Detector-Free Weakly Supervised Grounding by Separation

Assaf Arbelle*[1], Sivan Doveh*[1,4], Amit Alfassy*[1,6], Joseph Shtok[1], Guy Lev[1], Eli Schwartz[1,3],
Hilde Kuehne[1], Hila Barak Levi[4], Prasanna Sattigeri[1], Rameswar Panda[1,2], Chun-Fu Chen[1,2],
Alex Bronstein[6], Kate Saenko[2,5], Shimon Ullman[4], Raja Giryes[3], Rogerio Feris[1,2], Leonid Karlinsky[1]
IBM Research[1], MIT-IBM Watson AI Lab[2], Tel-Aviv University[3],
Weizmann Institute of Science[4], Boston University[5], Technion[6]

Abstract

Nowadays, there is an abundance of data involving images and surrounding free-form text weakly corresponding to those images. Weakly Supervised phrase-Grounding (WSG) deals with the task of using this data to learn to localize (or to ground) arbitrary text phrases in images without any additional annotations. However, most recent SotA methods for WSG assume an existence of a pre-trained object detector, relying on it to produce the ROIs for localization. In this work, we focus on the task of Detector-Free WSG (DF-WSG) to solve WSG without relying on a pre-trained detector. The key idea behind our proposed Grounding by Separation (GbS) method is synthesizing 'text to image-regions' associations by random alpha-blending of arbitrary image pairs and using the corresponding texts of the pair as conditions to recover the alpha map from the blended image via a segmentation network. At test time, this allows using the query phrase as a condition for a non-blended query image, thus interpreting the test image as a composition of a region corresponding to the phrase and the complement region. Our GbS shows an 8.5% accuracy improvement over previous DF-WSG SotA, for a range of benchmarks including Flickr30K, Visual Genome, and ReferIt, as well as a complementary improvement (above 7%) over the detector-based approaches for WSG.

1. Introduction

As multi-modal text + images data sources become abundant, so grows the importance of natural free-form text supervision [57] over the more traditional image labels or image bounding boxes annotation methods. Such multi-modal data (i.e. image-text pairs) can be almost effortlessly and autonomously collected from web pages and documents with illustrations, user captioned personal photos, transcribed videos, and many more. However, such form of automatic supervision poses significant challenges for learning. First, it is noisy in a sense that some of the text words are not relevant to the image; second, it is not well localized in a sense that it is unknown which parts of the image correspond to which parts of the text. In contrast, in traditional annotation the training signal is highly localized: isolated and cropped object images are commonly used in classification, and bounding boxes or polygons around the objects in detection and/or segmentation. However, these annotations are commonly manual and are costly to collect.

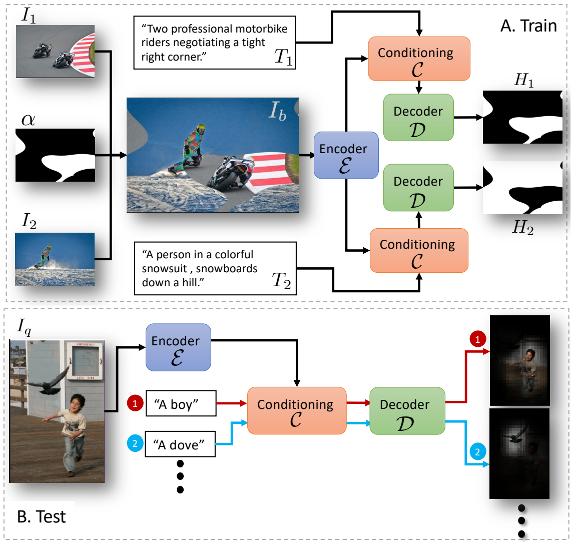
A. Train
I_1
"Two professional motorbike riders negotiating a tight right corner."
T_1
α
I_b
I_2
"A person in a colorful snowsuit , snowboards down a hill."
T_2
Encoder $\mathcal{E}$
Conditioning $\mathcal{C}$
Decoder $\mathcal{D}$
Decoder $\mathcal{D}$
Conditioning $\mathcal{C}$
H_1
H_2
B. Test
I_q
Encoder $\mathcal{E}$
1 "A boy"
2 "A dove"
Conditioning $\mathcal{C}$
Decoder $\mathcal{D}$
1
2

Figure 1. Illustration of our compositional approach. (a) The model is trained to decompose random alpha-blendings of pairs of images conditioned on their associated texts; (b) At test time, the model interprets any image as a composition of two image regions, related and unrelated to the conditioning query phrase, thus grounding the phrase to the image pixels.

The above discussion highlights the importance of weakly (and autonomously) supervised multi-modal (images + text) learning in general, and Weakly Supervised Grounding (WSG) in particular. In WSG, the model is expected to learn to localize (highlight) image regions corresponding to text phrases. In a sense, WSG is a detection task where the traditional 'noun object labels' are replaced

*Equal contribution

by an unbounded set of things describable using natural language. Moreover, the WSG model is expected to learn from image + free-form corresponding text (e.g. caption) pairs without any annotations for correspondence of text words or phrases to image regions.

While earlier WSG methods [1, 28, 71, 78] were 'detector-free', all the more recent state-of-the-art (SotA) methods rely on the existence of pre-trained object detectors being the source of the localization RoIs for grounding [11, 23, 67, 8, 9, 46]. Although this 'detector-based' setup benefits from higher performance compared to Detector-Free WSG (DF-WSG) methods, in a sense it shifts away from the true WSG, as the detector is trained using bounding boxes (which are forbidden in WSG). The use of a detector is indeed plausible when the set of objects supported by the detector significantly overlaps the set of objects (nouns or their taxonomy siblings) appearing in the WSG texts. However, if we need to train for WSG in a different domain (e.g. news [43] or technical documents) or for a significantly different set of objects, we are likely to be required to collect a large set of bounding boxes to train a new detector. Experimental evidence for this appears, for example, in a recent detector-based WSG work [11][1], where it was noted that using the 80-categories COCO-trained detector for the Flickr30K and Visual Genome (VG) WSG benchmarks performs poorly, as opposed to their best WSG result obtained with the VG trained detector that supports many more relevant categories.

In this work, we propose an approach for WSG that does not rely on pre-trained detectors and thus addresses the DF-WSG task. Our approach is based on the idea of image and text compositionality. Having an image + corresponding text pair, we can consider the image as a composition of image regions glued together (like puzzle pieces) to form the whole image, each corresponding to a phrase of the text. While for a given single image + text pair the composition parts are not known (due to the WSG setting), we can easily simulate a more complex composition by comprising it from any two random image + text pairs. To do so, we can blend the images of the two pairs using a random alpha map α, thus making the respective texts of the pairs correspond to the known α and $1 - \alpha$ mapped complementary regions of the blended image. In this way, we can create a reliable localized synthetic training signal for the DF-WSG model that learns to perform text grounding by learning to separate the blended image to its α-mapped constituents conditioned on the respective texts (Figure 1a). At test time, we can apply the trained model on a non-blended query image, which when conditioned on the query phrase is expected to decompose the image to constituents related and not related to the conditioning text (Figure 1b). In addition to the separation loss, we further propose two regularization loss terms which are important for improving the model performance on non-blended test images. These losses help to prevent the model from learning blending artifacts, as well as to prevent the model from making incorrect references.

Our Grounding by Separation (GbS) approach obtains a significant, up to 8.5%, improvement over previous DF-WSG SotA [1] for a range of phrase grounding benchmarks including Flickr30K, Visual Genome, and ReferIt. Moreover, our performance on these benchmarks is not only comparable to the detector-based WSG SotA [11, 23, 46], it is also complementary to them, as our approach is 'detector-free' and thus may better support classes that are unknown at the time of detector training. As a result, an ensemble of ours and detector-based SotA methods [23, 46] improves the Flickr30K detector-based WSG result by over 7%, underlining the benefits of our proposed GbS approach in situations where a detector is available.

To summarize, our key contributions are as follows: (i) We propose a novel GbS approach for training DF-WSG models (WSG without assuming a pre-trained detector) based on learning to separate randomly blended images conditioned on the corresponding texts at train time, and applying the learned model on single images with arbitrary text phrase conditioning at test time; (ii) we establish a new SotA for the DF-WSG task improving significantly the previous best result by up to 8.5% over a range of popular phrase-grounding benchmarks: Flickr30K, VG, and ReferIt; (iii) we provide an extensive ablation study and examine the relative contribution of the components of the GbS method; (iv) we obtain a new absolute SotA in WSG on Flickr30K via an ensemble of our DF-WSG and the best detector based WSG model, significantly improving the previous SotA result by over 7%.

2. Related Work

Joint analysis of natural images and text is a basic component of many downstream tasks, such as image captioning [10, 30, 52, 65, 66, 74, 75, 82], text based image retrieval[9, 35, 45, 46, 68, 76], visual question answering [2, 3, 8, 20, 21, 72], text grounding[1, 6, 7, 11, 23, 56, 67], and other general purpose multi-modal learning [26, 38, 39, 40, 57, 62, 63, 64]. Below we review the text grounding and source separation topics, as the most relevant to our work.

Fully supervised text grounding. In the fully supervised grounding setting the training annotations include pairs of phrases and their corresponding image bounding box location. As in object detection, methods employing these detailed annotation train a Region Proposal Network (RPN) to produce image ROIs which are candidates for the grounding target. In [56] joint visual-textual representation space is used for matching the ROIs with the query phrase; instead, [49] generate text captions for representing each ROI; finally [6] and [7] slightly modify the task and leverage additional 'context' phrases describing parts of image unrelated

[1] please see the footnote on page 6 in [11]

to the query phrase, using them too for ROI matching.

Detector-based WSG. Most recent methods for WSG (requiring only free-form text captions as image level annotations) assume the availability of a pre-trained object detector, which performs the ROI localization. These methods generally aim to create a joint visual-textual representation space, thus transforming the grounding task into a retrieval task: find the ROI whose embedding best matches the query phrase embedding. In [11] cosine similarity between ROI embeddings and image caption embedding is maximized directly; [23] generate negative text samples using linguistic tools and employ them in a contrastive learning objective between ROIs and the (positive) caption; [67] match the query phrase to ROI labels produced by multiple pre-trained object detectors; finally, in a growing body of literature [9, 26, 38, 39, 40, 45, 46, 62, 63, 64, 82] transformers are used for learning task-agnostic visual-textual representation space where grounding is implemented via retrieval of detector generated image ROIs closest to the query phrase.

Detector-Free WSG (DF-WSG). As opposed to detector-based WSG methods, DF-WSG methods perform dense localization for a given query phrase, thus generating attention heatmaps as opposed to ranking ROIs. The "Pointing Game" accuracy measure [79] is commonly used for DF-WSG evaluation. Lacking any localization information, DF-WSG methods often define and optimize some auxiliary task on the weakly supervised data. While the auxiliary task is not identical to the grounding objective, optimization of the task leads to the desired phrase grounding results. In [71] joint text and image parsing is employed to enforce structural similarities between the attended image regions and the text parse-tree; [28] employ an attention mechanism to find the common image region among subsets of images which share a specific concept (noun) in the caption; [78] employ an image & video captioning model salience maps with respect to the query phrase. The current DF-WSG SotA [1], maximizes the likelihood of the caption words in a distribution of image features collected at multiple network depths (scales), as well as optimizing the likelihood of image features in a distribution defined by words in the learned (shared) embedding space.

Source-separation methods. Our approach to DF-WSG can be considered as doing source-separation, as we separate a randomly generated blended image to its original image sources (conditioned on the texts). Unconditioned source-separation has been explored extensively using the classical vision methods [4, 15, 27, 36, 37, 55]. Audio-visual cues have been used for the separation of speakers [5, 13, 47, 51], musical instruments [16, 19, 44, 80, 81], and general sounds [18, 59, 73]. MixUp [77] proposed random image blending for augmentation, and unconditioned visual source separation has been examined in [17, 29, 48, 34, 24, 83, 84]. To the best of our knowledge, no previous work has employed (text conditioned) source separation as an objective for learning to perform the text grounding task, as well as for text driven attention in general.

3. Method

Let $P_1 = (I_1, T_1)$ and $P_2 = (I_2, T_2)$ be two random image ($I_\times$) + text ($T_\times$) pairs from the DF-WSG task training data. Assume w.l.o.g. that the images are of the same size ($|I_1| = |I_2|$) and let α be a random alpha-map of this size: $|\alpha| = |I_1|$ and $\alpha = \{0 \leq \alpha_{i,j} \leq 1 | 1 \leq i,j \leq |I_1|\}$. Let the blended image $I_b = \alpha \cdot I_1 + (1-\alpha) \cdot I_2$ be a per-pixel convex combination of I_1 and I_2, and let $\mathcal{M}(\mathcal{I}, \mathcal{T}) = \mathcal{H}$ be the GbS model we would like to train for the DF-WSG task, accepting an image $\mathcal{I}$ and text $\mathcal{T}$ as corresponding inputs and returning an output heatmap $\mathcal{H}$. This heatmap $\mathcal{H}$ is predicting the probability of each pixel of the image $\mathcal{I}$ to be related to the text $\mathcal{T}$, in a sense that the pixel belongs to the part of the image described by the text. Our idea is that while the linkage between text parts of T_1 and T_2 and the corresponding image regions of I_1 and I_2 is not known (due to the WSG setting), the association between the T_1 and T_2 components of the concatenated text $T_b = T_1 + T_2$ and the pixels of the blended image I_b (in a generated 'synthetic' pair (I_b, T_b)) is given by construction and can be used as a *synthetic* training signal for $\mathcal{M}$. Following this intuition we define our proposed GbS main objective (loss):

$$\mathcal{L}_{sep} = MSE(\mathcal{M}(I_b, T_1), \alpha) + MSE(\mathcal{M}(I_b, T_2), 1-\alpha) \tag{1}$$

where $MSE(x,y) = \frac{1}{|I_b|} \cdot \sum_{i,j} (x_{i,j} - y_{i,j})^2$ is the mean-square-error. In this formulation, the model $\mathcal{M}$ is learning to 'separate' the blended image I_b conditioned on the text. As mentioned above, any natural image can also be considered as an alpha blending of regions with different semantic meaning (e.g. an overlay of object segments, etc). According to this intuition, our goal is that following training, when provided with a random test image I_t and some corresponding query text T_q, computing $\mathcal{M}(I_t, T_q)$ would produce a heatmap H_q such that I_t could be considered as a result of a alpha-blending with H_q alpha-map between an image I_q corresponding entirely to T_q and the complement image $\hat{I_q}$ containing everything on I_t that is unrelated to T_q:

$$I_t = H_q \cdot I_q + (1 - H_q) \cdot \hat{I_q} \tag{2}$$

In the following sections, we provide the architecture specifics of the model we used in our experiments, as well as several additional regularization losses, namely $\mathcal{L}_{adv}$, $\mathcal{L}_{neg}$, and $\mathcal{L}_{i2t}$, which are introduced in sections 3.2, 3.3, and 3.4 respectively, and are instrumental to make the proposed construction work well in practice. Our overall loss $\mathcal{L}_{GbS}$ is the weighted sum of all of these losses:

$$\mathcal{L}_{GbS} = \mathcal{L}_{sep} + \gamma_{adv} \cdot \mathcal{L}_{adv} + \gamma_{neg} \cdot \mathcal{L}_{neg} + \gamma_{i2t} \cdot \mathcal{L}_{i2t} \tag{3}$$

An extensive ablation study examining our design choices

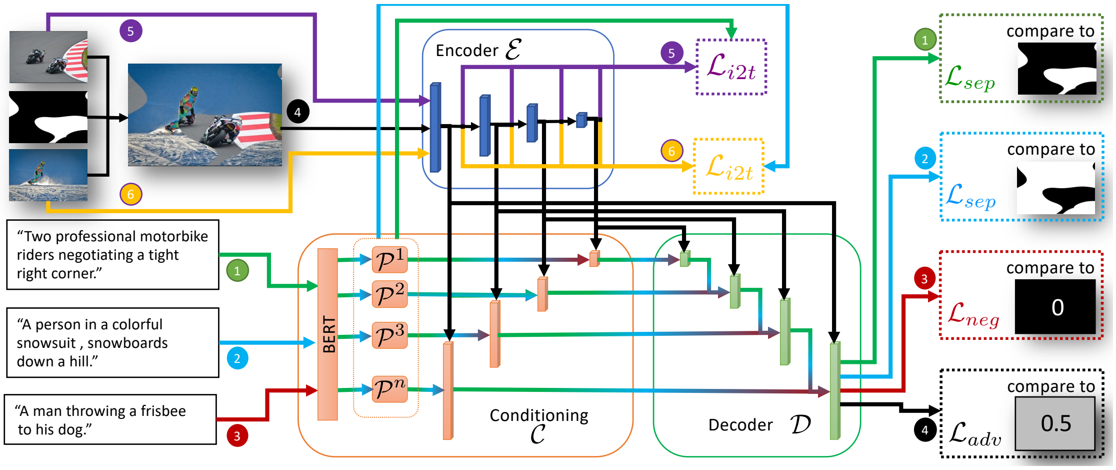

Figure 2. Detailed illustration of our model components and flow. Colored and numbered lines represent the flow of different inputs to the respective loss terms. The blended image (black line) flow from $\mathcal{E}$ to $\mathcal{D}$ on the way to the $\mathcal{L}_{adv}$ loss is direct and does not pass through $\mathcal{C}$.

is provided in section 4.4. The details of our GbS model are illustrated in Figure 2.

3.1. Model and text conditioning architecture

Our GbS model $\mathcal{M}(\mathcal{I},\mathcal{T}) = \mathcal{H}$ is comprised of an encoder $\mathcal{E}(\mathcal{I}) = E$, a text conditioning module $\mathcal{C}(E,\mathcal{T}) = C$, and a decoder $\mathcal{D}(C) = \mathcal{H}$ returning the final output:

$$\mathcal{H} = \mathcal{M}(\mathcal{I},\mathcal{T}) = \mathcal{D}(\mathcal{C}(\mathcal{E}(\mathcal{I}),\mathcal{T})) \qquad (4)$$

The encoder $\mathcal{E}(\mathcal{I}) = E$ is comprised of several (CNN) blocks with (stride r) pooling layers between them. We set $E = [E^1, \dots, E^n]$ to be a list of tensor outputs of n last blocks ordered in such a way that E^1 is the output of the last block (note that $|E^{i+1}| = r \cdot |E^i|$ due to pooling stride). **The text conditioning module** $\mathcal{C}(E,\mathcal{T}) = C$ is comprised of: (i) a text embedding model (e.g. BERT [12]) $\mathcal{N}(\mathcal{T}) = [W_1, ..., W_s]$ returning a list of word embeddings (in the context of the full text $\mathcal{T}$); followed by (ii) a projection module $\mathcal{P}^i(W_j) = W^i_j$, for each encoder $\mathcal{E}$ block $i \in [1, ..., n]$, adapting the word embeddings to the space of visual features of E^i; followed by (iii) averaging over the words $W^i = \frac{1}{s}\sum_j W^i_j$ to obtain the full text $\mathcal{T}$ embedding (again per block); and finally followed by (iv) the text attenuation module:

$$\mathcal{A}(E^i, W^i) = \exp\left(-\left|\frac{E^i}{||E^i||_2} - \frac{W^i}{||W^i||_2}\right|\right) \cdot E^i = C^i \qquad (5)$$

where all operations are element-wise, W^i is broadcasted to all the spatial locations of the tensor E^i, and the attenuation enhances locations of E^i that are closer to the projected text embedding for that block. The output of the text conditioning module is hence the per-block list: $C = [C^1, \dots, C^n]$. **The decoder module** $\mathcal{D}(C) = \mathcal{H}$ converts the text attenuated image encoding C into the final predicted heatmap $\mathcal{H}$. It is comprised of a series of ResNet [25] blocks $[D_1, \dots, D_n]$ such that the first block D_1 gets C_1 as the input: $O_1 = D_1(C_1)$, and similarly to U-Net [58] each subsequent block receives a combination of the up-scaled previous output and the input: $O_i = D_i(cat(C_i, U_r(O_{i-1})))$, where cat is channel-wise concatenation and U_r is the spatial up-scaling by the factor of r. We set the number of channels in O_i same as in C_i except for $O_n = \mathcal{H}$ which is the final output of the decoder $\mathcal{D}$ and has a single channel.

3.2. Unconditioned adversary loss: $\mathcal{L}_{adv}$

Naturally, images blended with a random alpha-map α differ from natural images and contain blending artifacts that could in turn be leveraged by the model $\mathcal{M}$ in order to produce the source separation. This is an unwanted behaviour that can increase the model's overfit to training data and decrease test performance. To reduce this effect we introduce an adversarial loss reducing the model's use of parameters that build upon these artifacts:

$$\mathcal{L}_{adv} = MSE(\mathcal{D}(\mathcal{E}(I_b)), 0.5 \cdot \mathbb{1}_{|\mathcal{H}|}) \qquad (6)$$

where $0.5 \cdot \mathbb{1}_{|\mathcal{H}|}$ stands for a uniform heatmap with 0.5 in all pixels indicating maximally uncertain prediction in case no text conditioning was provided.

3.3. Negative texts loss: $\mathcal{L}_{neg}$

We expect the model not only to produce correct α predictions matching the conditioning on the corresponding texts T_1 and T_2, but also to learn to 'reject' conditioning text that do not match. In other words, given a random unrelated text T_{neg} we want the model to produce close to zero prediction on the blended image I_b indicating that no pixel represents this text. We therefore define the negative loss to optimize for this requirement:

$$\mathcal{L}_{neg} = MSE(\mathcal{D}(\mathcal{C}(\mathcal{E}(I_b), T_{neg})), 0 \cdot \mathbb{1}_{|\mathcal{H}|}) \qquad (7)$$

3.4. Direct image-to-text alignment: $\mathcal{L}_{i2t}$

Our conditioning module $\mathcal{C}$ is using an attenuation strategy that is based on the similarity between the visual features $\{E^i\}$ returned by different depth blocks of the encoder $\mathcal{E}$ and the text embedding $\{W^i\}$ computed as explained in section 3.1. It is therefore likely that a stronger alignment between the $\{W^i\}$ and the $\{E^i\}$ would result in a more meaningful attenuation and in turn improved results. In this light, and inspired by ideas from [1] and [57], we added a direct image-to-text alignment loss $\mathcal{L}_{i2t}$ between (non-blended) batch images and their corresponding batch texts. First we compute a similarity between each pair of image #m and the text corresponding to (another or same) image #k in the batch:

$$Z_{k,m} = \max_i \left[\cos \left(\sum_{xy} \left[\cos_+ \left(W_k^i, E_m^{i,xy} \right) \cdot E_m^{i,xy} \right], W_k^i \right) \right] \tag{8}$$

where $\cos$ denotes the cosine similarity and the $\cos_+$ denotes its positive part, and $E_m^{i,xy}$ indicates the feature vector at spatial location (x, y) in the E_m^i tensor. Then we compute the $\mathcal{L}_{i2t}$ loss as:

$$\mathcal{L}_{i2t} = \sum_k \text{CE} \left[\text{softmax} \left(t_{i2t} \cdot Z_{k,\cdot} \right), k \right] + \tag{9}$$

$$\sum_m \text{CE} \left[\text{softmax} \left(t_{i2t} \cdot Z_{\cdot,m} \right), m \right] \tag{10}$$

where $Z_{k,\cdot}$ and $Z_{\cdot,m}$ stand for the text #k row and image #m column of the matrix Z respectively, t_{i2t} is the softmax temperature, and CE is the cross-entropy loss with respect to the index of the 'correct answer'. The 'correct answer' in this case is the respective row or column index itself, similarly to [57] we would like the text to best match its corresponding image in the batch - symmetrically when looking at the set of all batch images or all batch texts. Finally, direct matching of the text to the image also produces a heatmap predicting pixel correspondence to the query text. Therefore, in addition to $\mathcal{H}$ returned by the decoder $\mathcal{D}$ (slightly abusing notation, also referred to as $\mathcal{H}_{GbS}$ below), we define an additional output $\mathcal{H}_{i2t}$ from our model, which is the attention map produced by the direct matching:

$$\mathcal{H}_{i2t}(x, y) = \max_i \left[U_{|E^n|} \left(\cos_+ \left(W_k^i, E_m^{i,xy} \right) \right) \right] \tag{11}$$

here $U_{|E^n|}$ up-scales to spatial size of $|E^n|$. In our experiments (Section 4) we found that $\sqrt{\mathcal{H}_{GbS} \cdot \mathcal{H}_{i2t}}$, namely the per pixel geometric mean of $\mathcal{H}_{GbS}$ and $\mathcal{H}_{i2t}$, produces the best result and in the following we consider this geometric mean to be the main output of our GbS model $\mathcal{M}$.

4. Experiments

4.1. Datasets

MS-COCO 2014 [41] consists of $82,783$ training and $40,504$ validation images. Each image is associated with five captions describing it.

Flickr30k Entities [56] is based on Flickr30k [76] and contains $224K$ phrases describing localized bounding boxes in $\sim 31K$ images each described by 5 captions. For evaluation, we use the same 1k images from the test split as in [1].

VisualGenome (VG) [33] has $77,398$ train, 5000 validation, and 5000 test images. Each image comes with a set of free-form text annotated bounding boxes.

ReferIt has 20,000 images and 99,535 segmented image regions from the IAPR TC-12 [22] and the SAIAPR-12 datasets [7] respectively. Images also have an associated description for the entire image, and the image regions were collected in a two-player game [31] with approximately $130K$ isolated entity descriptions. We use the same $9K$ training, $1k$ validation, and $10K$ test images split as in [1].

Conceptual Captions 3M (CC3M) [60] has a total of $3,318,333$ training and $15,840$ validation images. The image raw descriptions are automatically harvested from the Alt-text HTML attribute associated with web images.

4.2. Implementation Details

All experiments were conducted on 4 Nvidia V100 GPU in multi-node DDP (1 GPU per node). We used the VGG [61] backbone from the torchvision [50] library, the PNASNet [42] from the TIMM library [69], and BERT [12] from the huggingface-transformers library [70]. As in [1], the VGG and PNASNet are ImageNet pre-trained. All experiments, unless otherwise noted, use the following configuration (found using MS-COCO validation set): (i) training batch of size 8 (pairs of images and text); (ii) half of the batch alpha maps are generated using Perlin noise [53] and half using a combination of two random Gaussians (more details in Sec. 4.4.3); (iii) the pre-trained BERT model is frozen; (iv) the projection modules $\mathcal{P}^i$ are a single fully connected layer; (v) we use $n = 2$ layers for the decoder $\mathcal{D}$ (Section 3.1); (vi) the decoder ResNet blocks D_i have 512 output planes (1 for the final output block) and stride 1; (vii) the pooling layers stride is $r = 2$; (viii) the losses weights are: $\gamma_{i2t} = 0.1$, $\gamma_{neg} = 1$, $\gamma_{adv} = 1$; (ix) the softmax temperature is $i_{i2t} = 10$; (x) we use the ADAM optimizer [32] and a linear LR schedule starting from $LR = 0.0001$ and dividing it by 10 every $50K$ steps; (xi) we use 50% dropout augmentation for the text, and random crop + 512×512 resize, color jitter, horizontal flip, and grayscale augmentations for the images.

The models where training for approximately 24 hours at 2.5 seconds per batch. Inference took 90 milliseconds per image-query pair on a single Nvidia V100 GPU.

4.3. Results

We follow the experimental protocol of [1], using the same data and splits for training, validation and testing. Specifically, in our experiments we evaluate our approach in two training setups: using either MS-COCO train split or the VG train split for training respectively. In both cases, as in [1], the resulting models are evaluated on the test splits of Flickr30K, VG, and ReferIt. Same as [1], we report the pointing-game accuracy [79] as our performance estimate in all of the experiments. Specifically, for the set of test 'image + query phrase' pairs, we report the percent of pairs for which the maximal point of the predicted heatmap for the pair was inside the ground truth annotation bounding box. Furthermore, we show the results of our method trained us-

Method	Backbone	Training	Test Accuracy		
			VG	Flickr30K	ReferIt
Baseline	Random	-	11.15	27.24	24.3
Baseline	Center	-	20.55	49.20	30.30
TD [78]	Inception-2	VG	19.31	42.40	31.97
SSS [28]	VGG	VG	30.03	49.10	39.98
MG [1]	VGG	VG	48.76	60.08	**60.01**
GbS (ours)	VGG	VG	**53.40**	**70.48**	59.44
MG [1]	PNASNet	VG	55.16	67.69	61.89
GbS (ours)	PNASNet	VG	**55.91**	**73.39**	**62.24**
FCVC [14]	VGG	MS-COCO	14.03	29.03	33.52
VGLS [71]	VGG	MS-COCO	24.40	-	-
MG [1]	VGG	MS-COCO	47.94	61.66	47.52
GbS (ours)	VGG	MS-COCO	**52.00**	**72.60**	**56.10**
MG [1]	PNASNet	MS-COCO	52.33	69.19	48.42
GbS (ours)	PNASNet	MS-COCO	**52.70**	**74.50**	**49.26**
GbS (ours) ensemble	-	MS-COCO	**54.55**	**75.60**	**58.21**

Table 1. Comparison with the state of the art DF-WSG methods evaluted using the "pointing game" accuracy on Visual Genome (VG), Flickr30K, and ReferIt. Our GbS method outperforms DF-WSG SotA when using corresponding backbones (VGG or PNASNet) by up to 10.4%. In **red**: best results with VGG; in **blue**: best results with PNASNet; in **bold black**: result of ensembling our GbS models.

Method	Overall	People	Animals	Vehicles	Instruments	Bodyparts	Clothing	Scene	Other
Ours (VGG)	72.6	82.5	**91.5**	81.1	56.6	34.8	58.6	70.9	59.9
Ours (PNASNet)	74.5	83.6	89.3	92.1	**83.3**	**53.2**	50.1	71.3	66.7
Align2Ground [11]	71.0	-	-	-	-	-	-	-	-
InfoGround (IG) [23]	**76.74**	83.2	89.7	87	69.7	45.1	**74.5**	80.6	67.3
12-in-1 [46]	76.4	**85.7**	82.7	**95.5**	77.4	33.3	54.6	**80.7**	**70.6**
IG + 12-in-1	81.1	87.1	90.4	95.5	74.2	61.5	74.0	79.9	73.5
Ours(VGG) + IG	83.9	88.8	96.1	93.4	74.4	**65.2**	77.2	82.4	76.8
Ours (PNASNet) + IG	83.4	87.3	95.2	95.4	75.2	62.3	**78.3**	81.5	77.2
Ours (VGG) + 12-in-1	**85.9**	**93.4**	**97.4**	96.4	79.6	52.6	78.0	**83.6**	78.3
Ours (PNASNet) + 12-in-1	84.9	93.3	96.5	**96.8**	**81.7**	54.8	71.4	82.3	**78.9**

Table 2. Detailed comparison with detector-based WSG methods [11, 23] on Flickr30K. The last two lines are an ensemble of our GbS models with the SotA InfoGround (IG) method [23]. In **blue** - best single model result, in **bold black** - best overall result. 12-in-1 [46] pointing accuracy results computed using official code. Align2Ground [11] did not provide detailed results and did not release their code.

ing the Conceptual Captions dataset setting a baseline for future work.

The evaluation results of our GbS and of other DF-WSG works (not using pre-trained detectors according to the definition of DF-WSG) are provided in Table 1. As can be seen, in both training regimes our GbS models significantly outperform the previous best results on all benchmarks using the matching backbones with $4-10.9\%$ absolute improvements for the lighter VGG backbone and $0.4-5.7\%$ absolute improvement for the much heavier PNASNet backbone. More specifically, we observe significant over 5.7% gains on Flickr30K in all training regimes, over 7% gains in ReferIt under MS-COCO training, and over 4% gains on VG in all training regimes using he VGG backbone. Moreover, when training using a large scale CC3M dataset, GbS (VGG backbone) attains **81.32%** Flickr30K grounding accuracy, which is 4.58% higher than the best result of even the detector based methods [23, 46], notably of that of [46] which was trained using a larger set of 4.4 million samples.

Interestingly, we also found that our proposed GbS approach is in fact *complementary* to the detector based WSG methods and can be effectively used to boost their performance. Any detector based method output can be converted to a heatmap by simple assignment of the bounding box scores to pixels of the bounding box (e.g. taking max for overlaps). As we show in Table 2, a simple geomet-

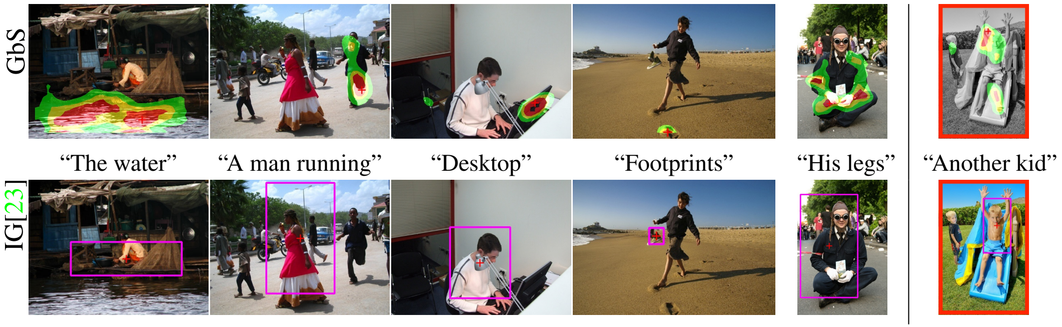

Figure 3. (Top) GbS heatmaps; (Bottom) IG [23] predicted boxes; (Middle text) grounding queries. We show cases where GbS handles phrases which are less familiar or ambiguous to the detector. On the right, where the query is ambiguous, both methods failed.

$\mathcal{L}_{neg}$	$\mathcal{L}_{adv}$	Pointing Accuracy
-	-	63.22
✓	-	66.8
-	✓	66.9
✓	✓	**72.6**

Table 3. The effect of the regularization losses $\mathcal{L}_{adv}$ and $\mathcal{L}_{neg}$.

Loss		Pointing Accuracy		
$\mathcal{L}_{i2t}$	GbS losses	$\mathcal{H}_{i2t}$	$\mathcal{H}_{GbS}$	Mean
✓	-	62.8	-	-
-	✓	-	69.5	-
✓	✓	68.2	67.1	**72.6**

Table 4. Different combinations of the direct image-to-text matching loss, GbS losses, and heatmap generation schemes. Here 'GbS losses' refers to our combination of $\mathcal{L}_{sep}$, $\mathcal{L}_{adv}$, and $\mathcal{L}_{neg}$.

ric average between the heatmap produced by our model and the heatmap resulting from the best performing detector based methods significantly boosts the pointing game accuracy of the latter indicating our model has learned to produce complementary predictions (e.g. for object categories less supported by the detector, such as some of the instruments and body parts in Table 2) boosting the combined performance by $\geq 7\%$ even over the detector-based SotA WSG approaches [11, 23, 46]. Additionally, we include a comparison to an ensemble of the two SotA detector-based methods InfoGround (IG) [23] and 12-in-1 [46] without our GbS model ('IG + 12-in-1' line in the table). As can be seen, an ensemble of our GbS model with any of these detector-based methods performs significantly better than an ensemble of the detector-based models between themselves (with the gain of 2.8% and 4.8% respectively). This shows that the gains obtained using an ensemble with GbS are not simply due to a combination of models, but rather likely stem from the GbS model being truly complementary to the detector-based methods. Some qualitative examples illustrating situations when not relying on the (more constrained) vocabulary of a pre-trained detector helps the grounding task are provided in Figure 3.

4.4. Ablations

We used the Flickr30K DF-WSG benchmark [56] with the 'pointing game' accuracy measure [79] for analyzing the relative contribution and importance of the different components of our GbS approach. All the ablation studies were carried out on (the lighter) VGG backbone trained on MS-COCO, and using our complete GbS approach with all of its components except the ones being examined in each respective ablation sub-section below.

4.4.1 Regularization losses $\mathcal{L}_{adv}$ and $\mathcal{L}_{neg}$

Table 3 evaluates the relative effect of our main regularization losses, namely: (i) the unconditioned adversarial loss $\mathcal{L}_{adv}$ responsible for suppressing the model parameters capitalizing on the artifacts of the synthetic blending we use for training our GbS models; and (ii) the negative text loss $\mathcal{L}_{neg}$ that encourages empty heatmap output when text is unrelated to the image. As we can see, each of these regularization losses adds above 3.5% to our GbS model performance affirming the benefits of their function. Moreover, jointly these two losses add more than 9% to the overall accuracy.

4.4.2 Image-to-text loss $\mathcal{L}_{i2t}$

Table 4 evaluates the benefit of the direct image-to-text matching loss $\mathcal{L}_{i2t}$ that is intended to improve the text and image features distributions alignment in order to facilitate better output of the conditioning module $\mathcal{C}$, as well as of an additional direct image-to-text attention output $\mathcal{H}_{i2t}$ resulting in the process of $\mathcal{L}_{i2t}$ computation. As the first row of Table 4 shows, training using only the $\mathcal{L}_{i2t}$ loss and using its corresponding output $\mathcal{H}_{i2t}$ (the only one available in this case) for grounding at test time is not sufficient for obtaining high performance. Significantly better accuracy (by almost 7%) is obtained via training using the $\mathcal{H}_{GbS}$ output and the GbS losses alone (without $\mathcal{L}_{i2t}$, second row). This indicates that the GbS losses contribute the most to the overall best result of 72.6% attained when using all the losses and outputs jointly (third row). We believe that the reason for this might be that the GbS losses employ a (synthetic) structured training signal (learning to predict localized masked regions of the blended image conditioned on the text), while $\mathcal{L}_{i2t}$ loss capitalizes on unstructured (bag-of-words like) contrastive (in the batch) text to image matching.

4.4.3 Blending alpha-map generation schemes

Perlin	Gaussian	Circle	Scale&Shift	Pointing Acc.
100%	0%	0%	0%	68.37
50%	50%	0%	0%	**72.6**
50%	0%	50%	0%	66.9
50%	0%	0%	50%	68.8
0%	100%	0%	0%	70.7
0%	50%	50%	0%	65.8
0%	50%	0%	50%	68.0
0%	0%	100%	0%	65.8
0%	0%	50%	50%	69.6
0%	0%	0%	100%	66.0

Table 5. Comparison of different variants of blending alpha-map generation including their mix (in % of the batch size).

Table 5 evaluates some choices for the blending alpha-map (α) generation scheme. Specifically, we examine the following alpha-map generators and their combinations (in portions of the batch): (i) the Perlin engine [53]; (ii) normalized pixel-wise combination of two random Gaussians: $\mathcal{G}\left[(x,y)|\mu_1,\sigma_1\right] / \sum_{j=1,2}\mathcal{G}\left[(x,y)|\mu_j,\sigma_j\right]$, with μ_j, and σ_j chosen at random and $\mathcal{G}$ being a Gaussian distribution; (iii) the Circle; and (iv) the Scale&Shift. The Circle refers to a binary circular mask (with randomly generated center and radius), and the Scale&Shift refers to a random scale and random relative shift blending of one of the images of the blended pair into the other image of the pair. We observed that increasing diversity via mixing different alpha-map generation schemes in most cases leads to increased performance compared to each of the schemes alone. Mixing the Perlin and Gaussian schemes attains the best result.

4.4.4 Language model ablation

In Table 6 we evaluate the effect of the choice of the language model (ELMO [54] or BERT [12]) used for the text embedding. Notably, even with the ELMO text embedding our proposed GbS approach retains significant performance gains (between 1.1% and 6.3%) above the results of [1] for the corresponding VGG backbone (winning in all cases except when testing on ReferIt after training on VG).

Language Model	Training	Pointing Accuracy		
		VG	Flickr30K	ReferIt
ELMO	VG	**53.65**	66.43	52.90
BERT	VG	53.40	**70.48**	**59.44**
ELMO	MS-COCO	49.03	67.9	49.37
BERT	MS-COCO	**52.00**	**72.60**	**56.10**

Table 6. Comparison of ELMO vs. BERT text encoders in our GbS model using the VGG backbone.

Condition Method	Pointing Accuracy
Distance	**72.60**
Attention	69.70
Projecton	69.54
Dist2Atten	69.43
Cosine	65.57

Table 7. Comparison of conditioning attenuation variants.

4.4.5 Attenuation for the text conditioning in $\mathcal{C}$

In Table 7 we evaluate several options for the type of the attenuation operation used in the conditioning module $\mathcal{C}$:
1. *Distance* (described by eq. (5)) attains the best result.
2. Alternatively, we also test the *projection* attenuation:

$$\mathcal{A}(E^i, W^i) = \cos_+\left(E^i, W^i\right) \cdot E^i. \tag{12}$$

3. The *attention* attenuation via using a self-attention block accepting the concatenated E^i and W^i (replicated to each pixel of E^i) and outputting a tensor of the same size as E^i.
4+5. Two 'scalar' attenuations that return a single channel tensors outputs, *dist2Atten*:

$$\mathcal{A}(E^i, W^i) = \exp(-\cos\left(E^i, W^i\right)) \tag{13}$$

and *cosine*:

$$\mathcal{A}(E^i, W^i) = \cos_+\left(E^i, W^i\right) \tag{14}$$

5. Conclusion

We propose Grounding by Separation (GbS) - a compositional approach for training text grounding models with weak (text only) supervision and without reliance on pre-trained detectors, as well as an architecture and a set of important regularization losses enabling our GbS approach to achieve a new SotA in Detector-Free WSG (DF-WSG). In addition, we show that our approach is complementary to the *detector-based* WSG by demonstrating significant improvement of the detector-based WSG accuracy SotA on Flickr30K when using our GbS model in a naive combination with a detector-based approach. Finally, in a comprehensive ablation study we clearly show the contributions of the novel GbS ideas to its success.

Interesting future work directions, which are beyond the scope of this work, include: adversarial optimization for the blending-alpha via back-propagation; recurrent generation of the blending alpha by applying the trained model to non-blended batch images and conditioning on random parts of associated text; exploring vision transformer backbones; and applications to multi-modal grounding outside the text domain (e.g. grounding of sound in still images).

Acknowledgments This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. FA8750-19-C-1001. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of DARPA. Raja Giryes was supported by ERC-StG grant no. 757497 (SPADE).

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, 3, 5, 6, 8

[2] Chris Alberti, Jeffrey Ling, Michael Collins, and David Reitter. Fusion of detected objects in text for visual question answering. *arXiv preprint arXiv:1908.05054*, 2019. 2

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433, 2015. 2

[4] Efrat Be'Ery and Arie Yeredor. Blind separation of superimposed shifted images using parameterized joint diagonalization. *IEEE Transactions on Image Processing*, 17(3):340–353, 2008. 3

[5] Guan-Lin Chao, William Chan, and Ian Lane. Speaker-targeted audio-visual models for speech recognition in cocktail-party environments. In *Interspeech*, 2016. 3

[6] Kan Chen, Rama Kovvuri, Jiyang Gao, and Ram Nevatia. Msrc: Multimodal spatial regression with semantic context for phrase grounding. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, pages 23–31, 2017. 2

[7] 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. 2, 5

[8] Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10800–10809, 2020. 2

[9] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*, pages 104–120. Springer, 2020. 2, 3

[10] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10578–10587, 2020. 2

[11] Samyak Datta, Karan Sikka, Anirban Roy, Karuna Ahuja, Devi Parikh, and Ajay Divakaran. Align2ground: Weakly supervised phrase grounding guided by image-caption alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2601–2610, 2019. 2, 3, 6, 7

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 4, 5, 8

[13] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T. Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. In *Proceedings of ACM SIGGRAPH 2018*, 2018. 3

[14] 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

[15] Kun Gai, Zhenwei Shi, and Changshui Zhang. Blind separation of superimposed moving images using image statistics. *IEEE transactions on pattern analysis and machine intelligence*, 34(1):19–32, 2011. 3

[16] Chuang Gan, Deng Huang, Hang Zhao, Joshua B Tenenbaum, and Antonio Torralba. Music gesture for visual sound separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10478–10487, 2020. 3

[17] Yosef Gandelsman, Assaf Shocher, and Michal Irani. " double-dip": Unsupervised image decomposition via coupled deep-image-priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11026–11035, 2019. 3

[18] Ruohan Gao, Rogerio Feris, and Kristen Grauman. Learning to separate object sounds by watching unlabeled video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 35–53, 2018. 3

[19] Ruohan Gao and Kristen Grauman. Co-separating sounds of visual objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3879–3888, 2019. 3

[20] Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. Vqa-lol: Visual question answering under the lens of logic. In *European Conference on Computer Vision*, pages 379–396. Springer, 2020. 2

[21] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2017. 2

[22] 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

[23] Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. Contrastive learning for weakly supervised phrase grounding. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 752–768, Cham, 2020. Springer International Publishing. 2, 3, 6, 7

[24] Tavi Halperin, Ariel Ephrat, and Yedid Hoshen. Neural separation of observed and unobserved distributions. In *International Conference on Machine Learning*, pages 2566–2575. PMLR, 2019. 3

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceed-*

ings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 770–778, 2016. 4

[26] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020. 2, 3

[27] Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000. 3

[28] Syed Ashar Javed, Shreyas Saxena, and Vineet Gandhi. Learning unsupervised visual grounding through semantic self-supervision. *arXiv preprint arXiv:1803.06506*, 2018. 2, 3, 6

[29] Vivek Jayaram and John Thickstun. Source separation with deep generative priors. In *International Conference on Machine Learning*, pages 4724–4735. PMLR, 2020. 3

[30] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 2

[31] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 5

[32] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[33] 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

[34] Donghoon Lee, Ming-Hsuan Yang, and Songhwai Oh. Generative single image reflection separation. *arXiv preprint arXiv:1801.04102*, 2018. 3

[35] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–216, 2018. 2

[36] Anat Levin and Yair Weiss. User assisted separation of reflections from a single image using a sparsity prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(9):1647–1654, 2007. 3

[37] Anat Levin, Assaf Zomet, and Yair Weiss. Learning to perceive transparency from the statistics of natural scenes. *Advances in Neural Information Processing Systems*, 15:1271–1278, 2002. 3

[38] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11336–11344, 2020. 2, 3

[39] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. VisualBERT: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 2, 3

[40] Liunian Harold Li, Haoxuan You, Zhecan Wang, Alireza Zareian, Shih-Fu Chang, and Kai-Wei Chang. Weakly-supervised visualBERT: Pre-training without parallel images and captions. *arXiv preprint arXiv:2010.12831*, 2020. 2, 3

[41] 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 *European conference on computer vision*, pages 740–755. Springer, 2014. 5

[42] Chenxi Liu, Barret Zoph, Maxim Neumann, Jonathon Shlens, Wei Hua, Li-Jia Li, Li Fei-Fei, Alan Yuille, Jonathan Huang, and Kevin Murphy. Progressive neural architecture search. In *Proceedings of the European conference on computer vision (ECCV)*, pages 19–34, 2018. 5

[43] Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. Visualnews : Benchmark and challenges in entity-aware image captioning, 2020. 2

[44] Francesc Lluís, Vasileios Chatziioannou, and Alex Hofmann. Music source separation conditioned on 3d point clouds. *arXiv preprint arXiv:2102.02028*, 2021. 3

[45] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*, 2019. 2, 3

[46] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10437–10446, 2020. 2, 3, 6, 7

[47] Rui Lu, Zhiyao Duan, and Changshui Zhang. Audio–visual deep clustering for speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(11):1697–1712, 2019. 3

[48] Daiqian Ma, Renjie Wan, Boxin Shi, Alex C. Kot, and Ling-Yu Duan. Learning to jointly generate and separate reflections. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 3

[49] 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 (CVPR)*, June 2016. 2

[50] Sébastien Marcel and Yann Rodriguez. Torchvision the machine-vision package of torch. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, page 1485–1488, New York, NY, USA, 2010. Association for Computing Machinery. 5

[51] Andrew Owens and Alexei A. Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 3

[52] Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. X-linear attention networks for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10971–10980, 2020. 2

[53] Ken Perlin. An image synthesizer. *ACM Siggraph Computer Graphics*, 19(3):287–296, 1985. 5, 8

[54] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018. 8

[55] Dinh Tuan Pham and Philippe Garat. Blind separation of mixture of independent sources through a quasi-maximum likelihood approach. *IEEE transactions on Signal Processing*, 45(7):1712–1725, 1997. 3

[56] 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. 2, 5, 7

[57] 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, 5

[58] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. 4

[59] Andrew Rouditchenko, Hang Zhao, Chuang Gan, Josh McDermott, and Antonio Torralba. Self-supervised audio-visual co-segmentation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2357–2361. IEEE, 2019. 3

[60] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia, July 2018. Association for Computational Linguistics. 5

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

[62] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019. 2, 3

[63] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7464–7473, 2019. 2, 3

[64] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 2, 3

[65] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE transactions on pattern analysis and machine intelligence*, 39(4):652–663, 2016. 2

[66] Cheng Wang, Haojin Yang, Christian Bartz, and Christoph Meinel. Image captioning with deep bidirectional lstms. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 988–997, 2016. 2

[67] Josiah Wang and Lucia Specia. Phrase localization without paired training examples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4663–4672, 2019. 2, 3

[68] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5005–5013, 2016. 2

[69] Ross Wightman. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019. 5

[70] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, Oct. 2020. Association for Computational Linguistics. 5

[71] 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, 3, 6

[72] Huijuan Xu and Kate Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *European Conference on Computer Vision*, pages 451–466. Springer, 2016. 2

[73] Xudong Xu, Bo Dai, and Dahua Lin. Recursive visual sound separation using minus-plus net. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 882–891, 2019. 3

[74] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. Boosting image captioning with attributes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4894–4902, 2017. 2

[75] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659, 2016. 2

[76] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 2, 5

[77] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. 3

[78] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neu-

ral attention by excitation backprop. *International Journal of Computer Vision*, 126(10):1084–1102, 2018. 2, 3, 6

[79] Jianming Zhang, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. In *European Conference on Computer Vision*, pages 543–559. Springer, 2016. 3, 5, 7

[80] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The sound of motions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1735–1744, 2019. 3

[81] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *Proceedings of the European conference on computer vision (ECCV)*, pages 570–586, 2018. 3

[82] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pretraining for image captioning and vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13041–13049, 2020. 2, 3

[83] Zhengxia Zou, Sen Lei, Tianyang Shi, Zhenwei Shi, and Jieping Ye. Deep adversarial decomposition: A unified framework for separating superimposed images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[84] Zhengxia Zou, Tianyang Shi, Zhenwei Shi, and Jieping Ye. Adversarial training for solving inverse problems in image processing. *IEEE Transactions on Image Processing*, 30:2513–2525, 2021. 3