LEARNING BETTER VISUAL REPRESENTATIONS FOR WEAKLY-SUPERVISED OBJECT DETECTION USING NATURAL LANGUAGE SUPERVISION

Anonymous authors
Paper under double-blind review

ABSTRACT

We present a framework to better leverage natural language supervision for a specific downstream task, namely weakly-supervised object detection (WSOD). Our framework employs a multimodal pre-training step, during which region-level groundings are learned in a weakly-supervised manner and later maintained for the downstream task. Further, to appropriately use the noisy supervision that captions contain for object detection, we use coherence analysis and other cross-modal alignment metrics to weight image-caption pairs during WSOD training. Results indicate that WSOD can better leverage representation learning by **(1)** learning a region-based alignment between image regions and caption tokens, **(2)** enforcing the visual backbone does not forget this alignment during the downstream WSOD task, and **(3)** suppressing instances that have weak image-caption correspondence during the WSOD training stage.

1 INTRODUCTION

Pre-training and fine-tuning have been integral parts of training deep neural networks for several computer vision tasks where supervision relies on densely-labeled data, such as object detection. Pre-training has commonly relied on datasets labeled at the image level, e.g. ImageNet (Deng et al., 2009). Recently, self-supervised pre-training strategies which leverage massive data readily available on the web have shown great success in learning visual representations. Prior works differ in the type of data used to formulate self-supervised objectives (e.g. different versions of the same image, images with corresponding captions). However, they aim to learn a generic visual space without any specific downstream task in mind, which raises a natural question: *"Can we benefit more from pre-training if the downstream task is already known?"*. In this work, we investigate the ways of learning better visual representations for weakly-supervised object detection (WSOD), including pre-training and fine-tuning on the downstream task.

Weakly-supervised detection has specific requirements: a model needs to ensure the features learned are capable of both distinguishing semantic categories, and estimating their boundaries, without explicit localization supervision. Most prior work formulates WSOD as a multiple instance learning (MIL) problem: an image is considered as a bag of regions, and a positive bag for a given class is supposed to have at least one region that contains an instance of that particular class. The model tries to maximize the likelihood of a positive region belonging to its ground-truth class, while minimizing the same likelihood for negative regions. While most prior WSOD approaches learn from image-level labels, some recent work (Ye et al., 2019) alleviates the need for supervision further, by learning from noisy, but more freely and naturally-available text appearing with images, e.g. web captions.

Since MIL can be seen of an instance discrimination task within a bag, high-level semantic information about *regions* should be captured in the visual representations used for initialization. Although current vision-language pre-training methods seem to be suitable for learning this semantic information about regions, they come with several drawbacks. First, most of these approaches (Chen et al., 2020b; Li et al., 2020; Lu et al., 2019), rely on a fully-functional object detector trained with box-level supervision to extract region features, which violates the core presumption of WSOD. Remaining approaches, however, may fail to provide such semantic information because: (1) even though their pre-text tasks require reasoning on both modalities, they do not learn an explicit align-

ment between image regions and text tokens, and (2) these approaches have architectural misalignments with WSOD that need to be addressed. Concretely, visual encoders used in such pre-training strategies are barebone CNNs followed by a pooling and a linear projection on each spatial location. However, region feature extractor of a WSOD architecture consists of a *specialized* pooling layer (e.g. RoI pooling (Girshick, 2015)) and series of *non-linear* projections. Integrating the region feature extractor into the pre-training architecture should result in bringing learned semantic information further towards the classification and detection streams used in WSOD models.

Apart from the problems that should be addressed in pre-training, we also hypothesize that transferring learned visual backbone bluntly to downstream WSOD task can cause several issues when supervision comes from noisy captions. First, manipulating the transferred rich visual space during MIL optimization for only a small number of semantic classes may result in overfitting to inconsistent labels and decreased generalization performance, considering image-level labels extracted from captions to supervise WSOD training will have a low recall due to the human reporting bias (Misra et al., 2016). Second, treating every instance equally within MIL optimization may add extra noise to the challenging weakly-supervised task. Noise arises because some object instances could be hard to learn from due to size, occlusion, and clutter. When the weak supervision comes from captions, further complications arise due to the loose, non-local association between the image and the corresponding captions (content in image is not precisely described in the captions).

In this work, we propose a framework for learning better representations for WSOD task from noisy but freely-available captions. We address the issues with pre-training by (1) learning an explicit alignment between image regions and text tokens, and integrating the region feature extractor into the pre-training architecture. Issues with the downstream WSOD task are addressed by (2) transferring the same region-token alignment objective from pre-training in order to preserve learned semantic space during MIL optimization as a form of regularization and (3) applying weighting mechanisms for reducing the effect of instances that have weak multimodal alignment with their paired captions, by relying on coherence analysis.

Our method achieves competitive performance on COCO, even though it uses image-level labels extracted from noisy captions by simple lexical matching. In addition to COCO, we focus on the important transfer setting, i.e. training on COCO, but evaluating on PASCAL VOC. This setting tests the true generalization ability of the detection models. We show our method's contribution is even more significant in the transfer setting. We also compare our method to a recent WSOD approach, Cap2Det (Ye et al., 2019), which also tries to learn an object detector from captions, and show our method achieves superior results, even though Cap2Det needs some ground-truth labels to train its text classifier for extracting image-level labels from captions. Most importantly, our method improves detection performance by absolute 2% (relative gain 8%) in the COCO $\rightarrow$ VOC transfer setting, without utilizing any ground-truth labels or additional data.

2 Related Work

2.1 Pre-training and self-supervised representation learning

The training recipe for dense computer vision tasks (e.g object detection) often entails initializing the network with the weights learned through pre-training on a large dataset (e.g. ImageNet), and then fine-tuning the obtained visual representation for the task of interest. This strategy can help learning downstream task in at least three ways. First, low-level convolutional filters for extracting primitive features, such as edge, blob and texture, would be already learned (Zeiler & Fergus, 2014). Second, the initialization weights would be capable of providing a semantic signal (e.g. 1,000 classes learned on ImageNet) that may well transfer to the downstream task. Third, pre-training may alleviate the risk of an optimizer being stuck in a bad local minimum, and provide a better starting point for learning than random initialization (Erhan et al., 2009).

While early approaches for pre-training relied on a dataset with labels at the image level (e.g. ImageNet), a vast amount of recent work has shown the benefits of self-supervised pre-text tasks. Some prior work use image-only pre-training with several pre-text objectives, maximizing representation similarity between different versions of the same image generated through augmentations (He et al., 2020; Chen et al., 2020a; Grill et al., 2020; Chen & He, 2021; Zbontar et al., 2021). Recent work in this line of research also investigates learning visual representations that can better

transfer to object detection task, utilizing region/pixel-level variants of similar pre-text objectives to ensure local consistency (Xie et al., 2021; Yang et al., 2021; Roh et al., 2021; Selvaraju et al., 2021). However, visual representations learned by this line of work suffer from lack of high-level semantic signal and leave bridging this semantic gap to downstream tasks. Another line of work utilizes both visual and textual data to alleviate this semantic gap, integrating pre-text tasks that require multimodal understanding (Zhang et al., 2020; Huang et al., 2020; Sariyildiz et al., 2020; Desai & Johnson, 2021; Yuan et al., 2021). Unlike prior work that treat each instance equally, Morgado et al. (2021) use instance weighting to suppress the effects of faulty positives and faulty negatives within multimodal contrastive learning. Recent work also use multimodal pre-training for zero-shot object detection (Zareian et al., 2021), learning an explicit pixel-level alignment between visual and textual tokens. Our approach is similar to Zareian et al. (2021) and Morgado et al. (2021) as we also perform multimodal pre-training with a specific task in mind, and integrate instance weighting. However, Zareian et al. (2021) learns a pixel-level alignment between visual and textual tokens during pre-training, and leaves learning how to localize object instances to the downstream task *utilizing box-level ground-truth labels for some categories*. We learn a region-level alignment between visual and textual tokens in pre-training and use the same region feature extractor in downstream WSOD, without utilizing any box-level ground-truth supervision. *Further, unlike our method, Zareian et al. (2021) do not utilize region-level alignment and cross-modal instance discrimination during the downstream task.* Morgado et al. (2021) formulates multimodal alignment within the same model to identify faulty positives and faulty negatives, while we exploit coherence and concreteness relations between modalities to quantify their alignment for instance weighting.

2.2 Weakly-supervised object detection via MIL

In the common multiple-instance learning (MIL) formulation, an image is considered as a bag of regions. If an image is labeled with class c, then there must be at least one region containing an instance of c. Oquab et al. (2015); Zhou et al. (2016) use global pooling layers to build class activation maps for instance localization. Bilen & Vedaldi (2016) introduce weakly-supervised deep detection networks (WSDDN) that rank region proposals using detection and classification streams. Kantorov et al. (2016) builds on WSDDN and integrates contextual information by exploiting region surroundings within two context models. Tang et al. (2017) formulate an iterative refinement module in which each iteration is supervised by its predecessor. Wan et al. (2018) learn spatial distribution of object classes by minimizing local and global entropy to reduce randomness of object locations. Ren et al. (2020) propose a spatial regularizer to combat part domination. All these works are similar to ours in terms of how they formulate WSOD problem. However, we use even weaker supervision utilizing image captions to learn an object detection model. In this perspective, our work is very similar to Ye et al. (2019) as their aim is also to learn an object detector using image captions. They extract image-level pseudo-labels from captions using a text classifier to guide WSOD training. Nonetheless, their method still requires ground-truth labels to train the text classifier. We learn an object detector from image-caption pairs without using any ground-truth labels.

3 Method

Our method aims to leverage natural language supervision in order to learn better visual representations for the downstream WSOD task. The contribution of language factors into three components of the method: (1) vision-language grounding over regions, (2) enforcing this grounding during both pre-training and weakly-supervised detection, and (3) weighting the contribution of image-text paired samples according to the degree of alignment in the pair. Our approach follows a standard pipeline (pre-training followed by the downstream detection task), but makes changes to both stages, to optimize the contribution of language for visual representation learning.

3.1 Multimodal pre-training

The purpose of pre-training is to leverage coarsely-aligned, co-occurring image-text pairs, as weak semantic signal for the subsequent detection. While image-text pre-training has been widely adopted for vision-language tasks (e.g., visual question answering), it is significantly less common for detection, especially WSOD. Our innovations include: (1) performing vision-language grounding on the region rather than pixel level, and (2) enforcing this grounding beyond the pre-training stage.

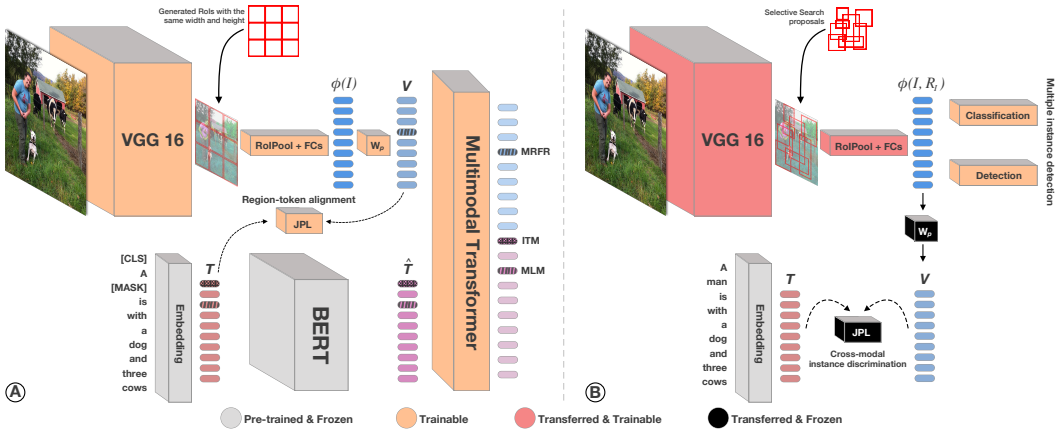

Figure 1: Demonstration of our overall framework for learning an object detector from noisy captions (best viewed in color). In pre-training (**A**), we learn an explicit region-token alignment along with other auxiliary objectives that require multimodal understanding. We transfer learned visual backbone along with the projection weights that are used for region-token alignment into downstream task (**B**), and perform cross-modal instance discrimination utilizing randomly-sampled captions to force our backbone not to forget the region-token alignment learned during pre-training.

Our multimodal pre-training architecture closely resembles PixelBERT (Huang et al., 2020). It takes a paired image and text, feeds them through separate encoders, and finally feeds these features into a multimodal transformer encoder to extract contextualized embeddings. However, our approach has *two major differences* with PixelBERT. First, we slice the extracted visual feature map into $n \times n$ spatial grid and apply RoI pooling followed by two non-linear transformations on each grid cell instead of treating each spatial location of the feature map as a visual token. The intuition is to learn better initialization weights for the region feature extractor which is the integral part of WSOD architecture. Second, our method learns an explicit alignment between image regions and caption tokens in a weakly-supervised manner contrasting other images and captions within the minibatch.

Our visual backbone ϕ takes an image I and extracts a $w \times h \times d^i$ feature map (d^i is the number of filters in the last conv layer). This feature map is sliced into $n \times n$ grid, and RoI pooling followed by two non-linear projections is applied on each grid cell to generate d^r-dimensional feature vectors for regions, resulting in $\phi(I) \in \mathbb{R}^{n^2 \times d^r}$. We employ a pre-trained BERT as the language backbone ψ, which takes a tokenized caption consisting of k words $T = [t_1, t_2, ..., t_k]$ and outputs contextualized token embeddings $\hat{T} = [\hat{t}_1, \hat{t}_2, ..., \hat{t}_k]$ where $t_i, \hat{t}_i \in \mathbb{R}^{d^w}$, resulting in $\psi(T) \in \mathbb{R}^{k \times d^w}$. We project each visual token into a d^w-dimensional space by multiplying visual tokens with $W_p \in \mathbb{R}^{d^r \times d^w}$ as we later feed both visual and textual tokens into a multimodal transformer encoder.

Next, we learn a linear joint-projection layer (JPL) that embeds both visual and textual tokens into a d^j-dimensional space in which we can measure cross-modal token similarity. JPL consists of a learnable weight matrix $W_{\text{JPL}} \in \mathbb{R}^{d^w \times d^j}$, which is same for both modalities. Even though one could measure cross-modal similarity on the d^w-dimensional space without applying a further projection, we found that doing so stabilizes training.

In the detection stage, we will need to learn associations between regions and semantic concepts. Thus, we want to mimic this objective in the pre-training stage. As we do not have ground-truth region-token associations to supervise our region-token alignment task, we learn it in a weakly-supervised manner contrasting other images and captions in the minibatch. Specifically, we calculate a global alignment score $\mathcal{S}(I, T)$ for an image-caption pair such that:

$$\mathcal{S}(I,T) = \frac{1}{k}\sum_{i=1}^{n^2}\sum_{j=1}^{k}(\mathbf{I} * \mathbf{T}^T)_{(i,j)} \quad (1)$$

$$\mathbf{I} = (\phi(I) * W_p) * W_{JPL} \quad (2a)$$

$$\mathbf{T} = T * W_{JPL} \quad (2b)$$

Using Eqs. 2a and 2b, we measure dot-product similarity between each region and textual token pair ($\mathbf{I} * \mathbf{T}^T$ in Eq. 1), and calculate image-level alignment by summing region-token dot-product similarity scores and dividing with the number of text tokens in the caption. We employ InfoNCE loss to enforce this alignment to be maximized for a positive pair, using other images and captions within the minibatch separately as negatives:

$$\mathcal{L}_{I \to T} = -\log \frac{\exp[\mathcal{S}(I, T_+)/\tau]}{\sum_{T \in B_T} \exp[\mathcal{S}(I, T)/\tau]} \quad (3) \qquad \mathcal{L}_{T \to I} = -\log \frac{\exp[\mathcal{S}(I_+, T)/\tau]}{\sum_{I \in B_I} \exp[\mathcal{S}(I, T)/\tau]} \quad (4)$$

where B_T and B_I denote the set of captions and images in the minibatch, respectively and τ is the temperature parameter and we set it to 1 in our experiments unless stated otherwise.

Several auxiliary objectives are used to train the multimodal transformer encoder to ensure our model learns not only how to align a token pair in isolation without knowing anything about other words and regions, but also reason about both modalities utilizing multimodal context attending to both visual and textual tokens. Following (Chen et al., 2020b; Huang et al., 2020), we employ masked-language modeling (MLM), image-text matching (ITM) and masked region feature reconstruction (MRFR) tasks. *While these objectives are not new, to the best our knowledge, we are the first to use them in a novel setting of learning visual representations for weakly-supervised object detection.*

MLM: Denoting visual regions $V = [v_1, v_2, ..., v_{n^2}]$ and caption tokens $\hat{T} = [\hat{t}_1, \hat{t}_2, ..., \hat{t}_k]$, we replace a random word $t_i \in T$ with [MASK]. The MLM objective is to correctly classify this masked token representation, attending to both caption tokens and visual regions:

$$\mathcal{L}_{\text{MLM}}(\theta) = -\mathbb{E}_{(V,\hat{T})\sim D} \log P_\theta(\hat{t}_i | \hat{t}_{j \neq i}, V) \quad (5)$$

ITM: We feed a special token, [CLS], along with visual and textual tokens to the multimodal transformer encoder to measure how well these two modalities align. We apply a fully-connected layer followed by sigmoid to get an alignment between 0 and 1. Denoting $y = 1$ for a positive image-caption pair (image and its paired caption in the dataset), we optimize a binary cross-entropy loss:

$$\mathcal{L}_{\text{ITM}}(\theta) = -\mathbb{E}_{(V,\hat{T})\sim D}[y \log \sigma_\theta(V, \hat{T}) + (1 - y) \log(1 - \sigma_\theta(V, \hat{T}))] \quad (6)$$

MRFR: We randomly sample a region i among n^2 regions and zero-out its feature vector v_i. Denoting the reconstructed feature vector as $h_\theta(\vec{0} \,|\, v_{j \neq i}, \hat{T})$, MRFR minimizes:

$$\mathcal{L}_{\text{MRFR}}(\theta) = \mathbb{E}_{(V,\hat{T})\sim D} \left\| v_i - h_\theta(\vec{0} \,|\, v_{j \neq i}, \hat{T}) \right\|_2^2 \quad (7)$$

The final loss that our pre-training method tries to minimize, $\mathcal{L}_{\text{PRE}}$, is the combination of the loss terms from individual tasks explained above:

$$\mathcal{L}_{\text{PRE}} = \frac{\mathcal{L}_{T \to I} + \mathcal{L}_{I \to T}}{2} + \mathcal{L}_{\text{MLM}} + \mathcal{L}_{\text{ITM}} + \mathcal{L}_{\text{MRFR}} \quad (8)$$

3.2 WEAKLY-SUPERVISED OBJECT DETECTION

Our method stands out from most prior weakly-supervised detection techniques in that it uses image-level labels that are not provided through crowdsourcing, but rather extracted from image captions. This technique opens the possibility to use freely-available web captions for supervision at the image level. However, in prior work, e.g. Ye et al. (2019), all image-caption pairs contribute equally to the detection loss. This is problematic because in some captions, the alignment between image and caption is more pure, hence the chance of false positives (objects mentioned in the caption but not shown) and false negatives (objects shown but not mentioned) is lower. Our contribution in this section is the investigation of image-caption weighting techniques that adjust the contribution of the paired samples used to extract labels at the image level.

Our method builds upon WSDDN (Bilen & Vedaldi, 2016), a well-established WSOD baseline, but adds a module to prevent the visual backbone from forgetting the rich semantic information

learned from captions in the pre-training stage, by transferring the visual backbone ϕ, W_p and W_{JPL}. Our method takes input image (I), set of captions (B_T) and region proposal set (R_I) generated by Selective Search (Uijlings et al., 2013). It performs multiple instance detection and cross-modal instance discrimination, detailed below. Unlike WSDDN and Ye et al. (2019), the final loss is weighed based on how well the input image and its paired caption are aligned.

Multiple instance detection. Given image I and region proposals R_I, we extract a d^r-dimensional feature vector for each region, resulting in $\phi(I, R_I) \in \mathbb{R}^{m \times d^r}$ where $|R_I| = m$ and $\phi(I, R_I) = [r_1, r_2, ..., r_m]$. These features are fed into two separate fully-connected layers that work in parallel to produce classification and detection scores for each region. Denoting the object classes to learn as $C = \{c_1, c_2, ..., c_k\}$, region r_i's classification and detection scores for a class c_a are calculated:

$$r^{\text{cls}}_{i,c_a} = \frac{\exp(r_i \cdot W_{\text{cls},a} + b_{\text{cls},a})}{\sum_{b=1}^{k} \exp(r_i \cdot W_{\text{cls},b} + b_{\text{cls},b})} \quad (9) \qquad r^{\text{det}}_{i,c_a} = \frac{\exp(r_i \cdot W_{\text{det},a} + b_{\text{det},a})}{\sum_{j=1}^{m} \exp(r_j \cdot W_{\text{det},a} + b_{\text{det},a})} \quad (10)$$

The model's prediction of c_a being present in the image is calculated as:

$$\hat{y}_{c_a} = \sum_{i=1}^{m} r^{\text{cls}}_{i,c_a} \, r^{\text{det}}_{i,c_a} \quad (11)$$

Lastly, given $y = \{y_{c_1}, y_{c_2}, ..., y_{c_k}\}$, which are the binary image-level labels extracted from captions and described shortly, our model optimizes the following for multiple instance detection:

$$\mathcal{L}_{\text{MID}} = -\frac{1}{k} \sum_{i=1}^{k} y_{c_i} \log(\hat{y}_{c_i}) + (1 - y_{c_i})(1 - \hat{y}_{c_i}) \quad (12)$$

Image-level label inference. We use EXACTMATCH, introduced in (Ye et al., 2019) as a baseline that does not require any supervision, to extract image-level object labels from captions. This approach entails applying lexical matching on captions to look for exact class labels.

Cross-modal instance discrimination (xID). We next introduce the cross-modal instance discrimination task that our method performs during WSOD training in order to preserve the rich semantic space learned from image captions during pre-training. We utilize a set of captions B_T which consists of a positive caption (i.e. paired with the input image) and randomly sampled negative captions. We first extract region features $\phi(I, R_I) = [r_1, r_2, ..., r_m]$ as explained previously, then map these to token space using transferred $W_p \in \mathbb{R}^{d^r \times d^w}$. We feed both visual ($V$) and textual ($T$) tokens into JPL to project them into the same joint-space learned in the pre-training stage. We calculate a global alignment score between image and caption as formulated in (Eqs. 1, 2a and 2b). Lastly, we employ InfoNCE loss to enforce this alignment score to be maximized between the input image and its paired caption among all captions in B_C, similar to Eq. 3:

$$\mathcal{L}_{\text{xID}} = -\log \frac{\exp[\mathcal{S}(I, T_+)/\tau]}{\sum_{T \in B_T} \exp[\mathcal{S}(I, T)/\tau]} \quad (13)$$

Instance weighting. The final loss that our method optimizes during WSOD training is the combination of two loss terms coming from aforementioned sub-tasks:

$$\mathcal{L}_{\text{WSOD}} = \beta(\mathcal{L}_{\text{MID}} + \lambda \mathcal{L}_{\text{xID}}) \quad (14)$$

Here λ weighs the importance of cross-modal instance discrimination within WSOD training and β is the weight for an individual instance. We found $\lambda = 0.1$ gives the best results. We experimented with the following to produce β values to weight supervision signal of image-caption pairs in WSOD training.

- $\beta_{\mathbf{ITM}}$: We use our pre-training architecture's own cross-modal alignment prediction from ITM output for an image-caption pair.

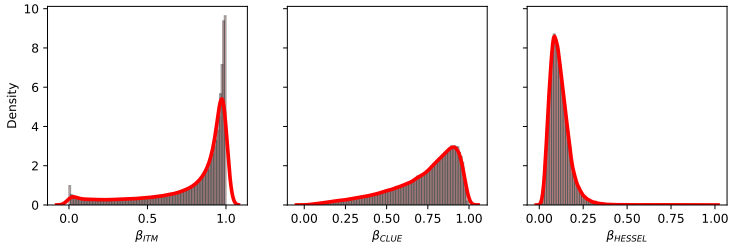

Figure 2: Distribution of β values in MS COCO.

- $\beta_{\mathbf{CLUE}}$: We use the output of a binary cross-modal coherence classifier trained on the CLUE dataset (Alikhani et al., 2020). CLUE investigates the different *purposes* of writing a caption for an image, and the different *coherence* relations between caption and image. The caption may literally describe the content of the image (resulting in a Visible relation), or may be complementary to the image, resulting in different relations, e.g. Story (describing circumstances) or Action (extending the activity presented in the image). CLUE contains image-caption coherence labels (Visible, Subjective, Action, Story, Meta) for a subset of Conceptual Captions (Sharma et al., 2018). We use these to train a Visible/Not model, and use the predicted probability of Visible, for a COCO image-caption pair, to weight its contribution in the detection loss. *The use of CLUE is innovative as it brings pragmatics and coherence analysis into object detection, and this has been done the first time.*
- $\beta_{\mathbf{HESSEL}}$: Hessel et al. (2018) assigns a concreteness score for words based on how similar the images paired with the same words are in a generic feature space. For each caption in our dataset, we extract Hessel et al. (2018)'s concreteness scores of each individual word that is in the particular caption and average these scores in order to compute the visual concreteness score for the caption. *To our knowledge, this is the first time concreteness scores have been used to weight supervision used for object detection.*

All β scores are bounded in $[0, 1]$, however their distributions vary greatly (see Figure 2). We shift their range with a constant ω so that cumulative effect of the weights would be the same as using no weighting at all. Concretely, we learn a shifting constant ω for each distribution that satisfies $\sum_{i=1}^{|D|} \omega + \beta^{(x_i)} = |D|$ where β^{x_i} denotes the weight of instance x_i and $|D|$ is dataset size.

4 Experiments

We evaluate all components of our method: **(1)** multimodal pre-training with region-token alignment, **(2)** transferring this alignment into WSOD training to preserve the learned visual space, and **(3)** weighting instances in WSOD training to suppress the effect of ones with weak image-caption alignment. We stress that all components can be easily integrated in existing multimodal pre-training and WSOD architectures. Hence, we focus on verifying our hypotheses with empirical evidence using simple baselines without bells and whistles, rather than outperforming the state-of-the-art.

4.1 Setup

Datasets and metrics. We use COCO (Lin et al., 2014) and PASCAL VOC2007 (Everingham et al., 2010). We pre-train our visual backbone on COCO utilizing its paired captions. We use both datasets to train and evaluate our WSOD method. Average precision (AP) is used as the performance metric for WSOD. We report $\text{AP}_{0.50}$ for both datasets, and also $\text{AP}_{0.50:0.95}$ for COCO.

Implementation details. All models and baselines were implemented in PyTorch (Paszke et al., 2019). We choose VGG-16 (Simonyan & Zisserman, 2015) as our visual backbone as it has been extensively used in the WSOD literature. Pre-training employs BERT (Devlin et al., 2019) from HuggingFace (Wolf et al., 2019), pre-trained on BookCorpus and English Wikipedia, as the language backbone. The multimodal transformer module is a 6-layer encoder (Vaswani et al., 2017) with 8 attention heads, and operates on 768-D feature space. We resize images to 490×490 and slice the generated feature map into 7×7 grid feeding RoI pooling with 49 static proposals, each of which

covers 70×70 on the input image. We randomly apply horizontal flipping with probability 0.5. For MRFR, we randomly choose one region among the $7 \times 7 = 49$ and zero out its feature vector. For MLM, we reduce dictionary size to $1,000$ to ease training, and apply masking only on the most frequent $1,000$ nouns, adjectives and verbs in COCO captions. We run pre-training for 20 epochs with mini batch size of 16, on an NVIDIA Quadro RTX 5000 (16GB). We start pre-training with an initial learning rate of $1e$-3 for multimodal transformer and $1e$-5 for visual backbone, then decay it at the end of 10th and 16th epochs by factor of 0.1.

For WSOD, we build on WSDDN (Bilen & Vedaldi, 2016) but replace spatial pyramid pooling (SPP) with RoI pooling, and remove the spatial regularizer following Tang et al. (2017). During training, we preserve the aspect ratio of images while randomly resizing their longest side to one of $\{480, 576, 688, 864, 1200\}$, and apply random horizontal flipping with probability 0.5. For cross-modal instance discrimination, we utilize 16 captions per image, one of which is positive. We train all WSOD models on a single GPU (Quadro RTX 5000 or GeForce GTX 1080Ti) with batch size 1. Training lasts 750K steps (~ 6.5 epochs) for COCO and 100K for VOC2007 (~ 20 epochs). We use initial learning rate of $1e$-3, and decay it by factor of 0.1 after 10th epoch for VOC2007 and 2nd for COCO. We control the random seed among experiments so that all models get training images in the same order. We held out a small validation set (200 images) from each dataset to evaluate model checkpoints, and pick the best performing checkpoint for complete testing. At inference time, we use both the original and horizontally-flipped version of the image, and resize their longest side to each of $\{480, 576, 688, 864, 1200\}$. Proposal scores are averaged across these 10 versions of the same image. We utilize max $1,000$ region proposals per image during training and inference.

4.2 RESULTS

Does region-token alignment help? To validate our pre-training hypothesis, we train two instances of WSDDN on VOC2007 train+val split using ground-truth labels, varying the initial weights. The first variant, WSDDN (W/O ALIGN), is initialized from the visual encoder of a multimodal pre-training architecture that shares the same auxiliary objectives with our approach but does not explicitly align visual regions and text tokens. The second variant, WSDDN (W/ ALIGN), it initialized using our pre-training architecture. Results in Table 1 show that learning an explicit region-token alignment in pre-training stage slightly improves the learned backbone's performance for WSOD.

Table 1: Effect of learning explicit region-token alignment in pre-training, testing on VOC2007.

	$\text{mAP}_{0.50}$ @ VOC
WSDDN (W/O ALIGN)	22.9%
WSDDN (W/ ALIGN)	**23.2%**

Does preserving transferred visual space help? Our hypothesis was that manipulating the rich visual space for a small number of semantic classes during WSOD training may cause problems such as overfitting and poor generalization ability especially when task supervision comes from noisy captions. To evaluate our hypothesis, we train two models: the first, WSDDN (W/O XID), is our WSDDN implementation and does not include any additions. The second model, WSDDN (W/ XID), adds cross-modal instance discrimination loss on top of WSDDN, utilizing a caption set among which one is positive. Both models start with the same visual backbone, learned through our pre-training schema with region-token alignment objective. They use EXACTMATCH labels, and are trained on the COCO 2017 train split, and tested on both COCO 2017 and VOC2007 test splits. Results in Table 2 show that integrating multimodal instance discrimination task improves the detection performance by 4% and 8% in $\text{mAP}_{0.50}$ and $\text{mAP}_{0.50:0.95}$ settings, respectively.

Table 2: Effect of integrating multimodal instance discrimination task in WSOD.

	$\text{mAP}_{0.50}$ @ COCO	$\text{mAP}_{0.50:0.95}$ @ COCO	$\text{mAP}_{0.50}$ @ VOC
WSDDN (W/O XID)	6.8%	2.5%	17.7%
WSDDN (W/ XID)	**7.1%**	**2.7%**	**21.4%**

Does instance weighting help? We train eight variants of our method using EXACTMATCH labels: pre-training with region-token alignment, with/without cross-modal instance discrimination in WSOD, using three different sources of β (as explained in Sec. 3.2) plus no-weighting versions.

Table 3: Contribution of different mechanisms to clean up the signal image-caption pairs provide. Best performer per group **bolded**; all weighting mechanisms outperforming the no-weighting baseline underlined. Gains (ratio of method vs baseline performance) obtained by our instance weighting are much more significant than incurred losses; losses are only in the within-dataset setting.

	$\text{mAP}_{0.50}$ @ COCO	$\text{mAP}_{0.50:0.95}$ @ COCO	$\text{mAP}_{0.50}$ @ VOC
OURS	**7.1%**	**2.7%**	21.4%
$+\beta_{\text{ITM}}$	7.0% (1% loss)	2.6%	21.5% (0% gain)
$+\beta_{\text{HESSEL}}$	7.0% (1% loss)	2.6%	21.6% (1% gain)
$+\beta_{\text{CLUE}}$	6.9% (3% loss)	2.6%	**21.9%** (2% gain)
OURS W/O XID	6.8%	2.5%	17.7%
$+\beta_{\text{ITM}}$	**7.4%** (9% gain)	2.8%	**19.1%** (8% gain)
$+\beta_{\text{HESSEL}}$	7.3% (7% gain)	2.7%	18.6% (5% gain)
$+\beta_{\text{CLUE}}$	7.3% (7% gain)	2.7%	18.2% (3% gain)

We observe in Table 3 that all of our instance weighting schemes help in the transfer setting, on PASCAL VOC. On COCO, using our full backbone (OURS, which uses xID, i.e. cross-modal instance discrimination), differences between weighting methods are smaller, and all weighting methods are equivalent and comparable to the no-weighting version. In the case where no instance discrimination is used, all weighting methods improve the no-weighting version (OURS W/O XID). Interestingly, the best overall setting differs between COCO (OURS W/O XID $+\beta_{\text{ITM}}$) and VOC (OURS $+\beta_{\text{CLUE}}$, which uses coherence analysis). **Importantly, gains obtained with any of our weighting methods are much more significant than losses** incurred in a single setting (top half, on COCO). This verifies the positive contribution of our weighting techniques.

Comparison with Cap2Det. We next compare our approach to a state-of-the-art method, namely CAP2DET (Ye et al., 2019), which also learns an object detector from image-caption pairs. CAP2DET utilizes a text classifier to extract image-level labels from an input caption to supervise WSOD training. However, it still needs image-level ground-truth labels to learn the text classifier. We train a WSDDN on COCO 2017 train split using the labels outputt by CAP2DET's text classifier (WSDDN(IM)-C2D). We also train WSDDN with ground-truth labels *as upper bound* (WSDDN(IM)-GT), and another using just the EXACTMATCH labels (WSDDN(IM)-EM). The three IM methods start with an ImageNet pre-trained visual backbone.

Table 4: Our method outperforms Cap2Det on all three settings without utilizing any ground-truth labels. Best performer except upper-bound (GT) is **bolded**. Gains shown for our methods (last two rows) are compared to the Cap2Det baseline (third row).

	$\text{mAP}_{0.50}$ @ COCO	$\text{mAP}_{0.50:0.95}$ @ COCO	$\text{mAP}_{0.50}$ @ VOC
WSDDN(IM)-GT	7.4%	3.1%	19.9%
WSDDN(IM)-EM	7.3%	2.7%	18.2%
WSDDN(IM)-C2D	6.2%	2.5%	20.0%
WSDDN (W/ XID, W/ ITM)	7.0% (13% gain)	2.6%	**21.5%** (8% gain)
WSDDN (WO/ XID, W/ ITM)	**7.4**% (19% gain)	**2.8**%	19.1% (4% loss)

Results in Table 4 clearly indicate that our method outperforms CAP2DET easily in all three settings, **without utilizing any ground-truth labels or any additional data.** It is worth mentioning that our method also outperforms WSDDN(IM)-GT in the COCO $\rightarrow$ VOC transfer setting, with 1.6% absolute (8% relative) improvement, while performing competitively in the $\text{mAP}_{0.50}$ setting on COCO.

REFERENCES

Malihe Alikhani, Piyush Sharma, Shengjie Li, Radu Soricut, and Matthew Stone. Cross-modal coherence modeling for caption generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6525–6535, 2020.

Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2846–2854, 2016.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020a.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15750–15758, 2021.

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*, pp. 104–120. Springer, 2020b.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE, 2009.

Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11162–11173, 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.

Dumitru Erhan, Pierre-Antoine Manzagol, Yoshua Bengio, Samy Bengio, and Pascal Vincent. The difficulty of training deep architectures and the effect of unsupervised pre-training. In *Artificial Intelligence and Statistics*, pp. 153–160. PMLR, 2009.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338, 2010.

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

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.

Jack Hessel, David Mimno, and Lillian Lee. Quantifying the visual concreteness of words and topics in multimodal datasets. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 2194–2205, 2018.

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.

Vadim Kantorov, Maxime Oquab, Minsu Cho, and Ivan Laptev. Contextlocnet: Context-aware deep network models for weakly supervised localization. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (eds.), *Computer Vision – ECCV 2016*, pp. 350–365, Cham, 2016. Springer International Publishing.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. What does BERT with vision look at? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5265–5275, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.469.

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*, pp. 740–755. Springer, 2014.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pp. 13–23, 2019.

Ishan Misra, C Lawrence Zitnick, Margaret Mitchell, and Ross Girshick. Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2930–2939, 2016.

Pedro Morgado, Ishan Misra, and Nuno Vasconcelos. Robust audio-visual instance discrimination. In *Computer Vision and Pattern Recognition (CVPR), IEEE/CVF Conf. on*, 2021.

Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 685–694, 2015.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32: 8026–8037, 2019.

Zhongzheng Ren, Zhiding Yu, Xiaodong Yang, Ming-Yu Liu, Yong Jae Lee, Alexander G Schwing, and Jan Kautz. Instance-aware, context-focused, and memory-efficient weakly supervised object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10598–10607, 2020.

Byungseok Roh, Wuhyun Shin, Ildoo Kim, and Sungwoong Kim. Spatilly consistent representation learning. In *CVPR*. IEEE, 2021.

Mert Bulent Sariyildiz, Julien Perez, and Diane Larlus. Learning visual representations with caption annotations. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, pp. 153–170. Springer, 2020.

Ramprasaath R Selvaraju, Karan Desai, Justin Johnson, and Nikhil Naik. Casting your model: Learning to localize improves self-supervised representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11058–11067, 2021.

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)*, pp. 2556–2565, 2018.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1409.1556.

Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. Multiple instance detection network with online instance classifier refinement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2843–2851, 2017.

Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Fang Wan, Pengxu Wei, Jianbin Jiao, Zhenjun Han, and Qixiang Ye. Min-entropy latent model for weakly supervised object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1297–1306, 2018.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16684–16693, 2021.

Ceyuan Yang, Zhirong Wu, Bolei Zhou, and Stephen Lin. Instance localization for self-supervised detection pretraining. In *CVPR*, 2021.

Keren Ye, Mingda Zhang, Adriana Kovashka, Wei Li, Danfeng Qin, and Jesse Berent. Cap2det: Learning to amplify weak caption supervision for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9686–9695, 2019.

Xin Yuan, Zhe Lin, Jason Kuen, Jianming Zhang, Yilin Wang, Michael Maire, Ajinkya Kale, and Baldo Faieta. Multimodal contrastive training for visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6995–7004, 2021.

Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *CVPR*, 2021.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. *arXiv preprint arXiv:2103.03230*, 2021.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pp. 818–833. Springer, 2014.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747*, 2020.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929, 2016.