Learning by Hallucinating: Vision-Language Pre-training with Weak Supervision

Tzu-Jui Julius Wang, Jorma Laaksonen
Aalto University, Finland
{tzu-jui.wang, jorma.laaksonen}@aalto.fi

Tomas Langer
Intuition Machines Inc.
tomas@intuitionmachines.com

Heikki Arponen
Systematic Alpha*
heikki.a.arponen@gmail.com

Tom E. Bishop
Glass Imaging*
tom@glass-imaging.com

Abstract

Weakly-supervised vision-language (V-L) pre-training (W-VLP) aims at learning cross-modal alignment with little or no paired data, such as aligned images and captions. Recent W-VLP methods, which pair visual features with object tags, help achieve performances comparable with some VLP models trained with aligned pairs in various V-L downstream tasks. This, however, is not the case in cross-modal retrieval (XMR). We argue that the learning of such a W-VLP model is curbed and biased by the object tags of limited semantics.

*We address the lack of paired V-L data for model supervision with a novel **V**isual **V**ocabulary based **F**eature **H**allucinator (WFH), which is trained via weak supervision as a W-VLP model, not requiring images paired with captions. WFH generates visual hallucinations from texts, which are then paired with the originally unpaired texts, allowing more diverse interactions across modalities.*

Empirically, WFH consistently boosts the prior W-VLP works, e.g. U-VisualBERT (U-VB), over a variety of V-L tasks, i.e. XMR, Visual Question Answering, etc. Notably, benchmarked with recall@{1,5,10}, it consistently improves U-VB on image-to-text and text-to-image retrieval on two popular datasets Flickr30K and MSCOCO. Meanwhile, it gains by at least 14.5% in cross-dataset generalization tests on these XMR tasks. Moreover, in other V-L downstream tasks considered, our WFH models are on par with models trained with paired V-L data, revealing the utility of unpaired data. These results demonstrate greater generalization of the proposed W-VLP model with WFH.

1. Introduction

Vision-language pre-training (VLP) has gained popularity as it shows great generalizability and transferability to many vision-language (V-L) downstream tasks. Pre-training is usually done on *webly-supervised* datasets, which are collected semi-automatically through the Internet and are hence noisy, e.g. the image and captions can be of weak mutual relevance. Furthermore, these uncurated image-text pairs may contain a wide spectrum of inappropriate contents that lead to some daunting biases when taken to train a model [3]. Despite trained on noisy datasets, these VLP models are shown to excel at various V-L downstream tasks [1, 45, 36, 30, 40, 28, 44, 6, 32, 11, 53, 22, 57]. More recent works, such as CLIP [41] and ALIGN [23], enjoy greater downstream improvements being pre-trained on even larger amounts of image-text pairs. These excellent prior works, on the one hand, offer a promising direction – a model properly pre-trained with massive amount of data, which could be imperfectly labeled, generalizes far better than one trained from scratch on a small dataset. On the other hand, the V-L research has been on a data-hungry path towards larger

(a) Fully-supervised pre-training

(b) This work: Weakly-supervised pre-training

Figure 1: Examples of different pre-training settings: (a) The fully-supervised setting is given image-caption pairs, whereas this work focuses on (b) The weakly-supervised setting which learns on unpaired images and captions.

*Work done at Intuition Machines Inc.

data collection efforts. This development could blur the other path more on trading-off the data efficiency and the generalization capability of V-L models.

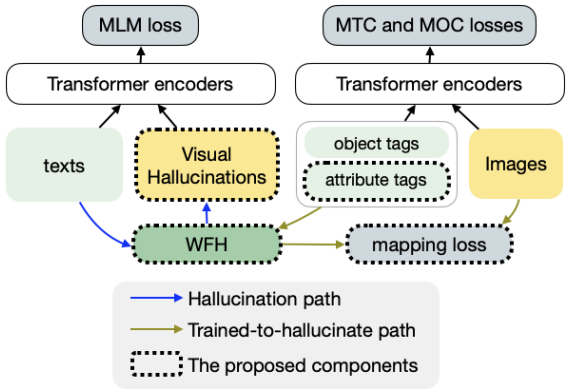


Figure 2: The proposed W-VLP model with the **V**isual **V**ocabulary based **F**eature **H**allucinator (WFH) at a glance. WFH is trained alongside to generate visual representations to pair with the textual counterparts. The components within the dotted frames distinguish us from the previous state-of-the-art W-VLP model, U-VisualBERT [31]. Please refer to Sec. 3.3 for the losses and their abbreviations.

Two different perspectives to enhance the data efficiency have been suggested. The first adopts the self-knowledge distillation principle, which guides the learning with soft labels predicted by the exponentially-averaged self, i.e. the same model with the parameters being updated by the exponential moving average [7, 15]. The second approach learns with limited access to paired images and texts [18, 31], thus largely reducing the effort in collecting a textual description for each image. This weakly-supervised setting makes VLP much more challenging since the aim of VLP is to learn to align V-L domains over paired data. Figure 1 illustrates the difference in the supervised and weakly-supervised settings.

Weakly-supervised VLP (W-VLP), though being a crucial step to unleash the potential of abundant web images and texts, is much less explored than supervised VLP (S-VLP) and only explored in some specific domains, e.g. medical imaging [10]. Interestingly, we find that the recently proposed W-VLP models, e.g. the unsupervised VisualBERT (U-VB) [31], largely fall short on cross-modal retrieval (XMR) tasks, motivating us to improve a W-VLP model particularly on XMR tasks. Concretely, our work enhances one of the pioneering W-VLP works, i.e. U-VB, by capitalizing more on the pre-trained visual attribute and object detectors with a novel Visual Vocabulary based Feature Hallucinator (WFH). WFH, depicted in Figure 2, is trained similarly to a W-VLP model without directly training on massive amounts of paired data. The central idea of WFH is to generate visual counterparts from textual representations with layers of Transformer encoders. The WFH-generated features are then paired with the originally unpaired texts.

It is worth clarifying that we do not claim the proposed model to be unsupervised (as is claimed for U-VB by its authors) but weakly-supervised. Both U-VB and our proposed model exploit knowledge from a pre-trained object detector for the follow-up unpaired training. Hence, they are exposed to some amounts of paired information, e.g. image regions and their object/attribute classes. We hereby consider that both models are learned under weak supervision.

We summarize the contributions as follows: (1) We present a novel WFH that enables more interactions across modalities during pre-training. (2) We propose a W-VLP model that accommodates object tags, attribute tags and the WFH-generated features. (3) The proposed model consistently outperforms the state-of-the-art weakly-supervised baseline, U-VisualBERT (U-VB), on the XMR tasks (i.e. text-to-image, image-to-text retrieval, and cross-dataset generalization), Visual Question Answering (VQA), Referring Expression Comprehension (REC), and Visual Entailment (VE) tasks, on totally six datasets. (4) We provide studies on, e.g. expressiveness of the word token embeddings and behavior of the attention probabilities in the Transformer encoder, to better understand the inner working of the W-VLP models. The introduced WFH is simple but shown effective given these quantified results.

2. Related Work

We introduce related work starting from the advancements in S-VLP methods followed by the W-VLP methods. We then explore more applications, e.g. image translation [59], medical image segmentation [47, 10], unsupervised machine translation [26] and unsupervised domain adaptation [34, 12, 46, 58], which advocate the usefulness of the unpaired data.

2.1. Supervised V-L Pre-training

Most recently proposed VLP models adapt Transformer [48, 9] for VLP with differences in architectures and training objectives. The VLP model architectures can be categorized into single- and two-stream models. The single-stream models, such as VisualBERT [30], ImageBERT [40], Unicoder-VL [28], VL-BERT [44], UNITER [6], Oscar [32], and SOHO [19] etc., adopt a unified Transformer sharing the parameters across modalities. The two-stream models, e.g. LXMERT [45] and ViLBERT [36], train a separate Transformer for each modality. These two separate Transformers cross-attend the representations from each layer of the other Transformer to learn cross-domain alignment through the attention mechanism. Though being architecturally simpler with less parameters to optimize, single-stream models are strongly comparable to two-stream models.

The usual training objectives of the VLP models are Masked Language Modeling (MLM) and Masked Region

Modeling (MRM), with variants such as Masked Object Classification (MOC) and Masked Region Feature Regression (MRFR). Image-Text Alignment (ITA), which classifies if the V-L inputs are aligned, is used to learn V-L alignment on the sentence level. The optimal transport method [8] can be used to learn fine-grained alignment across image regions and words. Oscar [32] introduces object tags detected from the images as the anchors [26] aligning word tokens and their visual groundings. VILLA [11] improves other V-L frameworks by adding adversarial perturbation to the V-L input spaces. More recent works have been advancing VLP by, e.g. training with larger datasets [41, 23] and enriching the image tags [57], which can benefit the framework such as Oscar. ALBEF [29] emphasizes cross-modal alignments in the early Transformer layers and learns from its momentum self to improve learning on noisy data.

2.2. Weakly-supervised V-L Pre-training

Aiming to pre-train a V-L model which learns to align V-L domains without image-text pairs, W-VLP is to save the substantial data collection effort. Hsu et al. [18] studied W-VLP in the context of medical imaging. Recently, Li et al. [31] proposed U-VB to be trained without accessing the image-text pairs. It learns cross-domain alignment with object tags served as anchors between domains and considered as "fake" data paired with the images. However, U-VB's learning could be confined by those tags which only amount to 1,600 object classes from Visual Genome (VG) [25] and bias the model to learn strong association between the visual and a limited amount of object tags' representations [52].

We thereby introduce a novel Visual Vocabulary based Feature Hallucinator (WFH), which aims to alleviate such a bias by generating regional visual representations to be paired with the textual description, e.g. a caption for an image. WFH generates diverse representations to offer a bridging signal across V-L domains. As a result, WFH greatly enhances U-VB over various V-L tasks.

2.3. Applications in Learning from Unpaired Data

Research interest in learning from unpaired data has grown in various applications. Along with the great advancement in Generative Adversarial Networks (GANs) [13], learning to translate images from one domain to another with different styles or artistic touches has been shown feasible without paired images [59]. Learning multi-modal representations for medical image analysis, e.g. organ segmentation, with unpaired CT and MRI scan images has also shown improvement in the segmentation accuracy compared to the models learned via a single modality [47, 10]. Unsupervised machine translation [26] and unsupervised domain adaptation [34, 12, 46, 58] share similarity with W-VLP in that they both learn to transfer or align domains without having access to paired data.

3. Our Proposed WFH Model

The proposed W-VLP model with Visual Vocabulary based Feature Hallucinator (WFH), sketched in Figure 3a, consists of a single-stream Transformer $\mathcal{T}_\theta$ which takes multi-modal inputs and shares parameters, i.e. those associated with *queries*, *keys*, and *values*, across modalities. Two sets of inputs are separately fed into $\mathcal{T}_\theta$. The first set $S_1 = \{(t_l, \mathbf{h}_l)\}_{l=1}^{L}$ consists of L text tokens t_l, each of which corresponds to a *hallucinated* visual representation $\mathbf{h}_l$, which we introduce in a later section. Another set of inputs $S_2 = \{(r_b, o_b, a_b)\}_{b=1}^{B}$ consists of (1) $B = 36$ regions of interest $\{r_b\}_{b=1}^{B}$ generated from a pre-trained object detector $\mathcal{O}$, the predicted object class probabilities given by $\mathcal{O}$, and (2) the sampled object tag $o_b \sim P_b^{obj}$ and attribute tag $a_b \sim P_b^{attr}$, where P_b^{obj} and P_b^{attr} are the predicted probabilities over the object and attribute classes [1] obtained from $\mathcal{O}$, respectively. $\mathcal{T}_\theta$ adopts the same Transformer architecture as in U-VB.

3.1. Model Architecture

This section focuses on formulating V-L inputs, WFH, the pre-training objectives and the losses. The differences with U-VB are emphasized.

3.1.1 V-L Inputs from S_1 set

Each language token t_l from the token sequence $\{t_l\}_{l=1}^{L}$ is obtained by tokenizing a sentence and embedded as

$$\mathbf{t}_l = T(T_{BERT}(t_l)) \in \mathbb{R}^{768}, l = 1, ..., L, \quad (1)$$

where $T_{BERT}(\cdot)$ is the BERT's embedding and T is a linear embedding layer. Each hallucinated visual representation is generated from the proposed WFH $\mathcal{H}_\phi$, i.e.

$$\mathbf{h}_l = \mathcal{H}_\phi(t_l | \{t_i\}_{i=1}^{L}, D) \in \mathbb{R}^{2048}, \; l = 1, ..., L, \quad (2)$$

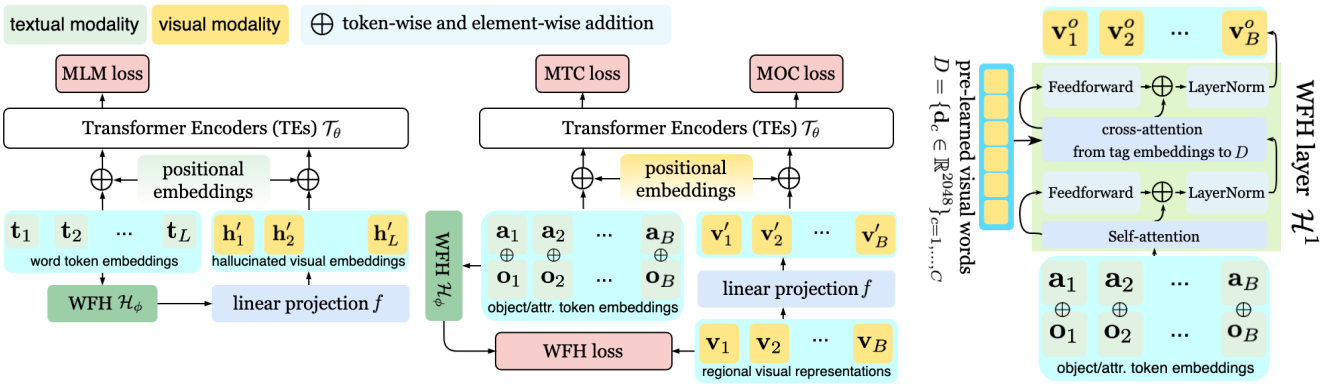
$$\mathbf{h}'_l = f(\mathbf{h}_l) = W_f \mathbf{h}_l + b_f \in \mathbb{R}^{768}, \; l = 1, ..., L, \quad (3)$$

where $D = \{\mathbf{d}_c \in \mathbb{R}^{2048}\}_{c=1}^{C}$ is the pre-learned visual dictionary. $\mathcal{H}_\phi(\cdot)$ is the hallucinator function which we will formally introduce later in Eqs. (8) and (9). f is a linear projection function parameterized by learnable weights $W_f \in \mathbb{R}^{768 \times 2048}$ and biases $b_f \in \mathbb{R}^{768}$. $\mathbf{t}_l$ and $\mathbf{h}'_l$ are respectively added with the token positional embedding $\mathbf{p}_l^W \in \mathbb{R}^{768}$ obtained by linearly transforming $l \in [1, ..., L]$, which is the token's position in the sequence. ϕ denotes WFH's parameters.

3.1.2 V-L Inputs from S_2 set

We denote the regional visual representations as $\{\mathbf{v}_b \in \mathbb{R}^{2048}\}_{b=1}^{B}$. Each $\mathbf{v}_b$ is extracted from r_b by $\mathcal{O}$ and trans-

[1] We refer the object and attribute classes to VG's object and attribute classes. The examples of object classes are *dishwasher*, *cat*, *ocean*, etc; attributes are *blank*, *matallic*, *talking*, etc.



(a) The proposed WFH model architecture and the training losses.

(b) Illustration of a 1-layer WFH.

Figure 3: The proposed weakly-supervised V-L pre-training model architecture with the cross-domain Visual Vocabulary based Feature Hallucinator (WFH).

formed to $\mathbf{v}'_b$ via

$$\mathbf{v}'_b = f(\mathbf{v}_b) \in \mathbb{R}^{768}, \tag{4}$$

which is consistent with the size of BERT's token embeddings. o_b and a_b are pre-processed similarly: tokenized by BERT's WordPiece tokenizer [9] and transformed to

$$\mathbf{o}_b = T(T_{BERT}(o'_b)) \in \mathbb{R}^{768}, b = 1, ..., B, \tag{5}$$

$$\mathbf{a}_b = T(T_{BERT}(a'_b)) \in \mathbb{R}^{768}, b = 1, ..., B, \tag{6}$$

respectively. o'_b and a'_b are the tokenized tags[2]. For the visual inputs, $\mathbf{v}'_b$ is added with the image positional embedding, $\mathbf{p}^I_b \in \mathbb{R}^{768}$, linearly transformed from a vector of the normalized bounding box x- and y-coordinates, width, and height of r_b. For the language inputs, we have $\mathbf{o}_b + \mathbf{a}_b + \mathbf{p}^I_b$. We fuse $\mathbf{o}_b$ and $\mathbf{a}_b$ embeddings by summing while other ways, e.g. appending the tokens, increases the number of input tokens, incurring much higher training complexity to the Transformer, which is quadratic to the number of tokens.

3.1.3 Differences with U-VB

Figure 2 highlights the differences between the proposed WFH model and U-VB. First, the WFH model additionally augments each object tag embedding with its attribute embedding. Second, U-VB processes t_l alone without a visual counterparts of any kind, unlike us pairing t_l with its hallucinations; in other words, the language part of U-VB's training is just fine-tuning BERT's weights on the given texts.

3.2. Learning Visual Hallucinator

As shown in Figure 3b, A WFH layer takes a textual token whose visual counterpart is to be hallucinated. The token is processed to account for its context in the sequence with a self-attention mechanism. The self-attention outputs then attend across modalities to each pre-learned $\mathbf{d}_c \in D$ to hallucinate visual representations. One can stack more WFH layers to model more complex interactions. The visual dictionary D is learned off-line by simple K-means with momentum updates [19] on the regional visual representations extracted from Conceptual Captions (CC) [42] images. Please find how D is learned in detail in the supplementary material (SM). Formally, the input to WFH can be a sequence of textual tokens $\{\mathbf{t}_l\}_{l=1}^L$ or $\{\mathbf{o}'_b\}_{b=1}^B$ with each $\mathbf{o}'_b$ obtained via

$$\mathbf{o}'_b = \mathbf{o}_b + \mathbf{a}_b, \; b = 1, ..., B. \tag{7}$$

When given $\{\mathbf{t}_l\}_{l=1}^L$, WFH generates visual representations $\{\mathbf{v}^t_l\}_{l=1}^L$. When given $\{\mathbf{o}'_b\}_{b=1}^B$, WFH generates $\{\mathbf{v}^o_b\}_{b=1}^B$. For example, generating $\mathbf{v}^o_b$ (as illustrated in Figure 3b) can be formulated as

$$\mathbf{v}^o_b = \mathcal{H}_\phi(\mathbf{o}'_b, \{\mathbf{o}'_i\}_{i=1,i\neq b}^B, D) \in \mathbb{R}^{2048}, \tag{8}$$

$$\mathcal{H}_\phi = \mathcal{H}^J \circ \mathcal{H}^{J-1} \circ \cdots \circ \mathcal{H}^1, \tag{9}$$

where $\circ$ is the function composition. For $j = 1, ..., J$,

$$\mathcal{H}^j(\cdot, \cdot, D) = \Big\|_{m=1}^M \{A^{j,m}_x(Q^{j,m}, K^{j,m}, V^{j,m})\}, \tag{10}$$

$$Q^{j,m} = \{W^{j,m}_Q \mathbf{o}''^j_i\}_{i=1}^B, \tag{11}$$

$$\mathbf{o}''^j_b = A^j_s(\mathbf{o}'^j_b | \{\mathbf{o}'^j_i\}_{i=1}^B) \in \mathbb{R}^{768}, \; \mathbf{o}'^1_b = \mathbf{o}'_b, \tag{12}$$

$$K^{j,m} = \{W^{j,m}_K \mathbf{d}_c\}_{c=1}^C, \tag{13}$$

$$V^{j,m} = \{W^{j,m}_V \mathbf{d}_c\}_{c=1}^C. \tag{14}$$

$\|$ indicates the concatenation of vectors, J is the number of WFH layers and in each layer A^j_s is the *self-attention* and $A^{j,m}_x$ is the m^{th} attention head (of totally M heads) in the *cross-attention* layer.

[2]Note that each tag can be tokenized to more than one tokens with the WordPiece tokenizer. We keep the same subscript b as in r_b, o_b, a_b for notational simplicity.

3.2.1 Self-attention A_s^j

A_s^j produces contextual textual representations $\mathbf{o''}_i^j$, which are input to each $A_x^{j,m}$ to construct the *query*. A_s^j is the multi-head attention mechanism introduced in [48] with $M = 12$ heads. Each head in A_s^j produces a $^{768}/_{12} = 64$-dimensional vector, and $\mathbf{o''}_b^j$ is the concatenation of those 12 vectors.

3.2.2 Cross-attention $A_x^{j,m}$

The textual *queries* in Q^j learn to align with the visual *keys* in $K^{j,m}$ and generate visual representations by $\mathcal{H}^j$. $W_Q^{j,m}$, $W_K^{j,m}$, $W_V^{j,m}$ are the learnable weight matrices for *queries*, *keys*, and *values*, respectively. $W_Q^{j,m} \in \mathbb{R}^{768/M \times 768}$, $W_K^{j,m} \in \mathbb{R}^{768/M \times 2048}$, and $W_V^{j,m} \in \mathbb{R}^{2048/M \times 2048}$ with $M = 16$ for $j = J$, i.e. in the last layer; otherwise, $W_V^{j,m} \in \mathbb{R}^{768/M \times 2048}$ with $M = 12$. Concatenating M vectors of each head produces the output of a WFH layer. One generates $\mathbf{v}_i^t$ given $\{t_l\}_{l=1}^L$ with the same process.

3.2.3 WFH Objectives

WFH is learned (1) implicitly through achieving Masked Language Modeling (MLM) task [31] with the captions, and (2) explicitly with the mapping loss:

$$L_{\phi,\text{WFH}} = \frac{1}{B}\sum_{i=1}^{B} ||\mathbf{v}_i^o - \mathbf{v}_i||_2^2, \tag{15}$$

where we regress $\mathbf{v}_i^o$, hallucinated from the object/attribute tag embedding $\mathbf{o}_i'$, to $\mathbf{v}_i \in \mathbb{R}^{2048}$, the visual representation extracted from $\mathcal{O}$. This objective ensures that the hallucinated features stay close to the visual domain in which the real $\mathbf{v}_i$ features reside.

3.2.4 Design Consideration for WFH

An alternative to Eq. (15) is to learn a direct projection matrix for mapping the textual representations to the visual ones. However, this is challenging in practice as it involves transforming high-dimensional distributions from one modality to another. Instead, the WFH's hallucination process is only to retrieve a visual representation from the space constructed by D, hence it avoids the direct mapping across domains and produces better hallucinations.

3.2.5 Visualizing Hallucinations

Please refer to the SM for the visualization, which shows that the hallucinated features (1) are contextual, and (2) appear to serve as the bridging representations across the V-L domains.

3.3. Loss Functions

$\mathcal{T}_\theta$ and WFH $\mathcal{H}_\phi$ are trained with the overall loss $L_{\theta,\phi}$ for an unpaired image and caption:

$$L_{\theta,\phi} = L_{\theta,\text{MLM}} + L_{\theta,\text{MTC}} + L_{\theta,\text{MOC}} + L_{\phi,\text{WFH}}, \tag{16}$$

which is composed of equally weighted Masked Language Modeling (MLM), Masked Tag Classification (MTC), Masked Object Classification (MOC) and WFH losses. MLM is to predict the masked tokens in the sequence $\{t_l\}_{l=1}^L$. Our MLM is conditioned on the hallucinated $\mathbf{v}_i^t$ along with the word tokens given at the input. MTC is to predict the masked object tag tokens. MOC is to predict the object class for the masked, i.e. zeroed-out, visual features. We closely follow U-VB's MTC and MOC including, e.g. tokens are masked at 15% of probability. The attribute tokens are always not masked. $L_{\phi,\text{WFH}}$ is the proposed WFH's objective given in Eq. (15). It is worth noting that WFH's learning is not only explicitly guided by $L_{\phi,\text{WFH}}$, but also implicitly by $L_{\theta,\text{MLM}}$ to generate useful features that also benefit the MLM task.

3.4. V-L Downstream Tasks

3.4.1 XMR Tasks

Following the same methodology as other VLP works, we fine-tune $\mathcal{T}_\theta$ with two additional projection layers for XMR. Specifically, $\mathcal{T}_\theta$ outputs contextual representations $\{\mathbf{t}_l'\}_{l=1}^L$ for the caption and $\{\mathbf{v}_b''\}_{b=1}^B$ for the image given a paired image and caption. We predict the matching score s by

$$s = \gamma \cdot \cos(f_t(\bar{\mathbf{t}}), f_v(\bar{\mathbf{v}})), \tag{17}$$

$$\bar{\mathbf{t}} = \frac{1}{L}\sum_{l=1}^{L} \mathbf{t}_l',\ \bar{\mathbf{v}} = \frac{1}{B}\sum_{b=1}^{B} \mathbf{v}_b''. \tag{18}$$

$f_t(\cdot)$ and $f_v(\cdot)$ are linear projections which do not change the dimensionalities of their inputs. Note that we summarize the image and caption with mean-pooled token embeddings instead of a single token embedding from, e.g. the class token `[CLS]` as in other VLP works [36]. We find that using mean-pooled embeddings leads to slightly better performance, aligning the finding in [37]. As in [36], the training objective is a 4-way classification that involves selecting three distracting choices for each image-text pair.

3.4.2 VQA, VE, and REC Tasks

For VQA, the model predicts the distribution over N_a answers on $\mathbf{c} \in \mathbb{R}^{1536}$,

$$\mathbf{c} = \bar{\mathbf{t}}\,||\,\bar{\mathbf{v}},, \tag{19}$$

i.e. the concatenation of $\bar{\mathbf{t}}$ and $\bar{\mathbf{v}}$, which is fed to a linear projection layer whose width is 1,024, followed by a GeLU [17] activation function and a N_a-way classification layer. Similarly for VE, the model predicts the answer distribution via passing c to a linear layer. For REC, to predict the visual grounding score for each image region, the model feeds $\mathbf{v}_b''$ (from Eq. (18), $b = 1, ..., B$) to a linear layer that outputs 768 neurons, which are then processed by GeLU and another linear layer producing the final matching score.

4. Experiments

This section introduces datasets for pre-training and fine-tuning for various V-L tasks. We detail the experiment settings followed by comparisons with other W-VLP methods.

4.1. Datasets and Tasks

4.1.1 Pre-training Dataset

The object detector that generates object and attribute tags is trained on VG. The visual vocabulary devised in WFH is fixed and pre-learned on the regional representations extracted from CC images with the same object detector. All the W-VLP models throughout the experiments are pre-trained on CC by randomly selecting a batch of captions and images which are not paired. Particularly, we use 2.7M images and captions from CC.

4.1.2 XMR Datasets and Tasks

The pre-trained models are fine-tuned on image-caption pairs from MSCOCO [33] or Flickr30K [39] to study their transferability. For Flickr30K, we follow the training/validation/test splits as in [24]. For MSCOCO, we follow the splits as in [27, 40, 28], and report the numbers by averaging over five folds of the test set, i.e. the "COCO 1K test set". We consider the following tasks: (1) image-to-Text Retrieval (TR), (2) text-to-Image Retrieval (IR), and (3) cross-dataset TR and IR, i.e. fine-tuned on COCO and tested on Flickr30K and vice versa.

4.1.3 VQA, REC, and VE Datasets

We assess the VQA task on two popular datasets: VQAv2 [2, 14] and GQA [21]. VE and REC are evaluated on SNLI-VE [51] and RefCOCO+ [55] datasets, respectively.

4.2. Model Parameters and Training Details

We develop our projects upon VOLTA [4], which is built with PyTorch [38] and aims for speeding up multi-modal machine learning research by establishing baselines within a controlled setup, e.g. models trained on same amount of text-image pairs across different VLP models. The object detector uses ResNet-101 [16] as the backbone. We follow the U-VB architecture, where each Transformer layer has $M = 12$ attention heads and the dimensionality of the hidden state is 768. The size of the pre-learned visual dictionary C is chosen from $\{1024, 1536, 3072\}$; the number of WFH layers J from $\{1, 2, 3\}$; γ in Eq. (17) from $\{8, 16, 32\}$. Throughout the experiments, the methods annotated with WFH are always trained with the attribute tokens added unless otherwise specified.

All W-VLP models are trained with eight 16GB-V100 GPUs with batch size of 400 for 12 epochs. AdamW [35] is used as the optimizer with weight decay as 0.01. The learning rate is adjusted with the warmup period being 10% of the total epochs. It is peaked at 1.5625×10^{-4} and linearly reduced to 0. The pre-training takes roughly one day for each model. At the fine-tuning stage, the models are trained with batch sizes of 64, 256, 256, and 192 for XMR, VQA, REC, and VE, respectively, with two 16GB-V100 GPUs. AdamW is used with weight decay being 0.0001.

4.3. Quantitative Results

In what follows, we present results of each task considered and provide studies on (1) the effects brought by varying variables in the proposed WFH, (2) spectral analysis [49] on the text token matrices, and (3) different patterns of the attention probabilities over the Transformer layers from attention heads learned by models with and without WFH. We aim for providing better understanding of how our and U-VB models perform differently in the latter two studies.

4.3.1 XMR Tasks

The main results of models on Flickr30K and MSCOCO are shown in Tables 1 and 2, respectively. UNITER's V-L paired results serve as an upper bound for the W-VLP models. Results from SCAN [27], SCG [43], PFAN [50], and GPO [5] are listed only for reference as they are task-specific and not proposed as generic VLP models. The models with † are replicated with VOLTA. Our work and U-VB share the same architecture, which adds a few additional task-specific layers on top of the pre-trained Transformer layers.

On Flickr30K, we first show the recalls of our model trained with 1-layer WFH without attribute tokens. While most recall values are comparable with U-VB, we observe a clear gain in R@1 on TR. Adding attribute tokens improves R@{1,5} on TR and all recall values on IR over U-VB. We obtain the best results with 2-layer WFH with 3.7% and 6.2% gains in R@1 on IR and TR, respectively. We take this model to continue the comparisons with other models in the rest of the experiments. On MSCOCO, the proposed model consistently surpasses U-VB in every recall value.

4.3.2 Cross-dataset Generalization on XMR

When trained and tested with different datasets, both models suffer from perceivable drops in recall values compared to when trained and tested with the same dataset. Nevertheless, the proposed model significantly outperforms U-VB as shown in Table 3. Notably, the improvements in recall over U-VB are always higher when both models are trained on Flickr30K than on MSCOCO, i.e. 14.5% and 16.5% gains in R@1 compared to 3.93% and 7.33% on IR and TR, respectively. This indicates that the proposed model could generalize better to a smaller dataset, e.g. Flickr30K (29K training images), which is about three times smaller than MSCOCO (82K training images).

Table 1: Comparing models on Flickr30K. We borrow results of U-VisualBERT (U-VB) from [31], in which it is trained on 3M images and 5.5M captions (CC + BookCorpus). The models with † and our WFH models are replicated and implemented with VOLTA, and all of which are trained only on 2.7M CC images and captions for fair comparison. From one row after U-VB †, it shows results of our proposed WFH models of different configurations, i.e. without attribute tokens involved (-attr) and different numbers of WFH layers employed, with the dictionary size $C = 1024$. The best models are highlighted in bold and the second best is underlined. Meta Sum is the sum of R@1, 5, and 10. The results shown with $\pm$ are the means and the standard deviations obtained with five pre-training runs with different random seeds.

Models	Text-Image Retrieval				Image-Text retrieval			
	R@1	R@5	R@10	Meta Sum	R@1	R@5	R@10	Meta Sum
SCAN [27]	48.6	77.7	85.2	211.5	67.4	90.3	95.8	253.5
SCG [43]	49.3	76.4	85.6	211.3	71.8	90.8	94.8	257.4
PFAN [50]	50.4	78.7	86.1	215.2	70.0	91.8	95.0	256.8
GPO [5]	60.8	86.3	92.3	239.4	80.7	96.4	98.3	275.4
UNITER [6] †	62.2	85.9	91.6	239.7	77.8	92.2	96.0	266.0
U-VB [31]	55.4	82.9	89.8	228.3	-	-	-	
U-VB †	54.4±0.3	81.7±0.4	88.8±0.3	224.2±1.1	67.8±0.3	90.7±0.5	94.9±0.8	253.5±1.1
U-VB (+attr) †	52.5	81.3	88.3	222.1	65.5	89.7	94.8	250
Ours: 1-layer WFH (-attr)	54.6	82.9	89.0	226.5	69.9	90.0	94.3	254.2
Ours: 1-layer WFH	55.0	82.7	89.8	227.5	71.7	**91.4**	94.8	257.9
Ours: 2-layer WFH	**56.4**±0.3	**83.2**±0.7	**89.9**±0.3	**229.5**±0.9	**72.0**±0.4	91.3±0.5	**95.6**±0.7	**258.9**±1.0

Table 2: Comparing models on MSCOCO, 1K test set. We replicate U-VB (U-VB †) as U-VB's authors did not report results on MSCOCO.

Models	Text-Image Retrieval			Image-Text Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
SCAN	58.8	88.4	94.8	72.7	94.8	98.4
SCG	61.4	88.9	95.1	76.6	96.3	99.2
PFAN	61.6	89.6	95.2	76.5	96.3	99.0
GPO	64.8	91.6	96.5	80.0	97.0	99.0
U-VB †	59.0±0.4	88.0±0.2	94.4±0.2	73.0±0.4	93.4±0.3	97.3±0.3
WFH	**61.9**±0.6	**89.4**±0.5	**95.3**±0.1	**73.9**±0.1	**94.6**±0.3	**98.0**±0.4

Table 3: Comparing models on cross-dataset generalization.

Models	Text-Image Retrieval			Image-Text Retrieval		
	Flickr30K train - MSCOCO test					
	R@1	R@5	R@10	R@1	R@5	R@10
U-VB †	37.0	69.3	80.9	45.6	73.0	83.3
WFH	**42.3**	**73.3**	**84.3**	**53.1**	**79.2**	**87.2**
	MSCOCO train - Flickr30K test					
	R@1	R@5	R@10	R@1	R@5	R@10
U-VB †	45.2	72.1	81.0	54.6	80.0	87.9
WFH	**47.0**	**73.8**	**82.4**	**58.6**	**83.9**	**90.5**

4.3.3 VQA, REC, and VE Tasks

Table 4 mainly compares our proposed WFH models and U-VB. SOTA referred in the table represents the state-of-the-art task-specific models which do not follow the same "pre-trained and fine-tuned" paradigm. We refer SOTA as MCAN [56] on VQAv2, NSM [20] on GQA, MAttNet [54] on RefCOCO+, and EVE-Image [51] on SNLI-VE as suggested in [6]. Please note that since our model, along with U-VB, aims at being generic and versatile for different V-L downstream tasks, direct comparisons with those task-specific models are not the primary focus of this work. Instead, we compare against U-VB (U-VB †) pre-trained in the VOLTA environment, while referring the readers to the reported results from the original U-VB work [31].

The proposed WFH consistently outperforms U-VB across all four tasks. Interestingly, comparing with VisualBERT (VB [4]), which is pre-trained with text-image pairs, both WFH and U-VB achieve competitive results on VQAv2 test-dev split and SNLI-VE, while WFH offers much clear improvements on GQA (+1.81 points on test-dev split) and on RefCOCO+ (+1.86 points on test split).

4.3.4 What Can Attention Probabilities Tell Us?

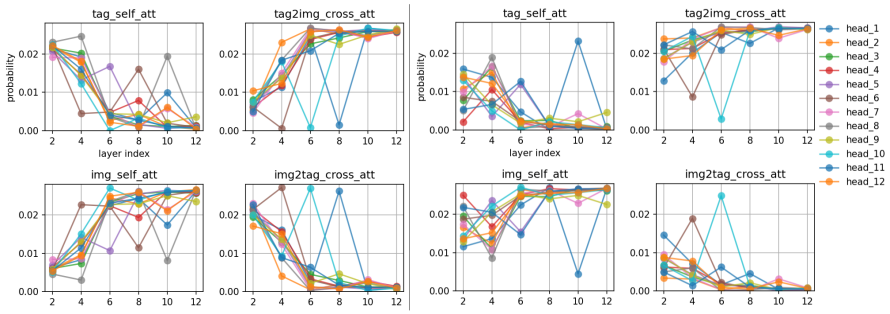
In Figure 4a we dive into the probability distribution of each attention head learned over the $M = 12$ Transformer layers across models. Both models exhibit similar patterns – the attention probabilities on vision ("img_self_att" and "tag2img_cross_att") keep increasing while those on language from tags ("tag_self_att" and "img2tag_cross_att") decrease. This suggests that both models gradually find alignments across modalities by adapting the textual domain to the visual domain. More intriguingly, our model demonstrates much higher language-to-vision cross-attention probabilities ("tag2img_cross_att"), already from the early layers. This indicates that our model could benefit from the earlier cross-domain alignments, which are shown to be beneficial for a VLP model, supporting the similar finding in [29]. Thereby, we would like to emphasize that although the proposed WFH method is simple, it leads to fundamental changes in how two modalities behave and results in better transferability.

4.3.5 What Can Spectral Analysis Tell Us?

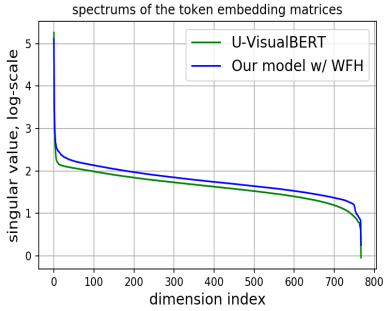
We also compare the spectra of the word token embedding matrices in Figure 4b, i.e. the weight matrix involves in $T_{BERT}(\cdot)$ in Eqs. (5) and (6), of U-VB and our WFH model. Our model's spectrum decaying slower indicates that the word token embeddings our model generates are likely more

Table 4: Comparing models fine-tuned for the VQA, REC and VE tasks. SOTA refers to the state-of-the-art task-specific models specified in Sec. 4.3.3. Any model with † refers to the replications realized in the VOLTA framework. VB w/o pt † represents the VisualBERT baseline without pre-training on image-text pairs. WFH refers to the same model as in Tables 2 and 3. The better models between U-VB † and WFH are highlighted in bold.

	VQAv2 test-dev				GQA	RefCOCO+			SNLI-VE
	overall	yes/no	number	other	test-dev	test set	testA	testB	test
SOTA	70.63	86.82	53.26	60.72	63.17	-	75.13	66.17	71.16
VB w/o pt †	66.07	82.74	46.51	56.29	53.55	67.81	75.41	58.91	74.56
VB †	68.20	-	-	-	56.58	69.70	-	-	75.67
U-VB [31]	70.74	-	-	-	-	-	79.11	64.19	-
U-VB †	67.78	84.15	49.71	57.89	56.53	70.53	77.82	62.00	75.02
WFH	**68.41**	**84.82**	**50.50**	**58.46**	**58.39**	**71.56**	**79.06**	**62.73**	**75.91**



(a) Left: U-VB. Right: The proposed WFH model.



(b) Spectrums from two pre-trained models.

Figure 4: Analyses on W-VLP models. (a) Probabilities of attention heads in selected Transformer layers. (b) Spectrums, i.e. the ordered singular values ordered by magnitude, of the text token embedding matrices.

expressive [49]. As such, the attention layers could be exposed with more diverse visual-textual embeddings from which they learn the alignment across the domains. This is especially crucial as the lack of paired V-L information is what the weakly-supervised model has to battle against.

4.3.6 WFH with Different Configurations

Please refer to the SM for the studies, which analyze (1) how different ways of utilizing attribute tokens and (2) differently configured WFHs, e.g. with varying visual dictionary sizes, affect the downstream tasks.

4.4. Qualitative Studies on XMR

Please refer to the SM The studies are conducted on Flickr30K through the XMR tasks on which we compare how capable the considered models are in terms of aligning attributes, entities, and activities etc., across V-L domains.

5. Conclusion

We proposed a novel W-VLP model that amends the lack of supervision from V-L pairs with a cross-domain hallucinator, dubbed as WFH, which generates bridging representations to interact with the textual modality.

Empirically, we found that the WFH model (1) learns more expressive word token embeddings, and (2) exhibits cross-domain alignments in the earlier Transformer layers. In retrieval tasks, it made consistent improvements, especially on the challenging cross-dataset generalization tests where it achieved at least 14.5% gains in R@1 over U-VB. The effectiveness of WFH was further confirmed in other V-L downstream tasks.

Next, we will study how much the WFH models generalize to the downstream tasks given varying amounts of supervision, e.g. the number of different tags used. In addition, the current W-VLP models cannot be considered fully unpaired because they rely on a trained object detector, which is trained on image and the class labels – a form of paired image-text data. We will explore ways to address this limitation to facilitate unpaired vision-language pre-training.

Acknowledgement

This work has been supported by the Academy of Finland in projects 317388, 329268 and 345791. Special thanks to Aalto Science IT and CSC – IT Center for Science, Finland for providing computing resources.

References

[1] 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] 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.

[3] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*, 2021.

[4] Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. Multimodal pretraining unmasked: A meta-analysis and a unified framework of vision-and-language BERTs. *Transactions of the Association for Computational Linguistics*, 2021.

[5] Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. Learning the best pooling strategy for visual semantic embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15789–15798, 2021.

[6] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. *arXiv preprint arXiv:1909.11740*, 2019.

[7] Ruizhe Cheng, Bichen Wu, Peizhao Zhang, Peter Vajda, and Joseph E Gonzalez. Data-efficient language-supervised zero-shot learning with self-distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3119–3124, 2021.

[8] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26:2292–2300, 2013.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, 2019.

[10] Qi Dou, Quande Liu, Pheng Ann Heng, and Ben Glocker. Unpaired multi-modal segmentation via knowledge distillation. *IEEE transactions on medical imaging*, 39(7):2415–2425, 2020.

[11] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. *Advances in Neural Information Processing Systems*, 2020.

[12] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, David Balduzzi, and Wen Li. Deep reconstruction-classification networks for unsupervised domain adaptation. In *European Conference on Computer Vision*, pages 597–613. Springer, 2016.

[13] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

[14] 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[15] 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.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[17] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

[18] Tzu-Ming Harry Hsu, Wei-Hung Weng, Willie Boag, Matthew McDermott, and Peter Szolovits. Unsupervised multimodal representation learning across medical images and reports. *arXiv preprint arXiv:1811.08615*, 2018.

[19] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12976–12985, 2021.

[20] Drew Hudson and Christopher D Manning. Learning by abstraction: The neural state machine. *Advances in Neural Information Processing Systems*, 32, 2019.

[21] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.

[22] Yuqi Huo, Manli Zhang, Guangzhen Liu, Haoyu Lu, Yizhao Gao, Guoxing Yang, Jingyuan Wen, Heng Zhang, Baogui Xu, Weihao Zheng, et al. Wenlan: Bridging vision and language by large-scale multi-modal pre-training. *arXiv preprint arXiv:2103.06561*, 2021.

[23] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *arXiv preprint arXiv:2102.05918*, 2021.

[24] 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.

[25] 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.

[26] Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *International Conference on Learning Representations*, 2018.

[27] 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.

[28] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, Daxin Jiang, and Ming Zhou. Unicoder-vl: A universal encoder for vision

and language by cross-modal pre-training. In *AAAI*, pages 11336–11344, 2020.

[29] Junnan Li, Ramprasaath R Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *arXiv preprint arXiv:2107.07651*, 2021.

[30] 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.

[31] Liunian Harold Li, Haoxuan You, Zhecan Wang, Alireza Zareian, Shih-Fu Chang, and Kai-Wei Chang. Unsupervised vision-and-language pre-training without parallel images and captions. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5339–5350, Online, June 2021. Association for Computational Linguistics.

[32] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020.

[33] 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.

[34] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR, 2015.

[35] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018.

[36] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23, 2019.

[37] Xiaofei Ma, Zhiguo Wang, Patrick Ng, Ramesh Nallapati, and Bing Xiang. Universal text representation from bert: An empirical study. *arXiv preprint arXiv:1910.07973*, 2019.

[38] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[39] 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.

[40] Di Qi, Lin Su, Jia Song, Edward Cui, Taroon Bharti, and Arun Sacheti. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. *arXiv preprint arXiv:2001.07966*, 2020.

[41] 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. *Image*, 2:T2, 2021.

[42] 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, 2018.

[43] Botian Shi, Lei Ji, Pan Lu, Zhendong Niu, and Nan Duan. Knowledge aware semantic concept expansion for image-text matching. In *IJCAI*, volume 1, page 2, 2019.

[44] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*, 2019.

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

[46] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017.

[47] Vanya V Valindria, Nick Pawlowski, Martin Rajchl, Ioannis Lavdas, Eric O Aboagye, Andrea G Rockall, Daniel Rueckert, and Ben Glocker. Multi-modal learning from unpaired images: Application to multi-organ segmentation in ct and mri. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 547–556. IEEE, 2018.

[48] 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*, pages 5998–6008, 2017.

[49] Lingxiao Wang, Jing Huang, Kevin Huang, Ziniu Hu, Guangtao Wang, and Quanquan Gu. Improving neural language generation with spectrum control. In *International Conference on Learning Representations*, 2019.

[50] Yaxiong Wang, Hao Yang, Xueming Qian, Lin Ma, Jing Lu, Biao Li, and Xin Fan. Position focused attention network for image-text matching. *arXiv preprint arXiv:1907.09748*, 2019.

[51] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*, 2019.

[52] Xu Yang, Hanwang Zhang, Guojun Qi, and Jianfei Cai. Causal attention for vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9847–9857, 2021.

[53] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil: Knowledge enhanced vision-language representations through scene graph. *arXiv preprint arXiv:2006.16934*, 2020.

[54] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1307–1315, 2018.

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

[56] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6281–6290, 2019.

[57] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021.

[58] Yabin Zhang, Bin Deng, Kui Jia, and Lei Zhang. Label propagation with augmented anchors: A simple semi-supervised learning baseline for unsupervised domain adaptation. In *European Conference on Computer Vision*, pages 781–797. Springer, 2020.

[59] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.