Multimodal Contrastive Training for Visual Representation Learning

Xin Yuan[1]*, Zhe Lin[2], Jason Kuen[2], Jianming Zhang[2], Yilin Wang[2]
Michael Maire[1], Ajinkya Kale[2], and Baldo Faieta[2]
[1]University of Chicago [2]Adobe Research
{yuanx, mmaire}@uchicago.edu {zlin, kuen, jianmzha, yilwang, akale, bfaieta}@adobe.com

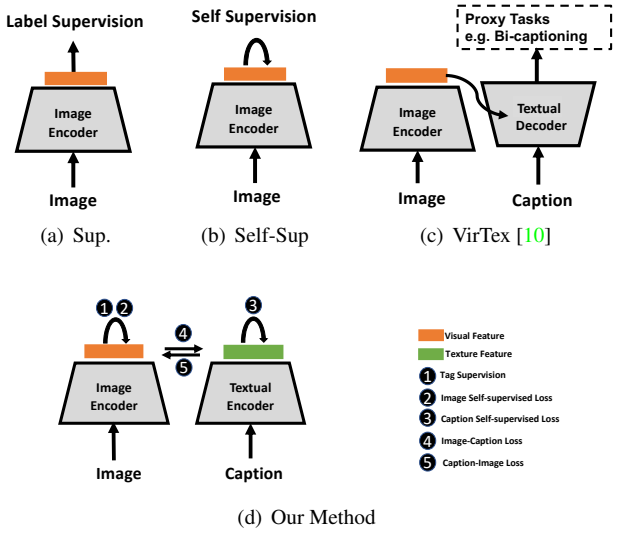
Abstract

We develop an approach to learning visual representations that embraces multimodal data, driven by a combination of intra- and inter-modal similarity preservation objectives. Unlike existing visual pre-training methods, which solve a proxy prediction task in a single domain, our method exploits intrinsic data properties within each modality and semantic information from cross-modal correlation simultaneously, hence improving the quality of learned visual representations. By including multimodal training in a unified framework with different types of contrastive losses, our method can learn more powerful and generic visual features. We first train our model on COCO and evaluate the learned visual representations on various downstream tasks including image classification, object detection, and instance segmentation. For example, the visual representations pre-trained on COCO by our method achieve state-of-the-art top-1 validation accuracy of 55.3% *on ImageNet classification, under the common transfer protocol. We also evaluate our method on the large-scale Stock images dataset and show its effectiveness on multi-label image tagging, and cross-modal retrieval tasks.*

1. Introduction

Visual representation learning is crucial for many computer vision tasks including image classification [9, 50, 27, 30], tagging [16, 23], object detection [17, 47, 40], semantic and instance segmentation [41, 26]. Supervised pre-training over large-scale datasets [9] yields useful visual features which lead to state-of-the-art performance on those tasks. Yet, fine-grained class labeling efforts [9] are prohibitively heavy. Self-supervised learning methods [4, 12, 59, 25, 5, 6] do not require any annotations, but still require either extremely large training sets or longer training epochs.

In addition to labels, image data usually comes with additional information including tags and captions, which is typically generated by internet users and therefore easier to acquire. More importantly, such multimodal information comes with higher-level abstract concepts, which offer the potential for drawing useful connections across different modalities [22, 31, 34, 20, 15].



(a) Sup. (b) Self-Sup (c) VirTex [10]

(d) Our Method

Figure 1. Main idea of the proposed method. Different from (c) VirTex [10], our method not only learns the cross-modal correlation between images and captions, but also exploits intrinsic data properties in a self-supervised manner within each modality.

Our objective is to learn visual representation from multimodal data in a unified training framework. The framework design should have the following essential properties: (1) fully exploits data potential within each unlabeled modality in a self-supervised manner; (2) bridges the heterogeneity gap by comparing different modalities in a common semantic space with similarity preservation objectives; (3) can be easily extended to take any new incoming modality. We aim to learn high-quality visual features, which benefit from not only the additional semantic information learned by cross-modal correlation modeling, but also the intrinsic data properties provided by each modality itself.

Some recently proposed methods [46, 35, 18, 19, 10, 49,

*This work has been done during the first author's internship at Adobe.

2] also focus on generating high-quality visual representations from scratch using multimodal data. For example, VirTex [10] makes a trade-off between the data efficiency and annotation effort by relaxing the extremeness of the unsupervised setting and embracing caption annotations which are relatively easy to acquire. However, as shown in Figure 1, VirTex is still trained in a single-path manner by solving a cross-modal proxy task, which is not sufficient to exploit the full potential within each individual modality.

In this paper, we take a unified view of both intra- and inter-modal similarity preservation in multi-modal data, based on which we develop a new visual representation learning framework, as shown in Figure 1. To be specific, an intra-modal training path is used to capture the intrinsic patterns of augmented data examples in a prediction task. A inter-modal training scheme is used to enhance the visual features by embracing the cross-modal interactions. With carefully designed contrastive losses, features in all modalities are adjusted via backpropagation in multiple training paths. We summarize our contributions as two-fold:

- **Unified Multi-modal Training.** Our multi-modal training framework can exploit intrinsic data properties within each modality and extract semantic information from cross-modal correlation simultaneously. In addition, as shown in Figure 1 and 2, our framework is symmetric for all modalities, which suggests it has the flexibility to incorporate any new modality.

- **Broad Transferability.** The visual representations pre-trained by our method can be transferred to many downstream computer vision tasks and achieve excellent performance under the common transfer learning protocols.

We demonstrate these advantages through making experimental comparisons between supervised, self-supervised and learning from text methods through extensive experiments on classification, tagging, cross-modal retrieval, object detection, and instance segmentation.

2. Related Work

Self-supervised learning. Many self-supervised methods [3, 54, 29, 60, 28, 25, 5, 6, 53, 57] utilize contrastive objectives for instance comparison in order to facilitate visual representation learning. For example, [57] use a memory bank which stores previously-computed representations and the noise-contrastive estimation (NCE) [24] to tackle the computational challenges imposed by the large number of instance classes. MoCo [25] further improve such a scheme by storing representations from a momentum encoder in dynamic dictionary with a queue. SimCLR [5] propose a simple framework under the large-batch setting, removing the needs of memory representations. MoCov2 [6] borrow the multi-layer perceptron (MLP) head design from [5] and show significant improvements. Our method shares the same spirit with these methods, in that we both use contrastive visual representation learning. However, we embrace multimodal data in multiple training paths to better align the visual features with additional semantic information.

Joint visual-textual pretraining. Vision and language (VL) methods [22, 31, 34, 20, 56, 42, 52, 7, 36, 21] are representatives that embrace multi-modal information for many computer vision tasks, such as image captioning and cross-modal retrieval. Such methods aim at mapping text and images into a common space, where semantic similarity across different modalities can be learned by ranking-based contrastive losses. Many works [51, 37] also focus on learning a joint visual-textual space via a fine-tuned BERT for VL tasks. However, these methods depend on the pre-trained image feature extractors or object detectors instead of learning from scratch on target datasets. Recently, [1] utilizes the NCE [24] and MIL-NCE losses [43] to learn representations using across video, language and audio modalities. Such methods share the similar cross-modal similarity preservation concept with ours in loss design. However, they only consider cross-modal correlation mining, while ours focuses on both intra-modal and inter-modal learning. Additionally, we demonstrate data efficiency by using a smaller dataset for pre-training visual representations.

Language-guided visual representation learning. Recently, several works [46, 35, 18, 19, 10, 49] propose pre-training approaches that use semantically dense captions to learn visual representations from scratch. For example, Virtex [10] jointly trains a convolutional network and a Transformer [55] from scratch to generate natural language captions for images, *i.e.* formulating a bi-directional image captioning proxy task. ICMLM [49] introduces image-conditioned masked language modeling (ICMLM) as a proxy task to learn visual representations over image-caption pairs. Both methods yield visual representations with good scalability and transferablity. However, the proxy tasks in these methods work in a single-path manner by merely conditioning on the visual representation while ignoring the intrinsic properties of visual data itself.

3. Method

Figure 2 shows the overall architecture for the proposed multi-modal contrastive training framework. The system is composed of two contrastive training schemes: intra-modal (orange and green paths) and intra-modal (yellow and blue paths) contrastive learning with different types of losses which are shown in Figure 3. The intra-modal training scheme is based on existing self-supervised representation

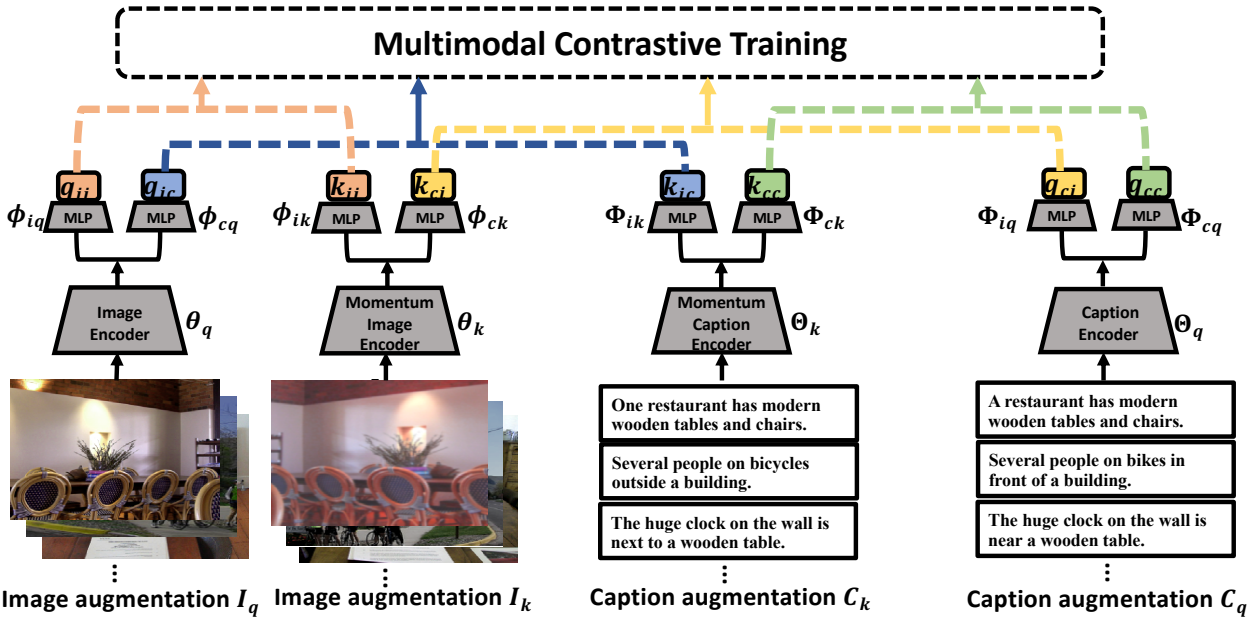


Figure 2. Framework of our proposed method is composed of two contrastive training schemes: intra-modal(**orange** and **green** paths) and inter-modal (**yellow** and **blue** paths) contrastive learning. In intra-modal contrastive learning, we train encoders for each individual modality in a self-supervised manner. In inter-modal contrastive learning, we compare different modalities in a common embedding space with bi-directional similarity preservation objectives. *Best viewed in color.*

learning framework MoCo-v2[6] that captures the intrinsic patterns of augmented image examples. However, self-supervised methods lack of the ability to learn semantic information from higher-level concepts [32]. We address such limitation by (1) designing additional textual encoder and momentum encoder to capture semantic information from augmented sentences. (2) involving the tag information in the contrastive loss to improve the visual representations.

The inter-modal training scheme is designed to enhance the visual feature by embracing the cross-modal interactions. We first embed the visual and textual features into a common space. Then, we design a visual-semantic contrastive loss to force the features of semantically-similar input examples to be closer. As such, visual features will be adjusted according to the captions via back propagation, and vice versa. Note that we use distinct MLP layers for cross-modal feature embedding so that the two intra-modality and inter-modality training schemes do not interfere with each other. Through the combinations of these two training schemes, we can learn powerful and generic visual representations. Although the proposed method also generates useful textual features as a by-product, it is not the main focus of this paper. After the multi-modal contrastive training has completed, the visual encoder can be directly applied to, or fine-tuned for, various downstream tasks.

3.1. Intra-modality Contrastive Learning

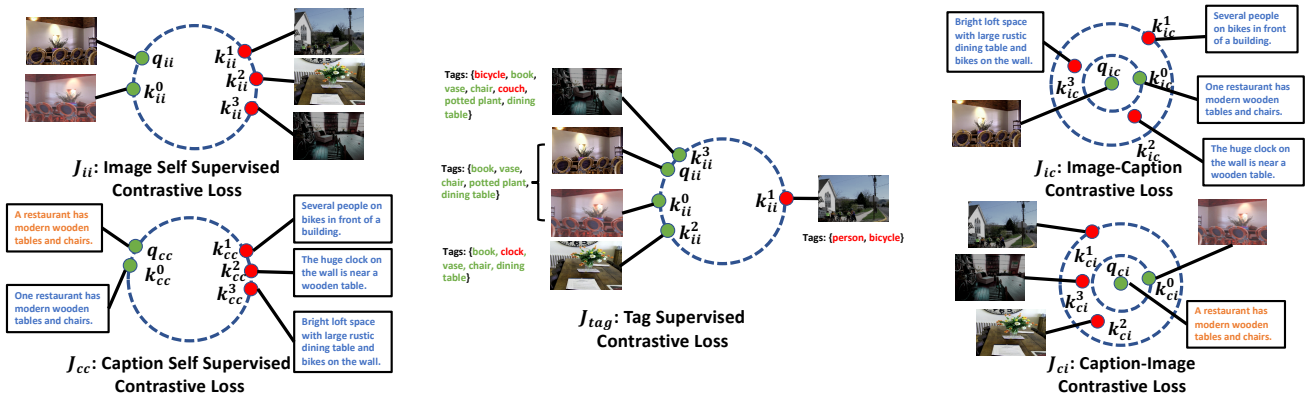
We first denote the multi-modal dataset as $D = \{(I_j, c_j, t_j)\}$, which comprises N image-caption-tags tuples. Note that t_j is a K-dim binary vector where $t_j^{(k)}$ is an indicator of the occurrence of a specific k-th tag in I_j. Our intra-modality contrastive training aims to preserve the similarity within the augmented variants of the same image or caption through self-supervised learning. As a running example, we formulate intra-modal visual/textual contrastive learning based on the MoCo-v2 framework.

Visual Contrastive Learning. We parameterize the image encoder as $f_{iq}(\cdot; \theta_q, \phi_{iq})$ and momentum image encoder as $f_{ik}(\cdot, \theta_k, \phi_{ik})$, where θ and ϕ are the weights of convolutional neural network (CNN) backbone and 2-layer MLP head, respectively. The weights θ_k, ϕ_{ik} are updated with momentum coefficient m: $\theta_k \leftarrow m\theta_k + (1-m)\theta_q, \phi_{ik} \leftarrow m\phi_{ik} + (1-m)\phi_{iq}$. The notation differs from MoCo-v2, which takes the encoder weights as a whole, since we require to map backbone features into different spaces, decoupling the feature embeddings from intra-modal and inter-modal training paths. For augmented examples $I_j^\dagger, I_j^\star$ from the same input image I_j in a minibatch, image encoder and momentum encoder embed them to *query* and *key* features:

$$q_{ii}^j = f_{iq}(I_j^\dagger; \theta_q, \phi_{iq}) \quad (1)$$

$$k_{ii}^j = f_{ik}(I_j^\star; \theta_k, \phi_{ik}) \quad (2)$$

Then, a dynamic set of *key* features with length K is maintained by iterative *dequeue* and *enqueue* operations. For *query* feature q_{ii} in the current mini-batch, *key* feature k_{ii} in the queue is denoted as k_{ii}^+ if it can form a positive pair with q_{ii}, *i.e.* they originate from the same image. The visual self-supervised contrastive loss shown in the top of Figure 3(a)



(a) Visual(Textual) Contrastive Loss (b) Tag-supervised Contrastive Loss (c) Visual-Semantic Contrastive Loss

Figure 3. Different types of contrastive loss in our method. These losses enable similarity preservation in both intra- and inter-modal training, encouraging features of semantically-similar input examples to be closer.

is defined as:

$$J_{ii} = -log\frac{exp(q_{ii} \cdot k_{ii}^{+}/\tau)}{\sum_{j=0}^{K} exp(q_{ii} \cdot k_{ii}^{j}/\tau)} \tag{3}$$

where $\cdot$ computes similarity scores between example pairs and τ is temperature hyperparameter.

Self-supervised learning frameworks consider all augmented examples originated from other images as negative samples, even if two images share very similar semantic concepts (e.g. they have a big overlap of tags) [32]. To encourage more closely semantic-aligned visual representation learning, we design an additional loss term in the visual-contrastive training path using the tag annotations provided by the dataset, as shown in Figure 3(b). For a *query* image I_j, in addition to the sample originating from the same input source, we also consider I_p as positive samples if I_p shares some common tags with I_j. Formally, we extend the *key* set from $\{k_{ii}^{+}\}$ to:

$$P = \{k_{ii}^{p} \mid \forall p : t_p \cdot t_j > \epsilon\} \tag{4}$$

where the dot product computes the similarity score between two tag lists and ϵ is a threshold hyperparameter. We thus define the tag supervised loss term by modifying Eq. 3:

$$J_{tag} = -\frac{1}{|P|}\sum_{p \in P} log\frac{exp(q_{ii} \cdot k_{ii}^{p}/\tau)}{\sum_{j=0}^{K} exp(q_{ii} \cdot k_{ii}^{j}/\tau)} \tag{5}$$

where $|P|$ denotes the set size of P. Note that J_{tag} degenerates to J_{ii} when there are no samples in *queue* sharing common semantics with *query* sample, *i.e.* $P = \{k_{ii}^{+}\}, |P| = 1$.

Textual Contrastive Learning. To learn useful semantic information from higher-level concepts, we design the textual encoder and momentum encoder to extract features from augmented captions. For the textual encoder architecture, we use BERT-like transformer architecture [11] with a 2-layer MLP head as a running example. Formally, we parameterize the textual encoder as $f_{cq}(\cdot; \Theta_q, \Phi_{cq})$ and momentum textual encoder as $f_{ck}(\cdot, \Theta_k, \Phi_{ik})$, where Θ and Φ is the weights of transformer and 2-layer MLP head, respectively. We utilize back-translation [13] for caption data augmentation. As in visual contrastive learning, we also maintain the same notion of *key*, *query* and *queue* in the textual contrastive training scheme. Given different augmented examples $c_j^{\dagger}, c_j^{\star}$ from the same input caption source c_j in a minibatch, the textural encoder and momentum encoder respectively embed them to *query* and *key* features. We formulate the embedding and mapping of caption modality as:

$$q_{cc}^{j} = f_{cq}(c_j^{\dagger}; \Theta_q, \Phi_{cq}) \tag{6}$$

$$k_{cc}^{j} = f_{ck}(c_j^{\star}; \Theta_k, \Phi_{ck}) \tag{7}$$

As shown in bottom of Figure 3(a), we aim to predict the positive *key* feature k_{cc}^{+} from the *queue* which originates from the same input source with q_{cc}. The contrastive loss is thus defined as:

$$J_{cc} = -log\frac{exp(q_{cc} \cdot k_{cc}^{+}/\tau)}{\sum_{j=0}^{K} exp(q_{cc} \cdot k_{cc}^{j}/\tau)} \tag{8}$$

where the dot product denotes similarity score and τ is a temperature parameter.

3.2. Inter-modality Contrastive Learning

To utilize the semantic information from captions for better visual feature learning, we enable cross-modal interactions via an inter-modality contrastive training scheme. We first embed the representations of the image and the caption into a common space and then use a ranking-based contrastive loss [20, 15] to learn both the visual and textual model parameters. In particular, we use CNN and BERT-like transformer as representation model backbones, with additional distinct branches of MLP layers with larger output dimensions. Note that such distinct MLPs design is not

ad-hoc, but based on the observation during experiments that using either unified or separate MLPs with the same-sized embedding space degrades the downstream task performance. As shown in Figure 3(c), the objective functions encourage the similarities of ground-truth caption-image pairs to be greater than those of all other negative pairs, instead of merely solving a *hard* prediction task.

Image-to-Caption Contrastive Learning. Given an image-caption pair (I_j, c_j), we generate the *query* feature using image encoder and *key* feature using momentum textual encoder, then map them to the common space:

$$q_{ic}^{j} = f_{iq}(I_j^{\dagger}; \theta_q, \phi_{cq}) \tag{9}$$

$$k_{ic}^{j} = f_{ck}(c_j^{\star}; \Theta_k, \Phi_{ik}) \tag{10}$$

where ϕ_{cq}, Φ_{ik} denote distinct MLP layers parameters from ϕ_{iq}, Φ_{ck} in Eq. 1 and Eq. 7. We denote the positive *key* feature k_{ic}^{+} from the *queue* which originates from the positive image-caption pair with q_{ic}, *i.e.* image is described by the caption. In the common space, we aim to simultaneously minimize the distance between q_{ic} and k_{ic}^{+} and maximize the distances between q_{ic} and all other negative *key* features from the *queue*. We thus formulate the image-caption contrastive loss as:

$$J_{ic} = \sum_{j=1}^{K} [\alpha - q_{ic} \cdot k_{ic}^{+} + q_{ic} \cdot k_{ic}^{j}]_{+} \tag{11}$$

where α is the margin, the dot product denotes similarity score, and $[x]_{+}$ represents $max(x, 0)$.

Caption-to-Image Contrastive Learning. Similar with image-to-caption contrastive learning, we generate the *query* feature using textual encoder and *key* feature using momentum image encoder as:

$$q_{ci}^{j} = f_{cq}(c_j^{\dagger}; \Theta_q, \Phi_{iq}) \tag{12}$$

$$k_{ci}^{j} = f_{ik}(I_j^{\star}; \theta_k, \phi_{ck}) \tag{13}$$

where Φ_{iq}, ϕ_{ck} denote distinct MLP layers parameters from Φ_{cq}, ϕ_{ik} in Eq. 6 and Eq. 2. The caption-to-image contrastive loss which aims at optimizing the distance between caption *query* and image *queue* is defined as:

$$J_{ci} = \sum_{j=1}^{K} [\alpha - q_{ci} \cdot k_{ci}^{+} + q_{ci} \cdot k_{ci}^{j}]_{+} \tag{14}$$

in which α is the margin, the dot product is similarity score, and $[x]_{+}$ represents $max(x, 0)$.

To this end, we formulate the final loss for our multi-modal contrastive training method as:

$$J = \lambda_{ii} J_{ii} + \lambda_{tag} J_{tag} + \lambda_{cc} J_{cc} + \lambda_{ic} J_{ic} + \lambda_{ci} J_{ci} \tag{15}$$

where λ_{ii}, λ_{tag},λ_{cc}, λ_{ic}, and λ_{ci} are trade-off parameters among different contrastive losses. Note that our method doesn't require all images to have tags. For some images with only tags or captions, the feature learning is guided by $\lambda_{tag} J_{tag} + \lambda_{ii} J_{ii}$ or $\lambda_{ii} J_{ii} + \lambda_{ci} J_{ci} + \lambda_{ic} J_{ic} + \lambda_{cc} J_{cc}$.

4. Experiments

We evaluate performance on multiple downstream tasks.

4.1. Experimental Setup

Pretraining Datasets. We train our models on the image-caption-tag tuples of the 2017 split of COCO [39] and Stock [58] image datasets. COCO has 123K images (118K and 5K for training and validation, respectively) with 5 captions for each image. For COCO dataset, we create tag sets for each image using the object list from the instance annotations. (80 objects in total). Stock is a large-scale dataset with 5.8 million training images and 50K test images. For each image, we have a title and a list of tags associated with it. We choose 18157 most frequent tags as our vocabulary during training.

Implementation Details. For image modality, we use ResNet-50 [27] as the backbone. We apply average pooling at the last layer of the backbone to generate a 2048-d feature vector. We follow the data augmentation scheme in [6] to generate a 224 $\times$224 image crop by random resizing, color jittering, horizontal flipping, grayscale conversion and gaussian blurring. For caption modality, we use BertTokenizer and Bert$_{base}$ model [11] to generate the 768-d pooled output, which is further processed by 2-layer MLP heads. For the text branch, we use back-translation [13] for caption data augmentation. More specifically, the given English sentence is randomly translated to French or German then back to English during training using the machine translation models [44, 13]. For all encoders, we use two distinct 2-layer MLP heads (hidden layer is 2048-d, with ReLU) to generate the last representations for contrastive training. The last layer of each encoder consists of a 128-d intra-modal representation and a 1024-d inter-modal representation. We normalize all feature vectors before calculating their dot products in the contrastive losses, where τ is set as 0.07. Momentum encoders' updating parameter m is set as 0.999 while K is 32768 and 65536 for COCO and Stock, respectively. For trade-off parameters in the final loss, we set λ_{ii} as 1.0, λ_{ic} as $1e^{-4}$, λ_{ci} as $1e^{-4}$ and λ_{cc} as 1.0. For λ_{tag}, we set it to 1.0 for our method with tag otherwise set it to 0. For loss term J_{tag}, we choose the threshold ϵ as 2 to extend the positive sample sets according to the tag information. For both J_{ic} and J_{ci}, we choose 0.2 as the margin parameter α. Finally, all networks are trained from scratch using SGD with weight decay of $1e^{-4}$, momentum of 0.9 and batch size of 512 on 8 V100 GPUs. We train image encoder and text encoder with an initial learning rate of 0.03 and $4e^{-5}$, and adopt the cosine learning schedule. Training epochs are 200 and 60 for COCO and Stock, respectively.

Downstream Tasks. Since our focus is visual representation learning, we only evaluate the pre-trained ResNet-50 backbone whose weights are denoted as θ_q while detaching all other modules in our framework. In particular, to evaluate the COCO pretrained visual backbone, we perform ImageNet [9] classification, PASCAL VOC [14] object detection, COCO instance segmentation tasks under the same setting and transfer learning protocols as in [25, 10]. Similarly, we also evaluate the Stock pretrained model on multiple downstream tasks including image tagging and cross-modal retrieval on Stock's test set to demonstrate effectiveness of our method in a large-scale dataset setting.

4.2. Evaluation of the COCO pretrained model

Linear Classification on ImageNet. We evaluate the learned visual representations (generated from our model pretrained on the COCO dataset) on ImageNet-1K (IN-1K) classification task. We choose ResNet-50 architecture as our visual backbone for all competing methods. The competing methods include three types: (1) supervised classification-based, (2) self-supervised, and (3) text-supervised. For supervised methods, we report the performance of competing methods trained on both full-sized IN-1K (1.28M images) and IN-100 (100K images) under label supervision. We construct IN-100 by randomly sample 100 images per class, which has the similar amount of images with COCO train set (118K). For self-supervised methods, we compare with MoCo and MoCo-v2, both pretrained on COCO dataset with unlabeled image data. For text-supervised methods, we adopt recently proposed VirTex and ICMLM$_{tfm}$, and compare to the best performance from their original papers.

We also evaluate another variant of our method – with additional tag supervision. Following the same transfer learning protocols as the competing methods, we train our linear classifier on features extracted from our frozen visual backbone on ImageNet-1K (IN-1K). In particular, we perform linear classification with a fully connected layer plus softmax on 2048-d global average pooled features extracted from the last layer of the COCO pre-trained visual backbone. We train on the IN-1K training set and report top-1 validation accuracy on the validation split, without performing any specific data augmentation. We train with batch size of 256 on 8 V100 GPUs for 60 epochs. We use SGD with momentum 0.9, weight decay 0 and learning rate 30, which is divided by 5 at 30th, 40th, and 50th epoch.

As shown in Table 1, our pretrained visual backbone outperforms self-supervised methods, which demonstrates that we effectively enhance the visual feature quality by utilizing the semantic information from other modalities. Our method trained with captions outperforms both VirTex and ICMLM by 2.1 *p.p* and 3.0 *p.p* respectively. Note that ICMLM uses heavier data augmentations in the linear classifier training stage and VirTex uses a longer training schedule in its pre-training stage than ours, which might lead to a performance gain in the IN-1K classification task. However, our method still achieves better performance, which suggests that it benefits from our design for implicit similarity preservation within individual modalities. We also train the model from scratch without using pre-trained BERT model. The scratch model yields slight lower performance than the default one and still consistently outperforms VirTex. In addition to comparisons with the state-of-the-art, we also observe that we can further improve the performance by 0.4 *p.p* by leveraging COCO's tag information. More importantly, when comparing with supervised methods that use similar amounts of pre-training data in IN-100, our method still performs better, even though the IN-1K classification task may unfairly favor supervised methods [10].

Table 1. Fully-, Un- and Text-supervised methods trained with ResNet50 backbones. We report top-1 obtained by linear classifier (on IN-1K) using pre-extracted features.

Model	Pretrain Dataset	Supervision	Top-1(%)
IN-Sup	IN-1K	Label	76.1
IN-Sup	IN-100	Label	53.3
MoCo [25]	COCO	NA	44.5
MoCov2 [6]	COCO	NA	49.3
VirTex [10]	COCO	Caption	52.8
ICMLM$_{tfm}$ [49]	COCO	Caption	51.9
Ours	COCO	Caption	54.9
Ours(scratch)	COCO	Caption	54.6
Ours(with tag)	COCO	Caption+Tag	55.3

Table 2. Object Detection on PASCAL VOC.

Model	Pretrain Dataset/Epochs	AP_{50}	AP	AP_{75}
Random Init	NA/NA	60.2	33.8	33.1
In-Sup	IN-1K / 90	81.6	54.3	59.7
MoCo [25]	IN-1K / 200	81.5	55.9	62.6
MoCo-v2 [6]	IN-1K / 200	82.4	57.0	63.6
MoCo [25]	COCO / 200	75.4	47.5	51.1
MoCo-v2 [6]	COCO / 200	75.5	48.4	52.1
VirTex [10]	COCO / 1000	81.4	55.6	61.5
VirTex* [10]	COCO / 200	80.2	54.8	60.9
Ours	COCO / 200	80.8	55.6	61.9
Ours(scratch)	COCO / 200	80.7	55.4	61.5
Ours(scratch)	COCO / 1000	82.1	56.1	62.4
Ours(with tag)	COCO / 200	81.8	55.8	61.7

Object Detection on PASCAL VOC. In addition to evaluating features extracted from frozen visual backbones, we follow anther common protocol [10]: fine-tuning the entire visual backbone for the object detection task on PASCAL VOC. For the competing methods, we choose the *Random Init* method as a simple baseline where the visual backbone is randomly initialized and trained on the downstream task. For IN-1K pretrained methods, we compare with both supervised and unsupervised (MoCo, MoCo-v2) methods. We also compare with MoCo and MoCo-v2 models trained

Table 3. Instance Segmentation on COCO: We use Mask R-CNN with ResNet-50-FPN backbones.

Model	Pretrain Dataset / Epochs	Box AP	Box AP_{50}	Box AP_{75}	Mask AP	Mask AP_{50}	Mask AP_{75}
Random Init	NA / NA	36.7	56.7	40.0	33.7	53.8	35.9
In-Sup	IN-1K / 90	41.1	62.0	44.9	37.2	59.1	40.0
MoCo [25]	IN-1K / 200	40.8	61.6	44.7	36.9	58.4	39.7
MoCo-v2 [6]	IN-1K / 200	41.5	62.2	45.7	37.4	59.6	40.5
MoCo [25]	COCO / 200	38.5	58.5	42.0	35.0	55.6	37.5
MoCo-v2 [6]	COCO / 200	39.8	59.6	43.1	35.8	56.9	38.8
VirTex [10]	COCO / 1000	40.9	61.7	44.8	36.9	58.4	39.7
VirTex* [10]	COCO / 200	39.6	60.9	44.0	36.0	57.6	38.9
Ours	COCO / 200	41.1	61.8	44.9	36.9	58.2	40.0
Ours(with tag)	COCO / 200	41.2	61.9	44.9	37.1	58.9	40.1

on COCO with default hyperparameters. When comparing with VirTex, we report the performance from (1) the original paper (pre-trained with 1000 epochs), (2) the model pretrained using the official code [1] and the default hyperparameters except that the epochs are set to 200. We train Faster R-CNN [48] models with ResNet-50-C4 backbones on VOC trainval 07+12 split, while evaluating on test2007 split using Detectron2 [2]. In particular, during the fine-tuning stage of total 24K iterations, we warmup learning rate for first 100 iterations to 0.02, which is then decayed by 10 at 18K and 22K iterations. The batch size is 16 and BN layers are synchronized (SyncBN) [45] across 8 V100 GPUs.

As shown in Table 2, our method significantly outperforms self-supervised methods which use COCO for pre-training. This suggests that useful semantic information is captured by our unique intra-modal pre-training scheme, which yields the visual backbone with better transferrability for the VOC object detection task. When comparing to VirTex under the same pre-training epochs of 200, our method consistently performs better. We also observe that we achieve comparable APs with VirTex (1000 epochs) while using 5× fewer pre-training epochs. When both trained for 1000 epochs from scratch, our method still consistently outperforms VirTex. These results suggest we benefit from the intra-modal similarity preservation that is not included in Virtex's design, which merely relies on a cross-modal proxy task during pre-training.

Instance Segmentation on COCO. We evaluate our method on the instance segmentation task of COCO, while choosing the same protocol and competing methods for VOC object detection. Following the setting used by VirTex, we train Mask R-CNN [26] models with ResNet-50-FPN [38] backbones on train2017split and evaluate on val2017 split. We adopt the 2× schedule of total 180K iterations with the initial learning rate of 0.02, which is decayed by 10 at 120K and 160K iterations. The batch size is 16 across 8 V100 GPUs and SyncBN is used. As shown in Table 3, our method performs slightly better than VirTex. Both our model and VirTex outperform the self-supervised methods which are also trained with COCO. Both methods achieve comparable results with the methods that are pre-trained on IN-1K with 10× more data.

[1] https://github.com/kdexd/virtex

[2] https://github.com/facebookresearch/detectron2

Table 4. Cross-modal search on COCO 1K test-set.

Method	**Image-to-Text**			**Text-to-Image**		
	R@1	**R@10**	**Med r**	**R@1**	**R@10**	**Med r**
IN-sup	57.9	92.7	1.0	42.8	87.0	2.0
MoCo-v2 [6]	51.6	90.0	1.0	39.0	84.8	2.0
VirTex [10]	58.1	93.0	1.0	44.0	88.5	2.0
Ours	58.4	93.4	1.0	45.1	90.0	2.0

Cross-modal Retrieval on COCO. We evaluate the learned visual representation on both image-to-text and text-to-image retrieval tasks on COCO. For a fair comparison with other methods which are pre-trained without textual encoder, we adopt 1-layer GRU which is randomly initialized as the sentence encoder [15] for all methods. This enables us to focus on comparing only the visual backbones. All visual backbones are ResNet-50 and generate 2048-d global pooled features, which are then mapped to 1024-d by fully connected layers. For the GRU textural encoder, we set the word embedding size to 300 and the dimensionality of the embedding space to 1024. We train the visual and textual encoders using the VSE++ [15] loss on 113K images with 5 captions each and evaluate on 1K images. We use the Adam optimizer [33] and set the batch size as 128. We follow the transfer protocol as in [15]: train with a fixed image encoder with learning rate $2e^{-4}$ for 15 epochs, and then decrease the learning rate to $2e^{-5}$ for the next 15 epochs. As for evaluation, we use the same evaluation protocols as in [15]: (1) $R@K$, defined as the percentage of queries in which the corresponding image is contained in the first K retrieved results. The higher this value, the better. (2) Med r, which is the median rank of the first retrieved ground-truth sentence or image. The lower its value, the better. The results are in Table 4. We see that our method consistently performs better than all competing methods. This suggests that the learned visual representation not only generalizes better on image-based downstream tasks but also on the cross-modal task, demonstrating the effectiveness of the implicit cross-modal similarity preservation objective.

Table 5. Image tagging on Stock 50K test-set.

Method	mIOU@5	mIOU@10	mIOU@15	mIOU@20
Stock-sup	12.86%	13.93%	14.88%	15.74%
Ours	13.81%	14.69%	15.55%	16.42%

Table 6. Cross-modal search on Stock 10K test-set.

Method	Image-to-Text			Text-to-Image		
	R@1	R@10	Med r	R@1	R@10	Med r
Stock-sup	33.76	71.22	3	30.64	68.18	4
Ours	36.98	74.02	2	33.83	70.76	3

4.3. Evaluation on the Stock dataset

We evaluate our method on the Stock image dataset and evaluate on its image tagging and cross-modal retrieval task.

Image Tagging on Stock. Similar with image classification task, we train on features extracted from the frozen visual backbone. In particular, we map the tags to 4096-d feature using Pointwise Mutual Information (PMI) embedding [8]. We use a 2-layer MLP (hidden layer is 2048-d with ReLU) to map image backbone feature to 4096-d. We train the MLP using the cosine similarity loss between pre-extracted tagging features and image features in the common embedding space. We compare with the supervised method, denoted as Stock-sup, which directly trains the backbone on the tagging task. The evaluation metric is mIOU@K, which is measured by the average overlaps between top-K predicted tags (pred) and ground-truth tags (gt) over all test samples N, *i.e.* $\sum_{i=1}^{N} \frac{|pred \bigcap gt|}{|pred \bigcup gt|}/N$. The Stock-sup model is trained directly for the tagging task and has an extremely unfair advantage, compared to the model trained with our method. However, as shown in Table 5, our method consistently outperforms the Stock-sup model.

Cross-modal Retrieval on Stock: We evaluate our method on the Stock cross-modal retrieval task, following the protocol in COCO cross-modal retrieval, except that we use 10K images for testing. Note that the competing method Stock-sup pre-trains the backbone on the Stock tagging task. As shown in Table 6, our method consistently outperforms the supervised baseline by a large margin.

4.4. Ablation Study

Investigation on Separate MLP Design. Using distinct MLPs for intra- and inter-modal feature embedding with different dimensions is an essential design in our method. We show the effectiveness of this design by varying the MLP head in our method (1) unified MLPs: we use shared MLP layers with output dimension of $\{128, 1024, 1152\}$. (2) distinct MLPs but with same dimensions of $\{128, 1024, 1152\}$. We train both baselines on COCO images and captions, and evaluate on the IN-1K image classification. For baseline (1), we achieve 49.6%, 50.2% and 52.3% top-1 accuracy, respectively. For baseline (2), the performance is 53.6%, 53.8% and 53.9%. Separate design consistently yields better visual features than unified design for IN-1K. Our final design (128-d for intra-modal; 1024-d for inter-modal) performs best (54.9%), which suggests our unified framework benefits from using a larger-sized common space for cross-modal correlation modeling and a relatively small one for self-supervision.

Table 7. Cross-modal search on COCO without fine-tuning.

Method	Image-to-Text			Text-to-Image		
	R@1	R@10	Med r	R@1	R@10	Med r
RS	0.1	1.0	650	0.1	1.0	500
Inter-modal	13.2	42.2	10.0	9.5	36.7	16.0
Multimodal	24.1	66.0	5.0	18.6	58.4	7.0

Investigation on Tag Supervision. We show the effectiveness of the tag supervision term in visual learning by pre-training on COCO unlabeled images using (1) J_{ii}; (2) J_{ii}+J_{tag}. and evaluating on IN-1K image classification. Note that training with merely the J_{ii} term is equivalent to MoCo-v2, which achieves 49.3% top-1 accuracy as shown in Table 1. Training with the additional tag supervision term J_{tag} further improve the accuracy to 50.2%. The final model trained with all five losses achieves the best performance of 55.3%, benefiting from the inter-modal training that leverages both image and caption information. We also use the ICMLM-style POS tagging + noun filtering to form image labels, in which ImageNet classification accuracy drops from 55.3% to 54.5 % by ignoring tags.

Investigation on Learned Textual Features. Even though the main focus of this paper is visual representations, we still generate useful textual features as a by-product. We show this via a COCO cross-modal retrieval downstream task. We design three variants for comparison: (1) RS: we randomly pair the image and caption across 1K test data. (2) Inter-modal: we train image and textual encoders merely with J_{ic} plus J_{ci} on images and captions. MLPs output dimensions are all set as 1024. (3) Multimodal: we use the multimodal pre-trained visual/textual encoder for feature extraction. As shown in Table 7, we observe that even without fine-tuning, variant (2) and (3) still perform much better than random baseline (1), which suggests some useful textual information has already been captured by the model during the pre-training stage.

5. Conclusion

We propose a simple yet effective method to learn visual representations in a unified multimodal training framework, composed of two intra-modal and inter-modal learning paths with carefully designed contrastive losses. Extensive experiments across various datasets and tasks demonstrate that we learn high-quality visual features with better scalability and transferability. Since our framework is symmetric for all modalities (e.g. image and caption explored here), it has the flexibility to be extended to other modalities such as video and audio.

References

[1] Jean-Baptiste Alayrac, Adrià Recasens, Rosalia Schneider, Relja Arandjelovic, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks. *CoRR*, abs/2006.16228, 2020.

[2] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. In *NeurIPS 2020*, 2020.

[3] Philip Bachman, R. Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *NeurIPS*, 2019.

[4] Mathilde Caron, Piotr Bojanowski, Julien Mairal, and Armand Joulin. Unsupervised pre-training of image features on non-curated data. In *ICCV*, 2019.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. *CoRR*, abs/2002.05709, 2020.

[6] Xinlei Chen, Haoqi Fan, Ross B. Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *CoRR*, abs/2003.04297, 2020.

[7] 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 *ECCV*, 2020.

[8] Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Comput. Linguistics*, 16(1):22–29, 1990.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[10] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. *CoRR*, abs/2006.06666, 2020.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *NAACL-HLT*, 2019.

[12] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, 2015.

[13] Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. Understanding back-translation at scale. In *EMNLP*, 2018.

[14] Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *Int. J. Comput. Vis.*, 111(1):98–136, 2015.

[15] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. VSE++: improving visual-semantic embeddings with hard negatives. In *BMVC*, 2018.

[16] Vijetha Gattupalli, Yaoxin Zhuo, and Baoxin Li. Weakly supervised deep image hashing through tag embeddings. In *CVPR*. Computer Vision Foundation / IEEE, 2019.

[17] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[18] Lluis Gomez-Bigorda, Yash Patel, Marçal Rusiñol, Dimosthenis Karatzas, and C. V. Jawahar. Self-supervised learning of visual features through embedding images into text topic spaces. In *CVPR*, 2017.

[19] Albert Gordo and Diane Larlus. Beyond instance-level image retrieval: Leveraging captions to learn a global visual representation for semantic retrieval. In *CVPR*, 2017.

[20] Jiuxiang Gu, Jianfei Cai, Shafiq R. Joty, Li Niu, and Gang Wang. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In *CVPR*, 2018.

[21] Jiuxiang Gu, Jason Kuen, Shafiq Joty, Jianfei Cai, Vlad Morariu, Handong Zhao, and Tong Sun. Self-supervised relationship probing. *NeurIPS*, 33, 2020.

[22] Jiuxiang Gu, Gang Wang, Jianfei Cai, and Tsuhan Chen. An empirical study of language CNN for image captioning. In *ICCV*, 2017.

[23] Hao Guo, Kang Zheng, Xiaochuan Fan, Hongkai Yu, and Song Wang. Visual attention consistency under image transforms for multi-label image classification. In *CVPR*, 2019.

[24] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *AISTATS*, 2010.

[25] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

[26] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. In *ICCV*, 2017.

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[28] Olivier J. Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, S. M. Ali Eslami, and Aäron van den Oord. Data-efficient image recognition with contrastive predictive coding. *CoRR*, abs/1905.09272, 2019.

[29] R. Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Philip Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *ICLR*. OpenReview.net, 2019.

[30] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *CVPR*, 2017.

[31] Andrej Karpathy and Fei-Fei Li. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015.

[32] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *CoRR*, abs/2004.11362, 2020.

[33] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *ICLR*, 2015.

[34] Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *CoRR*, abs/1411.2539, 2014.

[35] Ang Li, Allan Jabri, Armand Joulin, and Laurens van der Maaten. Learning visual n-grams from web data. In *ICCV*, 2017.

[36] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. In *Arxiv*, 2019.

[37] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, 2020.

[38] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017.

[39] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV*, 2014.

[40] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: single shot multibox detector. In *ECCV*, 2016.

[41] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[42] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019.

[43] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *CVPR*, 2020.

[44] Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. Scaling neural machine translation. In *WMT*, 2018.

[45] Chao Peng, Tete Xiao, Zeming Li, Yuning Jiang, Xiangyu Zhang, Kai Jia, Gang Yu, and Jian Sun. Megdet: A large mini-batch object detector. In *CVPR*, 2018.

[46] Ariadna Quattoni, Michael Collins, and Trevor Darrell. Learning visual representations using images with captions. In *CVPR*, 2007.

[47] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016.

[48] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *NeurIPS*, 2015.

[49] Mert Bülent Sariyildiz, Julien Perez, and Diane Larlus. Learning visual representations with caption annotations. In *ECCV*, 2020.

[50] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *ICLR*, 2015.

[51] 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 *ICLR*. OpenReview.net, 2020.

[52] Hao Tan and Mohit Bansal. LXMERT: learning cross-modality encoder representations from transformers. In *EMNLP-IJCNLP*, 2019.

[53] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *CoRR*, abs/1906.05849, 2019.

[54] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018.

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

[56] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *CVPR*, 2016.

[57] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018.

[58] Jianming Zhang, Shugao Ma, Mehrnoosh Sameki, Stan Sclaroff, Margrit Betke, Zhe Lin, Xiaohui Shen, Brian L. Price, and Radomír Mech. Salient object subitizing. *Int. J. Comput. Vis.*, 124(2):169–186, 2017.

[59] Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization. In *ECCV*, 2016.

[60] Chengxu Zhuang, Alex Lin Zhai, and Daniel Yamins. Local aggregation for unsupervised learning of visual embeddings. In *ICCV*, 2019.