative setting. Starting from our baseline ALBEF [27], we sequentially add intra-chunk consistency, inter-chunk consistency, and refinement with affinity. We can see that intra-chunk consistency is beneficial, but it still tends to produce inconsistent localization for the relationship across chunks (*e.g.,* 'pizza' and 'mozzarella' in the first example, and 'the tow truck' and 'plane' in the last example). Inter-chunk consistency can address this issue and produce a more accurate localization. Finally, the refinement with affinity is effective for obtaining more complete regions of the target object.

We now analyze the layer we use for patch affinity in Section 3.3.1. We chose the $9^{\text{th}}$ layer (among the 0 to 11 layers) because it contains information that is sufficiently semantically meaningful yet not too biased towards the self-supervised objectives. We tried to use other layers for obtaining patch affinities. The affinities obtained from $8^{\text{th}}$, $9^{\text{th}}$, and $10^{\text{th}}$ layers resulted in mIoU scores of 28.61, 29.58, and 20.71, respectively.

**Relationships that our method can handle:** As mentioned in Section 3.2.2, relationships between chunks demonstrate various semantic meaning. We now analyze the relationships our method can handle. For a clear explanation, we categorize the relationships into three distinct types: inclusive, exclusive, and negative. Inclusive relationships imply that the two connected chunks refer to the same object (*e.g.,* A man wearing blue shirt). Exclusive relationships indicate that the two connected chunks refer to different objects (*e.g.,* Woman throwing a frisbee). Negative relationships indicate that the two connected chunks have an opposing connection (*e.g.,* Donut without a hole).

Since our inter-chunk consistency regularizes that the two masks of neighboring chunks are simply spatially close in the pixel-space regardless of the relationships, our method can handle both inclusive and exclusive relationships. We empirically show that our method can handle both inclusive and exclusive relationships. We create a subset of RefCOCO+ *val* by selecting captions containing top-3 most frequent inclusive and exclusive relationships. We specifically focus on the captions with the structure of "[chunk] [relationship] [chunk]" because more complex captions can often lead to ambiguity in determining inclusivity or exclusivity. Table 5 demonstrates that our method bring significant improvements for both inclusive and exclusive relationships.

Finally, we discuss negative captions. Because our method assumes closeness of two neighboring chunks, negative captions may introduce bias during the training of our model. To analyze this issue, we tested an additional simple technique: We exclude negative captions when computing $L_{\text{intra}}$ (Eq. 3) and $L_{\text{inter}}$ (Eq. 6), while still considering them for $L_{\text{ALBEF}}$ in Eq. 2. This resulted in a slight improvement in performance: On RefCOCO+, 25.68 $\rightarrow$ 25.76 (*val*), 26.11 $\rightarrow$ 26.24 (*testA*), and 25.94 $\rightarrow$ 26.02 (*testB*). This results indicate that negative captions were indeed hindering the training of our model. However, we have found that the impact was not excessively significant and can be easily mitigated by excluding them for $L_{\text{intra}}$ and $L_{\text{inter}}$. Failure cases on negative captions are shown later.

**Visualization of Attention Maps**: To demonstrate the effectiveness of our consistency regularization in capturing the

Table 5: Performance gain over baseline for inclusive and exclusive relationships.

| | All | Inclusive | Exclusive |
|---|---|---|---|
| Baseline | 25.22 | 24.46 | 19.62 |
| Ours | 29.23 +4.01 | 28.72 +4.26 | 23.57 +3.95 |

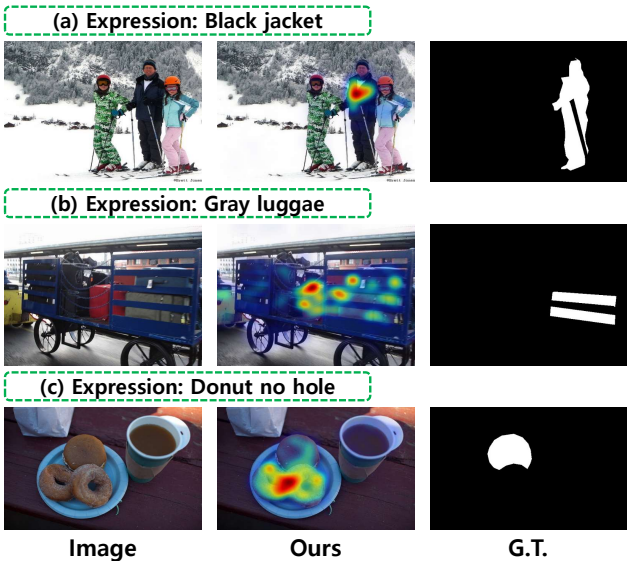


Figure 7: Examples of failure cases. (a) The expression describes only a part of the object. (b) The expression includes some typos. (c) The expression contains negative words.

relationship between words, we visualize the self-attention scores between word tokens in Figure 6. Our method produces notably higher self-attention scores between semantically related words than the baseline, as shown in Figure 6(a). Similarly, in Figure 6(b), while the baseline exhibits that the word tokens 'fur' and '##thest' primarily interact only with each other, our method allows these word tokens to interact with other semantically related words such as 'cup' and 'coffee'. Furthermore, the baseline demonstrates a low correlation between 'cup' and 'coffee', which causes Grad-CAM of 'cup' to potentially identify other cups, such as a cup of juice. In contrast, our method successfully captures the relationship between 'cup' and 'coffee' as indicated by the self-attention scores.

**Performance by expression sentence length:** We analyze the localization performance according to the length of given sentences. For the comparison, we choose the G-Ref dataset [36] because its average length of expression sentences is longer than RefCOCO and RefCOCO+. Note that we here omit refinement techniques to see the effectiveness of intra-chunk and inter-chunk consistency. We split the validation set by its sentence length and see the performance of each bin in Figure 8. Because long and complex sentence makes the referred object difficult to be accurately segmented, it is natural for long sentences to produce low performance. The baseline ALBEF [27] produces a 26.5%

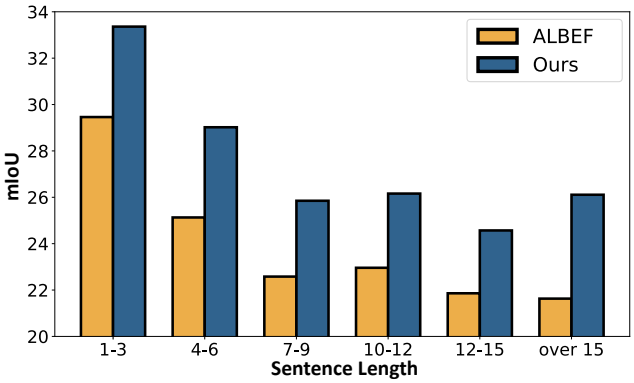


Figure 8: mIoU values for different lengths of expression sentences on the G-Ref dataset.

lower performance for long sentences (over 15 words) than short ones (1–3 words), while our method exhibits a less severe regression, 21.7%. This shows that our method can handle long sentences better than our baseline ALBEF [27].

**Failure examples:** We now analyze some failure cases where our method does not produce satisfactory results. As shown in Figure 7(a), ambiguous labels in the dataset are problematic, particularly for a person object. The expression tends to indicate only the part of the person (*e.g.,* black jacket), but the segmentation label covers the region of the whole person. Even though our method precisely localized the referred object by the given expression, the segmentation accuracy was low. The second example is a typo in the given expression. In Figure 7(b), 'luggage' is mistakenly given as 'luggae', so it produces a poor localization for the unknown word. Lastly, our method does not work properly for the negative terms (*e.g.,* 'no', 'not') as shown in Figure 7(c), because it is difficult to learn the negative meanings without explicit localization information of the referred object. We can partly address this issue with the utilization of the box labels, which are shown in the Appendix.

## 5. Conclusion

In this study, we proposed a novel method for learning referring image segmentation using only image-text pairs. Our approach leverages the linguistic structure of a given textual expression through intra-chunk and inter-chunk consistency to generate more precise localization maps. We then refine these maps using patch affinities obtained from self-attention maps of the visual Transformer and unsupervised object shape priors. Through extensive experiments, we showed that our proposed method outperforms the current state-of-the-art on three popular benchmarks. Moreover, we demonstrated the versatility of our approach by integrating it with various levels of supervision. In future work, it would be interesting to improve the robustness of our method by applying spelling correction to the expression and exploring new regularization techniques for negative constraints.

## References


[1] Hassan Akbari, Svebor Karaman, Surabhi Bhargava, Brian Chen, Carl Vondrick, and Shih-Fu Chang. Multi-level multimodal common semantic space for image-phrase grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12476–12486, 2019. 2

[2] Pablo Arbeláez, Jordi Pont-Tuset, Jonathan T Barron, Ferran Marques, and Jitendra Malik. Multiscale combinatorial grouping. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 328–335, 2014. 5, 6

[3] Assaf Arbelle, Sivan Doveh, Amit Alfassy, Joseph Shtok, Guy Lev, Eli Schwartz, Hilde Kuehne, Hila Barak Levi, Prasanna Sattigeri, Rameswar Panda, et al. Detector-free weakly supervised grounding by separation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1801–1812, 2021. 2

[4] Míriam Bellver Bueno, Amaia Salvador Aguilera, Jordi Torres Viñals, and Xavier Giró Nieto. Budget-aware semi-supervised semantic and instance segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2019*, pages 93–102, 2019. 7

[5] Ryan Burgert, Kanchana Ranasinghe, Xiang Li, and Michael S Ryoo. Peekaboo: Text to image diffusion models are zero-shot segmentors. *arXiv preprint arXiv:2211.13224*, 2022. 3

[6] Danqi Chen and Christopher D Manning. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 740–750, 2014. 2, 3

[7] Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11472–11481, 2022. 3

[8] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 1

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5

[10] Zi-Yi Dou and Nanyun Peng. Improving pre-trained vision-and-language embeddings for phrase grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6362–6371, 2021. 1, 2

[11] Guang Feng, Lihe Zhang, Zhiwei Hu, and Huchuan Lu. Learning from box annotations for referring image segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. 3, 6, 7

[12] Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. Contrastive learning for weakly supervised phrase grounding. In *European Conference on Computer Vision*, pages 752–768. Springer, 2020. 2

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[14] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *European Conference on Computer Vision*, pages 108–124. Springer, 2016. 2

[15] Zhiwei Hu, Guang Feng, Jiayu Sun, Lihe Zhang, and Huchuan Lu. Bi-directional relationship inferring network for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4424–4433, 2020. 6, 7

[16] Daniel P Huttenlocher, Gregory A. Klanderman, and William J Rucklidge. Comparing images using the hausdorff distance. *IEEE Transactions on pattern analysis and machine intelligence*, 15(9):850–863, 1993. 4

[17] Syed Ashar Javed, Shreyas Saxena, and Vineet Gandhi. Learning unsupervised visual grounding through semantic self-supervision. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 796–802, 2019. 2

[18] Eunji Kim, Siwon Kim, Jungbeom Lee, Hyunwoo Kim, and Sungroh Yoon. Bridging the gap between classification and localization for weakly supervised object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14258–14267, 2022. 2

[19] Namyup Kim, Dongwon Kim, Cuiling Lan, Wenjun Zeng, and Suha Kwak. Restr: Convolution-free referring image segmentation using transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18145–18154, 2022. 1, 2, 6

[20] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NeurIPS*, 2011. 6

[21] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. 1

[22] Jungbeom Lee, Jooyoung Choi, Jisoo Mok, and Sungroh Yoon. Reducing information bottleneck for weakly supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 34:27408–27421, 2021. 2

[23] Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5267–5276, 2019. 2

[24] Jungbeom Lee, Eunji Kim, Jisoo Mok, and Sungroh Yoon. Anti-adversarially manipulated attributions for weakly supervised semantic segmentation and object localization. *IEEE transactions on pattern analysis and machine intelligence*, 2022. 2

[25] Jungbeom Lee, Seong Joon Oh, Sangdoo Yun, Junsuk Choe, Eunji Kim, and Sungroh Yoon. Weakly supervised semantic segmentation using out-of-distribution data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16897–16906, 2022. 2

[26] Jungbeom Lee, Jihun Yi, Chaehun Shin, and Sungroh Yoon. Bbam: Bounding box attribution map for weakly supervised semantic and instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2643–2652, 2021. 2, 5

[27] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 2, 3, 4, 5, 6, 7, 8, 9

[28] Ruiyu Li, Kaican Li, Yi-Chun Kuo, Michelle Shu, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. Referring image segmentation via recurrent refinement networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2018. 2

[29] Yi Li, Huifeng Yao, Hualiang Wang, and Xiaomeng Li. Freeseg: Free mask from interpretable contrastive language-image pretraining for semantic segmentation. *arXiv preprint arXiv:2209.13558*, 2022. 3

[30] Zizhang Li, Mengmeng Wang, Jianbiao Mei, and Yong Liu. Mail: A unified mask-image-language trimodal network for referring image segmentation. *arXiv preprint arXiv:2111.10747*, 2021. 2

[31] Chenxi Liu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, and Alan Yuille. Recurrent multimodal interaction for referring image segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1271–1280, 2017. 2, 6

[32] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7086–7096, 2022. 3

[33] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 5

[34] Edgar Margffoy-Tuay, Juan C Pérez, Emilio Botero, and Pablo Arbeláez. Dynamic multimodal instance segmentation guided by natural language queries. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 630–645, 2018. 6

[35] Effrosyni Mavroudi and René Vidal. Weakly-supervised generation and grounding of visual descriptions with conditional generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15544–15554, 2022. 1, 2

[36] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *European Conference on Computer Vision*, pages 792–807. Springer, 2016. 9

[37] Khanh Nguyen, Debadeepta Dey, Chris Brockett, and Bill Dolan. Vision-based navigation with language-based assistance via imitation learning with indirect intervention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12527–12537, 2019. 1

[38] Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, Gokhan Tur, and Dilek Hakkani-Tur. Teach: Task-driven embodied agents that chat. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2017–2025, 2022. 1

[39] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18082–18091, 2022. 3

[40] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 2, 3

[41] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3

[42] Robin Strudel, Ivan Laptev, and Cordelia Schmid. Weakly-supervised segmentation of referring expressions. *arXiv preprint arXiv:2205.04725*, 2022. 1, 3, 6

[43] Mingjie Sun, Jimin Xiao, Eng Gee Lim, Si Liu, and John Y Goulermas. Discriminative triad matching and reconstruction for weakly referring expression grounding. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):4189–4195, 2021. 1, 2

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 3

[45] Liwei Wang, Jing Huang, Yin Li, Kun Xu, Zhengyuan Yang, and Dong Yu. Improving weakly supervised visual grounding by contrastive knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14090–14100, 2021. 2

[46] Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6629–6638, 2019. 1

[47] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11686–11695, 2022. 1, 2

[48] Fanyi Xiao, Leonid Sigal, and Yong Jae Lee. Weakly-supervised visual grounding of phrases with linguistic structures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5945–5954, 2017. 2

[49] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18134–18144, 2022. 3, 6, 7

[50] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18155–18165, 2022. 1, 2, 7

[51] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer, 2016. 5, 7

[52] Hanwang Zhang, Yulei Niu, and Shih-Fu Chang. Grounding referring expressions in images by variational context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4158–4166, 2018. 2

[53] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126(10):1084–1102, 2018. 2

[54] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision (ECCV)*, 2022. 3