# Similarity Maps for Self-Training Weakly-Supervised Phrase Grounding

Tal Shaharabany
Tel Aviv University
shaharabany@mail.tau.ac.il

Lior Wolf
Tel Aviv University
wolf@cs.tau.ac.il

## Abstract

*A phrase grounding model receives an input image and a text phrase and outputs a suitable localization map. We present an effective way to refine a phrase ground model by considering self-similarity maps extracted from the latent representation of the model's image encoder. Our main insights are that these maps resemble localization maps and that by combining such maps, one can obtain useful pseudo-labels for performing self-training. Our results surpass, by a large margin, the state of the art in weakly supervised phrase grounding. A similar gap in performance is obtained for a recently proposed downstream task called WWbL, in which only the image is input, without any text. Our code is available at* https://github.com/talshaharabany/Similarity-Maps-for-Self-Training-Weakly-Supervised-Phrase-Grounding.

## 1. Introduction

The most important technological foundation of the ongoing revolution in the field of jointly processing images with text is that of computing a similarity score between an image and a text phrase. Models such as CLIP [27] are able to learn powerful similarity functions based on large datasets that contain pairs of images and suitable captions. Building upon this technology, endless possibilities opened up for zero-shot and few-shot learning applications, including image captioning [19, 35], image editing [6, 13, 23, 25, 41], and image recognition [24, 40].

Phrase grounding is a related image-text task, in which, given an image and a text phrase, the method identifies the image region described by the phrase. One can train a phrase grounding network in a weakly supervised manner, by employing the CLIP similarity. Such a model would learn to produce masks that create a foreground region that is similar to the phrase and a background region that is not. To avoid trivial solutions, in which the mask includes the entire image, one should also include a regularization term.

Recently, Shaharabany et al. have obtained state-of-the-art results in phrase grounding using this scheme with the addition of a constraint that considers also the explainability map of Chefer et al. [8], when applied to the CLIP model. This map is expected to focus on foreground regions and can, therefore, provide an additional training cue.

In this work, we focus on the output of the image encoder of the phrase grounding network of [32]. This encoder creates a spatial image representation that is aligned with the text encoding of CLIP. In other words, each spatial location of the tensor that the image encoder outputs is associated with a vector that is correlated with a representation of the textual description of the corresponding image location.

As a result, spatial locations that are associated with the same image object have similar encoding and vice versa. Building upon this insight, we compute for each image location, the cosine-similarity to the local encoding of all other locations. These similarity maps (one per location) are text-agnostic since they involve only the image input of the phrase grounding network.

As we demonstrate, these maps capture different objects in the image, i.e., the similarity map that is obtained by correlating the image embedding at the spatial location $(x, y)$ provides a delineation of the object that can be found at this image location. This is a type of semantic "seed fill" effect, in which all image regions that are associated with the same object in the seed location are highlighted. These regions do not need to be connected, i.e., multiple persons in the image would be highlighted by a seed placed on one of them.

This localization-map effect is reminiscent of the emergence of segmentation maps in self-supervised transformer networks [5]. The two effects are, however, different. First, the effect in the transformer case relies on the attention maps, which do not exist in the networks we employ. Second, the attention maps rely on a single CLS token, and our effect is true to every spatial location. Third, the type of supervision and the task are both different.

The effect described is also different than the relevancy maps obtained by explainability methods such as Grad-CAM [31] or recent transformer explainability methods [8].

For example, our similarity maps do not require backtracking the relevancy scores all the way from the network's output. Also, the maps are per seed location and not per label.

By aggregating multiple similarity maps, we obtain fairly accurate foreground masks. Using these masks as pseudo labels, we refine the phrase grounding model using a supervised loss term. On all of the known phrase grounding benchmarks, we demonstrate that our refinement method leads to a marked improvement across multiple scores. Similarly, a sizable gap over the current state-of-the-art is also shown in the recently proposed computer vision task of "What is Where by Looking" (WWbL) [32], for which the phrase grounding network is coupled with a captioning model.

Our contributions are: (i) observing that spatial similarity maps of the latent space of phrase grounding networks capture the boundaries of objects, (ii) developing a method for aggregating multiple such similarity maps to obtain a comprehensive segmentation map, (iii) using the obtained maps to finetune the phrase grounding network, and (iv) surpassing the current state-of-the-art by a sizable gap in both phrase grounding and WWbL.

## 2. Related Work

In the weakly supervised phrase localization task, text phrases are associated with specific image locations [1, 16, 39]. Text embedding is often extracted from a pretrained language model and is aligned with an image representation to obtain a shared semantic domain [11, 17, 30]. Recent contributions employ CLIP [28] for linking text with image locations [20, 32]. Self-supervision methods were also used to obtain an initial map by means of explainability [3, 32]. However, we are unaware of any other phrase-grounding method that employs pseudo labels that are extracted from a phrase-grounding network.

Since our method employs the result of a phrase grounding model to collect data that is used to finetune the same model, our work falls under the category of self-training. Self-training methods use a model to infer pseudo-labels for unlabeled or partly labeled samples and use the resulting samples to further improve the same model [2]. For example, domain adaptation of detection networks can be performed by labeling samples of the target domain using a network trained on the source domain [29]. Another example is a teacher-student framework in which the teacher provides pseudo labels that are as accurate as possible, and the student is further regularized by noise injection [38]. Self-training can also be applied to multiple networks that are trained jointly, as in the case of a semi-supervised semantic segmentation method that employs two differently initialized networks that produce labels to one another [10]. We are not aware of any other self-training work in which the labels are produced by analyzing the embedding space and not using the network's output.

The recently introduced WWbL task extends phrase grounding by generating both the text and the associated image regions. The WWbL method [32] employs a multi-step procedure on top of phrase grounding, which combines Selective Search [36] with the image captioning method BLIP [19] in order to provide captioning to objects in the scene and then apply phrase grounding to these captions. WWbL, therefore, provides a novel application for phrase grounding methods, which does not require any text input, and serves, in our case, as a way to test the downstream success of our method.

## 3. Method

Given an input text $t$ and an RGB image $I \in R^{3 \times W \times H}$, our method's starting point is the current state-of-the-art architecture for weakly supervised phrase grounding [32]. We then apply a refinement procedure in order to improve this model.

Let $g$ be the phrase grounding network of [32], i.e.,

$$M = g(I, Z_t(t)) \tag{1}$$

where the output mask $M \in \mathbb{R}^{W \times H}$ contains values between 0 and 1, and $Z_t$ is the text encoder of CLIP [27], with an output dimension of 512.

Network $g$ is based on an encoder-decoder architecture adapted to support text-based conditioning. The image encoder of $g$, called $Z_I$, is a VGG network [33] with a receptive field of size $16 \times 16$, i.e., the image is downscaled four times. The output of $Z_I$ has 512 channels, to match the dimensions of the text encoder $Z_t$.

The map $Z_I(I) \in \mathbb{R}^{W/16 \times H/16 \times 512}$ is viewed as a spatial field of vectors in $\mathbb{R}^{512}$. Each such vector is normalized to have a norm of one. To link the image information with the text information, a dot product is computed between the normalized vector at each location in the map and the vector $Z_t(t)$ (CLIP also has normalized outputs). Following this cosine similarity step, a map $Z_s(I, t)$ is obtained with values between -1 and 1. $Z_s(I, t)$ has a single channel and the same spatial dimensions as $Z_I(I)$.

The next step of [32] is to spatially re-weight all channels of $Z_I(I)$ by multiplying them pointwise by $Z_s(I, t)$. The result is passed to the decoder $U$, which consists of three upsampling blocks, to obtain the single channel map $M$ of Eq. 1. This map is obtained after applying a Sigmoid and has values between 0 and 1.

Our proposed solution for phrase-grounding refinement is based on calculating self-similarity maps based on the data of the spatial map $Z_I(I)$. These maps are aggregated and serve as pseudo-labels for fine-tuning $g$ and obtaining a fine-tuned network $g^{++}$.

Let $J$ be the map that is obtained after normalizing $Z_I(I)$ along the channel index. A set of $W/16 \times H/16$ maps is

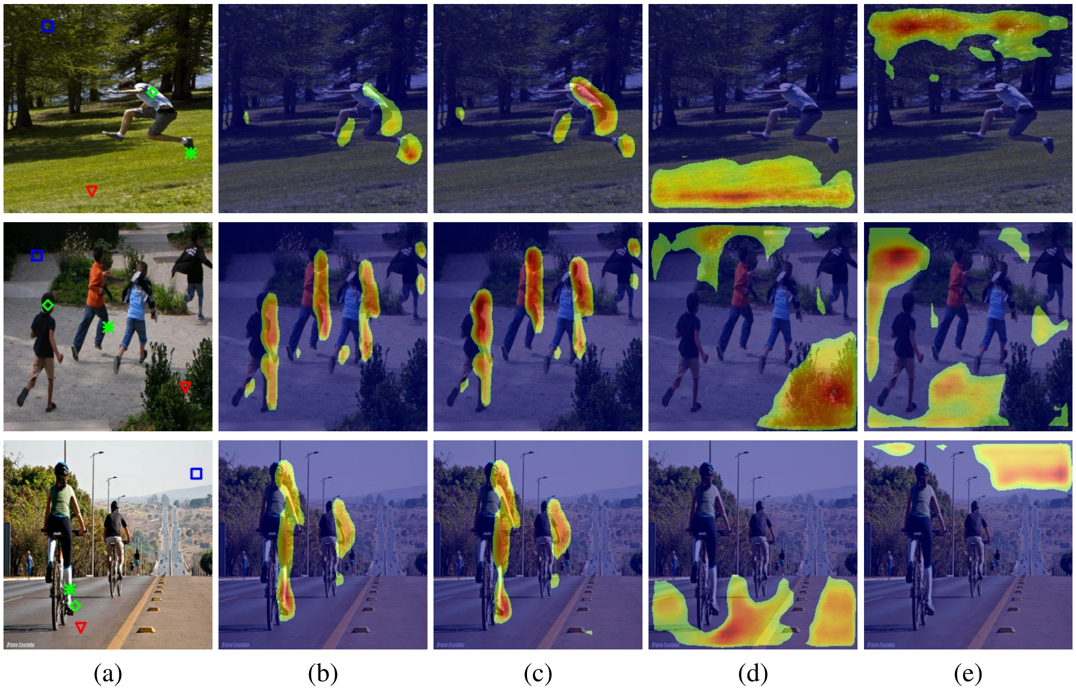


**Figure 1.** Sample self-similarity maps for given image locations. (a) input image. (b-c) self-similarity maps for the green star and green diamond image locations (resp.), which are inside the same object. (d-e) self-similarity maps for the locations marked by a red triangle and a blue square, respectively. Evidently, the similarity map for an image location highlights the object in that region, and two similarity maps that are associated with image locations on the same object produce similar maps.

obtained by considering every spatial location of $J$. Specifically, let $j_{x,y} \in \mathbb{R}^{512}$ denote the vector of all channels at one specific map location $(x, y)$. The similarity map $S^{x,y}$ has the same spatial dimensions as $J$ and at location $(x', y')$ has the scalar value $j_{x,y} \cdot j_{x',y'}$.

$S^{x,y}$ are text-agnostic as the text representation $Z_t(t)$ does not play a role in their creation. Sample maps, for specified image locations, are depicted in Fig. 1. As can be seen, the similarity map $S^{x,y}$ is consistent with the object in the corresponding image location $(16x, 16y)$ (upsampling four times, due to the encoder's downsampling effect).

This is not surprising, since, within the spatial region that is associated with a single object, the cosine similarity between $Z_I(I)$ and $Z_t(t)$ is expected to be similar, i.e., high if that image location is described by the text and low otherwise. Therefore, the embeddings of two spatial locations from the same object-associated region would be high. Conversely, the embeddings at image locations that are associated with different objects are expected to be low.

Next, our method filters the set of computed maps $\{S^{x,y}\}_{x,y}$ in order to identify the maps that are relevant to a given input text $t$. This is done by comparing the output map $M = g(I, Z_t(t))$ to each of the maps in the set.

Specifically, each map $S^{x,y}$ is binarized by threshold at a value of zero (recall that the values of the similarity maps are cosine similarities between -1 and 1). Similarly, the map $M$ is binarized using a threshold that is half of the map's maximal value. Since $M$ is in the size of the original image, its binarized version is downscaled by a factor of 16 in each dimension to match that of $S^{x,y}$.

The Intersection Over Union (IOU) score is then computed between the two binary maps, the one based on $S^{x,y}$ and the one derived from $M$. If this score is above a relatively conservative threshold of $\tau = 0.6$, the map $S^{x,y}$ is considered relevant to the text $t$. Finally, all relevant maps are averaged to obtain the map $\bar{M}$. In the case that no location $x, y$ produced a map $S^{x,y}$ that leads to an IOU above $\tau$, we select the map with the maximal IOU score.

Fig. 2 presents examples of the mask aggregation process, which extracts the mask $\bar{M}$ based on the text, image, and phrase grounding network $g$. As can be seen, the obtained masks $\bar{M}$ are often much more comprehensive than the original map $M$, without covering significantly more of the background regions. The appendix includes a pseudo-code outlining the steps involved in our approach.

The finetuning of $g$ to obtain the refined network $g^{++}$ employs the map $\bar{M}$ as the pseudo-label. Four-loss terms are used, given the input image $I$ and the text $t$. The pseudo-supervised loss is given by

$$L_{pseudo}(I, t, \bar{M}) = \|\bar{M} - g^{++}(I, Z_t(t))\|^2 \,, \qquad (2)$$

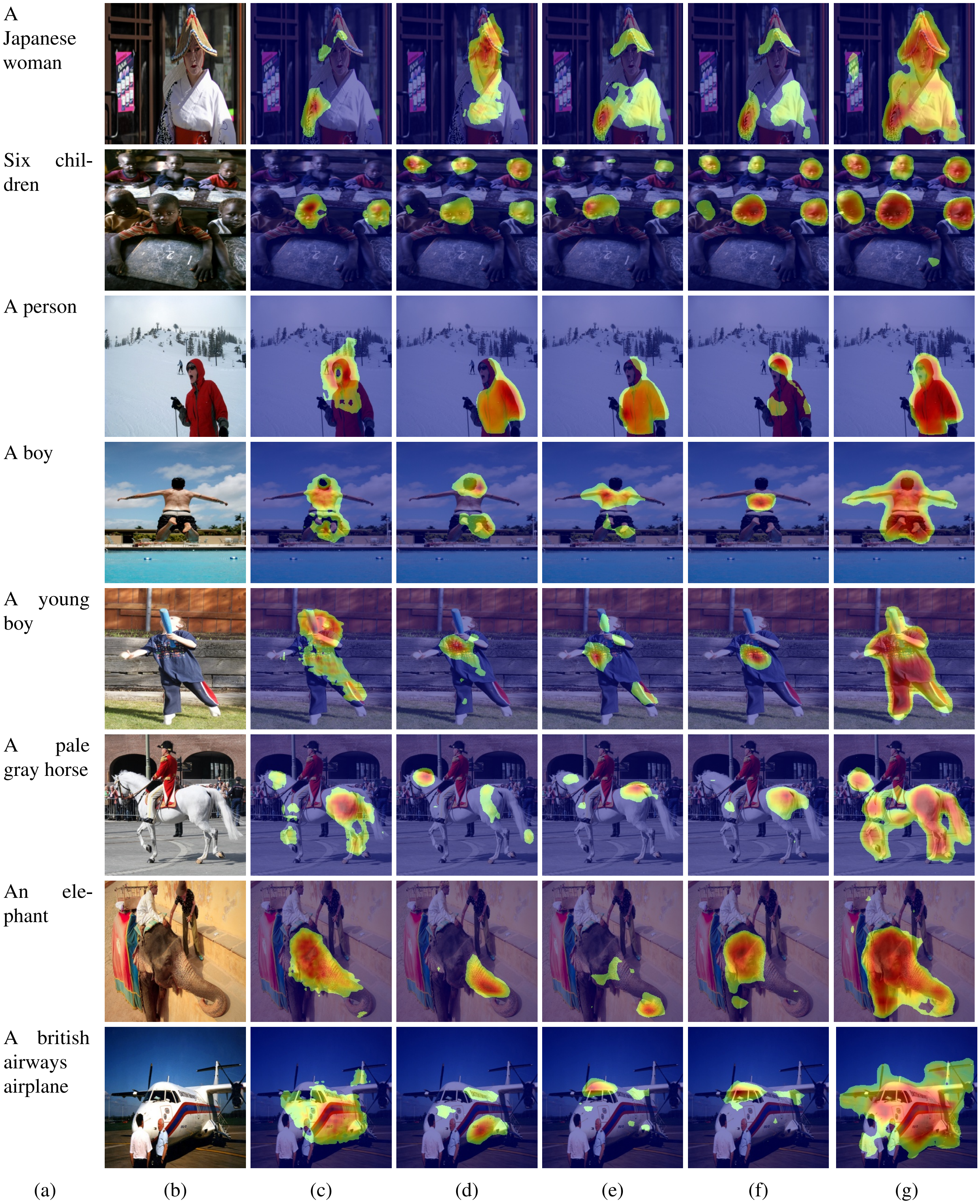


**Figure 2.** Sample self-similarity maps and their aggregation. (a) the input text $t$. (b) the input image $I$. (c) the output $M = g(I,t)$. (d-f) three self-similarity maps $S^{x,y}$ that overlap $M$. (g) the aggregated map $\bar{M}$. Evidently, the aggregated map is much more comprehensive.

where $\bar{M}$ is the pseudo label obtained from $g$ when applying it to image $I$ and text $t$.

The other three loss terms follow [32]. The foreground loss $L_{fore}(I, t)$ is given by

$$L_{fore}(I,t) = -CLIP(g^{++}(I, Z_t(t)) \odot I, t), \tag{3}$$

where $\odot$ is a pointwise multiplication. This loss maximizes the CLIP similarity between the foreground region of the obtained mask $g^{++}(I, Z^T)$ and the text $t$.

The background loss $L_{back}(I, t)$ is

$$L_{back}(I,t) = CLIP((1 - g^{++}(I, Z_t(t))) \odot I, t). \tag{4}$$

This loss minimizes the CLIP similarity between the background of image $I$, as denoted by the obtained mask, and the text $t$.

Finally regularization loss $L_{reg}(I)$ is applied to create compact maps

$$L_{reg}(I,t)) = \|g^{++}(I, Z_t(t))\| \tag{5}$$

The combined loss is defined as an unweighted sum of the four loss terms, in order to avoid adding hyper-parameters, i.e., $L(I, t, \bar{M}) = L_{pseudo}(I, t, \bar{M}) + L_{fore}(I, t) + L_{back}(I, t) + L_{reg}(I, t)$.

# 4. Experiments

We present our results for both weakly supervised phrase grounding (WSPG) and WWbL. The networks $g$ and $g^{++}$ are trained on either MSCOCO 2014 [21] or the Visual Genome (VG) dataset [18]. Evaluation is done on the test splits of Flickr30k [26], ReferIt [9, 14], and VG.

The training split of Akbari et al. [1] is used for MSCOCO 2014. It consists of 82,783 training images and 40,504 validation images, each image with five captions describing it. VG contains 77,398 training, 5000 validation, and 5000 test images. Each image is associated with both free-form text and annotated bounding boxes.

The Flickr30k Entities dataset [26], which is derived from Flickr30k, contains 224K phrases describing objects in more than 31K images. Each image is associated with five captions. For evaluation, we use the 1k images of the test split of Akbari et al [1]. ReferIt [9, 14] consists of 20K images and 99,535 manually captioned and segmented images. This data was collected in a two-player game with approximately 130K isolated entity descriptions. Here, too, we use the test split of Akbari et al. [1].

**Implementation details** The proposed network $g^{++}$ is obtained by fine-tuning network $g$ with the loss terms presented at Sec. 3. The visual backbone of $g$ is VGG16 [33], which is also used by the vast majority of the phrase grounding work.

Finetuning takes place using an SGD optimizer with a batch size of 32 and a learning rate of 0.0001. This learning rate is 3 times lower than the original learning rate of $g$. The optimizer momentum is 0.9 and weight decay regularization is 0.0001. Finetuning runs for 3000 iterations.

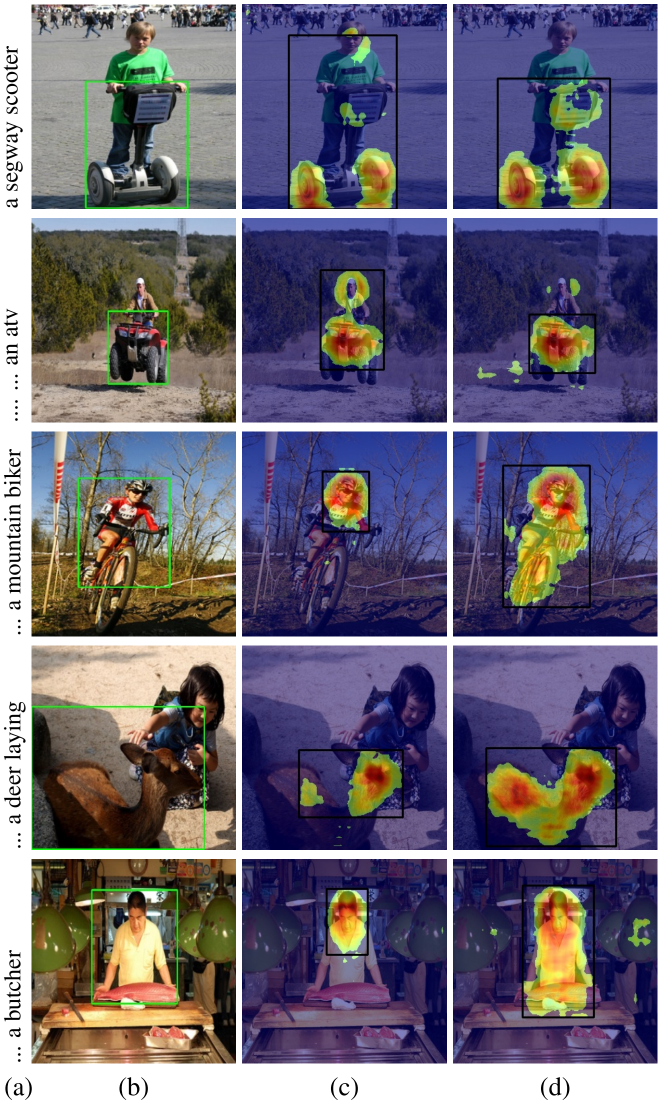
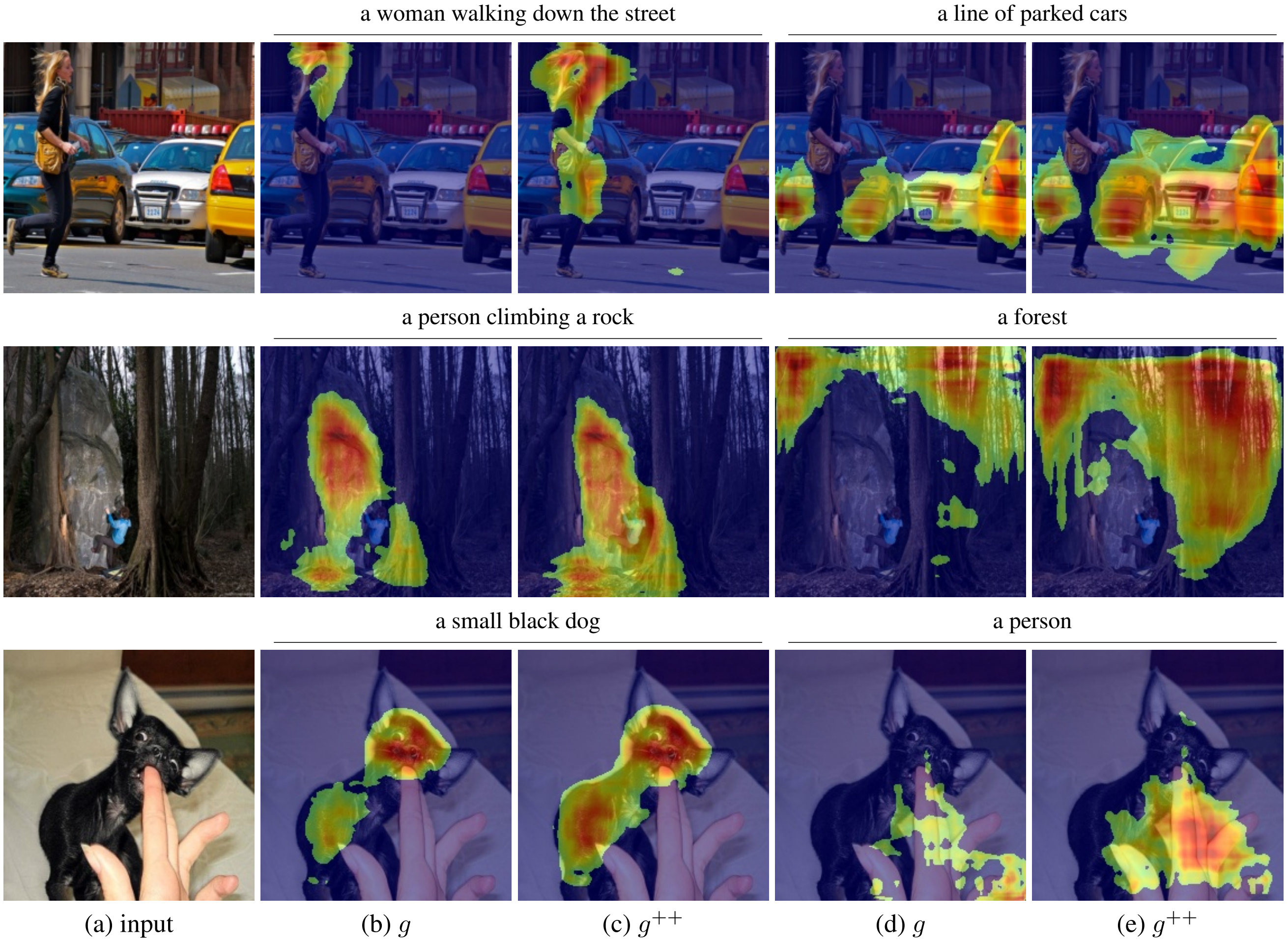


**Figure 3.** Sample phrase grounding results. (a) the phrase. (b) the input image and the ground truth bounding box. (c) results for network $g$ [32] shown as a heatmap. (d) same for the refined network $g^{++}$.

**Results** Phrase grounding tasks are evaluated with respect to the accuracy of the pointing game [39], which is calculated from the output map by finding the maximum-value location for the given query and checking whether this point is located in the region of the object.

Another metric shown ("BBox accuracy") extracts a bounding box from the output mask and compares it with the bounding-box annotations. An accurate prediction is defined as one for which the IOU is over 0.5. To extract the bounding



(a) input (b) $g$ (c) $g^{++}$ (d) $g$ (e) $g^{++}$

**Figure 4.** Samples WWbL results on images from Flickr1K comparing between $g$ [32](b,d) and our $g^{++}$(c,e). The automatically generated captions (these are independent of the phrase grounding network used) are on top of the relevant masks.

| Method | VG trained | | | MS-COCO trained | | |
|---|---|---|---|---|---|---|
| | VG | Flickr | ReferIt | VG | Flickr | ReferIt |
| FCVC [12] | - | - | - | 14.03 | 29.03 | 33.52 |
| VGLS [37] | - | - | - | 24.40 | - | - |
| TD [39] | 19.31 | 42.40 | 31.97 | - | - | - |
| SSS [16] | 30.03 | 49.10 | 39.98 | - | - | - |
| MG-BiLSTM [1] | 50.18 | 57.91 | 62.76 | 46.99 | 53.29 | 47.89 |
| MG-ELMo [1] | 48.76 | 60.08 | 60.01 | 47.94 | 61.66 | 47.52 |
| GbS [3] | 53.40 | 70.48 | 59.44 | 52.00 | 72.60 | 56.10 |
| CLIP+GAE [7] | 54.72 | 72.47 | 56.76 | 54.72 | 72.47 | 56.76 |
| g [32] | 62.31 | 75.63 | 65.95 | 59.09 | 75.43 | 61.03 |
| $g^{++}$ (ours) | **66.63** | **79.95** | **70.25** | **62.96** | **78.10** | **61.53** |

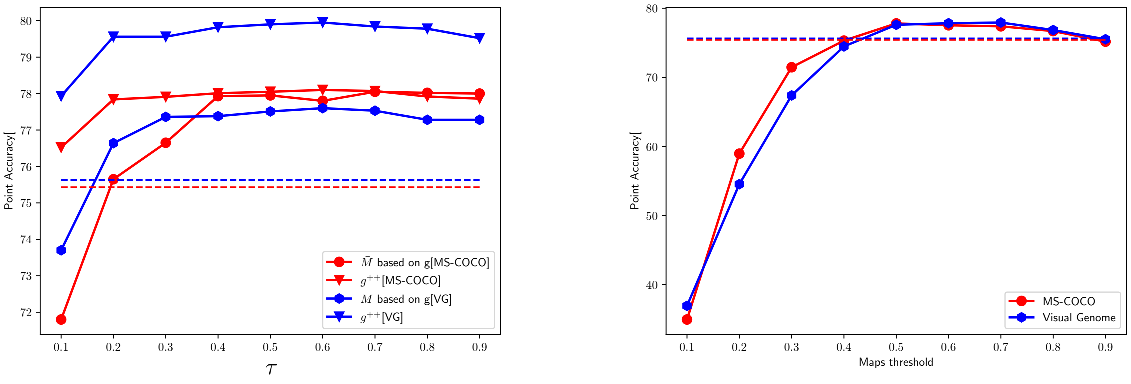
**Table 1.** Weakly Supervised Phrase Grounding (WSPG) results, showing the "pointing game" accuracy for multiple benchmarks.

box from an output map $M$, the procedure of Shaharabany et al. [32] is employed. First, a threshold of 0.5 is applied to M, and then the method of Suzuki et al. [34] is applied to find image contours. A set of bounding boxes is obtained by considering the enclosing box of each contour. These bounding boxes are scored by summing the values of $M$ within the contour and those with low scores are discarded. Following a non-maximal suppression step, the minimal bounding box that contains the remaining bounding boxes is used.

As mentioned, we use the same training/validation/test splits as Akbari et al. [1]. For ReferIt, Visual Genome and Flickr30K, each query is treated as a single sentence.

Tab. 1 lists the results for the Flickr30k, ReferIt, and VG for the weakly-supervised phrase grounding task. Evidently, our method is superior to all baselines, whether training takes place over VG or MS-COCO. In particular, there is a sizable improvement with respect to the network $g$ that we refine, which is currently the state-of-the-art [32].

In addition to the pointing game results, Tab. 2 presents bounding box accuracy for the phrase grounding task (this data is not available for most baselines). Here, too, our



(a) (b)

**Figure 5.** A parameter sensitivity analysis presenting the pointing game score for the Flickr dataset. (a) sensitivity to parameter $\tau$, in which the default value is 0.6, and (b) varying the threshold on the map $M$ as a fraction of the maximal value (default is 0.5). Shown in panel (a) is the performance of using $\bar{M}$ instead of $M$ of $g$ (same as the first ablation) and the performance of the full $g^{++}$ refinement. In (b), due to lack of time, only the former appears. The dashed horizontal lines denote the performance of the baseline model $g$.

| Task | Model | VG Trained | | | | | | MS-COCO Trained | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Point Accuracy | | | Bbox Accuracy | | | Point Accuracy | | | Bbox Accuracy | | |
| | | VG | Flickr | ReferIt | VG | Flickr | ReferIt | VG | Flickr | ReferIt | VG | Flickr | ReferIt |
| WWbL | MG [1] | 32.15 | 49.48 | 38.06 | 12.23 | 24.79 | 16.43 | 32.91 | 50.12 | 36.34 | 11.48 | 23.75 | 13.31 |
| | g [32] | 43.91 | 58.59 | 44.89 | 17.77 | 31.46 | 18.89 | 44.20 | 61.38 | 43.77 | 17.76 | 32.44 | 21.76 |
| | $g^{++}$ (ours) | **45.90** | **62.98** | **45.14** | **20.01** | **33.71** | **21.07** | **47.39** | **65.93** | **44.52** | **20.58** | **36.40** | **22.07** |
| WSPG | MG [1] | 48.76 | 60.08 | 60.01 | 14.45 | 27.78 | 18.85 | 47.94 | 61.66 | 47.52 | 15.77 | 27.06 | 15.15 |
| | g [32] | 62.31 | 75.63 | 65.95 | 27.26 | 36.35 | 32.25 | 59.09 | 75.43 | 61.03 | 27.22 | 35.75 | 30.08 |
| | $g^{++}$ (ours) | **66.63** | **79.95** | **70.25** | **30.95** | **45.56** | **38.74** | **62.96** | **78.10** | **61.53** | **29.14** | **46.62** | **32.43** |
| WSPG ablations | $\bar{M}$ based on $g$ | 63.10 | 77.60 | 66.61 | 24.07 | 26.40 | 33.33 | 61.19 | 77.80 | 61.15 | 21.56 | 22.17 | 27.41 |
| | Only $L_{pseudo}$ | 65.50 | 78.84 | 68.49 | 23.50 | 39.06 | 29.16 | 62.37 | 78.07 | 60.15 | 22.10 | 40.12 | 26.62 |
| | $L_{pseudo} + L_{reg}$ | 59.40 | 73.95 | 64.31 | 22.35 | 26.25 | 26.25 | 56.97 | 74.99 | 60.03 | 19.94 | 22.22 | 23.55 |
| | $L_{pseudo} + {}^{1}\!/_{3}L_{reg}$ | 65.80 | 78.94 | 68.68 | 30.03 | 43.46 | 37.27 | 60.44 | 76.81 | 58.83 | 27.05 | 44.99 | 30.89 |
| | $L_{pseudo} + L_{fore} + L_{back}$ | 65.47 | 79.51 | 69.77 | 25.71 | 44.96 | 34.29 | 62.61 | 78.05 | 60.86 | 25.51 | 45.90 | 30.66 |
| | No aggregation | 66.22 | 79.24 | 70.03 | 27.36 | 44.44 | 35.71 | 61.72 | 78.02 | 59.55 | 27.28 | 46.34 | 31.42 |
| | Segmentation encoder | 56.52 | 73.26 | 61.22 | 19.72 | 20.37 | 23.91 | 56.53 | 74.25 | 59.12 | 19.78 | 21.65 | 22.52 |
| | Classification encoder | 60.49 | 74.40 | 66.67 | 4.85 | 4.07 | 16.50 | 55.91 | 72.84 | 63.08 | 4.67 | 4.44 | 15.41 |

**Table 2.** WWbL and Weakly Supervised Phrase Grounding (WSPG) results for the test sets showing both “pointing game" accuracy and bounding box accuracy. The ablations show the performance of various reductions of our method on the WSPG task (see text for details).

method outperforms the baseline methods by a wide margin. Sample results are shown in Fig. 3, compared with $g$ [32].

The WWbL task is an open-world localization task, in which the only input is a given image: given an input image, the goal is to localize and describe all the elements composing the scene. To solve this, Shaharabany et al. [32] employ a multi-stage algorithm, in which selective search [36] is used to extract object proposals and BLIP is used to caption these regions. Similar captions are pruned using the Community Detection (Cd) method [4]. The phrase grounding model then assigns heatmaps to the filtered list of captions. Since phrase grounding is applied last, WWbL provides a direct way to compare phrase grounding methods on automatically generated captions.

The WWbL task is evaluated with the same two metrics (pointing game and bounding-box accuracy). For each ground-truth pair of bounding box and caption, the closest caption in CLIP space from the list of automatically generated captions is selected. The associated output map of the phrase grounding method is then compared to the ground truth bounding box using the pointing accuracy metric. In addition, bounding boxes are extracted for the output heatmaps, as described above.

WWbL results are listed in Tab. 2, with the leading WSPG results provided for reference. The table shows the performance obtained by $g$, $g++$, and a baseline that employs the phrase grounding method MG [1] as part of the same WWbL captioning procedure described above. Evidently, $g^{++}$ obtains the best results among the three across all benchmarks and the two scores. Fig. 4 presents samples of the results of $g^{++}$, compared to those of $g$ [32] for images from Flickr1K.

**Ablation Study** We present multiple ablations, each validating a different part of our approach. The results of these experiments are provided for the phrase grounding task and are also listed in Tab. 2.

The first ablation considers the pseudo-label generation process but instead of using it to finetune $g$ to obtain $g^{++}$, it uses it as the output of the phrase grounding task. In other words, this ablation directly provides $\bar{M}$ as the output. As can be seen, using $\bar{M}$ of $g$ improves upon $g$ in the pointing game accuracy but not in the bounding box accuracy. Moreover, it scores below our complete solution $g^{++}$ on all scores across all benchmarks.

The second ablation employs only the loss of the pseudo label $L_{psudo}$ but not the other three-loss terms. As a general trend, we can see that it improves the results of $g$ and obtains much of the benefit of $g^{++}$ (not on all datasets when considering the bounding box accuracy).

Both these ablations may indicate that the other three loss terms, including the CLIP-based loss terms $L_{fore}$ and $L_{back}$ and the regularization loss, help maintain a reasonable bounding box for the resulting mask. To check whether the regularization loss can elevate $L_{psuedo}$ to the level of $g^{++}$ performance, we conducted an ablation with both $L_{pseudo} + L_{reg}$. As can be seen, the addition of the regularization term to the pseudo-label loss hurts performance.

Since the regularization term in this experiment is balanced by a single loss (and not three), this could indicate that too much regularization takes place. We, therefore, repeat the experiment with a third of the regularization, which seems to improve, on most benchmarks and scores, above using only $L_{psuedo}$. Lastly, using the three other loss terms, without regularization leads to a performance that is sometimes better and sometimes worse than adding a third of the regularization term to the self-training loss, and still below that of the full method.

The next ablation "No aggregation" selects the similarity map $S^{x,y}$ that has the highest IOU with $M$ instead of averaging multiple maps that are similar to $M$. The rest of the training pipeline of $g^{++}$ is kept the same. Evidently, this ablation maintains much of the advantages of $g^{++}$ over other phrase grounding methods but obtains slightly lower accuracy than the complete method.

The last set of ablations employs an alternative encoder instead of the image encoder of $Z_I$. The purpose of these experiments is to check whether the emergence of segmentation maps of the latent space is a unique characteristic of the phrase grounding network we employ or if it is a general phenomenon. Specifically, we (1) employ the encoder of a Fully-Convolutional semantic segmentation network [22] with a ResNet-50 [15] backbone, which is also trained on the MS-COCO dataset, and (2) employ a latent representation of a VGG16 classification network trained on ImageNet (the same backbone of $g$). As can be seen from the ablation results, these two alternative encoders do not lead to an improvement in the refined network over the baseline network $g$. Sample self-similarity maps obtained from the alternative encoders are provided in the supplementary appendix, showing that the alternative maps do not provide the level of object-delineation provided by $Z_I$.

**Parameter sensitivity** Our method has very few parameters. There is a threshold on the cosine similarity values that is taken at the middle of the range, i.e., at zero, and a threshold on $M$ that is taken at half the maximal value. Both of which are natural choices. Additionally, there is a threshold $\tau$ on the IOU between the binarized $M$ and the binarized $S^{x,y}$ that is set to 0.6. This number is often used as a relatively conservative threshold for IOU.

Fig. 5 presents results when varying the binarization parameters. These are shown for $\bar{M}$ based on $g$ (same as the first ablation), which is a direct way to measure the quality of $\bar{M}$, and for the refined $g^{++}$ (the latter is shown only for $\tau$ due to lack of time). Our method is largely insensitive to both these thresholds, obtaining similar performance scores in a relatively wide range of values.

## 5. Conclusions

Image encoders provide a rich spatial representation of the input image. Despite the widespread use of such encoders, we are unaware of other contributions that apply image-processing techniques to this embedding space in order to obtain spatial information. Our work demonstrates that this is an oversight as the information in the embedding tensor is readily available for further exploitation. By using the informative correlations between the data in various image locations, we are able to extract segmentation pseudo-labels that are extremely beneficial in advancing the performance in the phrase grounding task beyond the state-of-the-art.

## Acknowledgments

This work was supported by a grant from the Tel Aviv University Center for AI and Data Science (TAD).

## 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, 5, 6, 7, 8

[2] Massih-Reza Amini, Vasilii Feofanov, Loic Pauletto, Emilie Devijver, and Yury Maximov. Self-training: A survey. *arXiv preprint arXiv:2202.12040*, 2022. 2

[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, 6

[4] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008. 7

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021. 1

[6] Hila Chefer, Sagie Benaim, Roni Paiss, and Lior Wolf. Image-based clip-guided essence transfer, 2021. 1

[7] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 397–406, October 2021. 6

[8] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 782–791, 2021. 1

[9] Kan Chen, Rama Kovvuri, and Ram Nevatia. Query-guided regression network with context policy for phrase grounding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 824–832, 2017. 5

[10] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2613–2622, 2021. 2

[11] Jason PC Chiu and Eric Nichols. Named entity recognition with bidirectional lstm-cnns. *Transactions of the association for computational linguistics*, 4:357–370, 2016. 2

[12] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1473–1482, 2015. 6

[13] Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. 1

[14] Michael Grubinger, Paul Clough, Henning Müller, and Thomas Deselaers. The iapr tc-12 benchmark: A new evaluation resource for visual information systems. In *International workshop ontoImage*, volume 2, 2006. 5

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 8

[16] Syed Ashar Javed, Shreyas Saxena, and Vineet Gandhi. Learning unsupervised visual grounding through semantic self-supervision. *arXiv preprint arXiv:1803.06506*, 2018. 2, 6

[17] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021. 2

[18] 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. 5

[19] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022. 1, 2

[20] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. *arXiv preprint arXiv:2112.03857*, 2021. 2

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, volume 8693 of *LNCS*, pages 740–755, 2014. 5

[22] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 8

[23] Oscar Jarillo Michel, Roi Bar-On, Richard Liu, Sagie Benaim, and Rana Hanocka. Text2mesh: Text-driven neural stylization for meshes. *ArXiv*, abs/2112.03221, 2021. 1

[24] Roni Paiss, Hila Chefer, and Lior Wolf. No token left behind: Explainability-aided image classification and generation. In *European Conference on Computer Vision*, pages 334–350. Springer, 2022. 1

[25] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021. 1

[26] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 5

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 1, 2

[28] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 2

[29] Aruni RoyChowdhury, Prithvijit Chakrabarty, Ashish Singh, SouYoung Jin, Huaizu Jiang, Liangliang Cao, and Erik Learned-Miller. Automatic adaptation of object detectors to new domains using self-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 780–790, 2019. 2

[30] Justyna Sarzynska-Wawer, Aleksander Wawer, Aleksandra Pawlak, Julia Szymanowska, Izabela Stefaniak, Michal Jarkiewicz, and Lukasz Okruszek. Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Research*, 304:114135, 2021. 2

[31] 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 *ICCV*, pages 618–626, 2017. 1

[32] Tal Shaharabany, Yoad Tewel, and Lior Wolf. What is where by looking: Weakly-supervised open-world phrase-grounding without text inputs. *arXiv preprint arXiv:2206.09358*, 2022. 1, 2, 5, 6, 7, 8

[33] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2, 5

[34] Satoshi Suzuki et al. Topological structural analysis of digitized binary images by border following. *Computer vision, graphics, and image processing*, 30(1):32–46, 1985. 6

[35] Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. Zero-shot image-to-text generation for visual-semantic arithmetic. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 1

[36] J.R.R. Uijlings, K.E.A. van de Sande, T. Gevers, and A.W.M. Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 2013. 2, 7

[37] 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. 6

[38] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[39] 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, 5, 6

[40] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *arXiv preprint arXiv:2109.01134*, 2021. 1

[41] Peihao Zhu, Rameen Abdal, John C. Femiani, and Peter Wonka. Mind the gap: Domain gap control for single shot domain adaptation for generative adversarial networks. *ArXiv*, abs/2110.08398, 2021. 1