Self-supervised Visual Attribute Learning for Fashion Compatibility

Donghyun Kim[1], Kuniaki Saito[1], Samarth Mishra[1], Stan Sclaroff[1], Kate Saenko[1,2], Bryan A. Plummer[1]
[1]Boston University, [2]MIT-IBM Watson AI Lab
{donhk, keisaito, samarthm, sclaroff, saenko, bplum}@bu.edu

Abstract

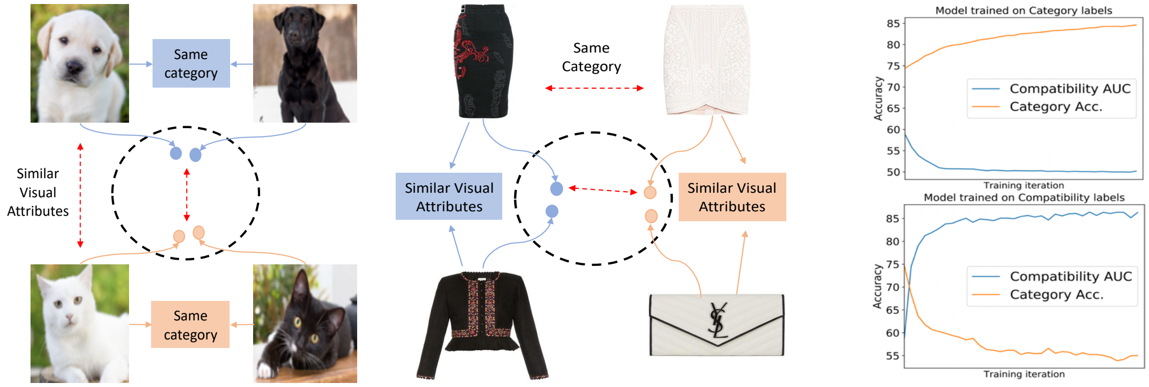
Many self-supervised learning (SSL) methods have been successful in learning semantically meaningful visual representations by solving pretext tasks. However, prior work in SSL focuses on tasks like object recognition or detection, which aim to learn object shapes and assume that the features should be invariant to concepts like colors and textures. Thus, these SSL methods perform poorly on downstream tasks where these concepts provide critical information. In this paper, we present an SSL framework that enables us to learn color and texture-aware features without requiring any labels during training. Our approach consists of three self-supervised tasks designed to capture different concepts that are neglected in prior work that we can select from depending on the needs of our downstream tasks. Our tasks include learning to predict color histograms and discriminate shapeless local patches and textures from each instance. We evaluate our approach on fashion compatibility using Polyvore Outfits and In-Shop Clothing Retrieval using Deepfashion, improving upon prior SSL methods by 9.5-16%, and even outperforming some supervised approaches on Polyvore Outfits despite using no labels. We also show that our approach can be used for transfer learning, demonstrating that we can train on one dataset while achieving high performance on a different dataset.

1. Introduction

Colors and textures information are important features for tasks like fine-grained classification [5, 8] and microscopy image classification [28] as well as applications like image search, recommendation, and outfit generation [4, 11, 16, 32, 34, 35, 37]. However, collecting annotations to train these models can be expensive, especially when they require domain expertise [29] or are constantly evolving like e-commerce datasets. Self-supervised learning (SSL) would appear to be a good fit to address this problem since they require no labels for training, but prior work focused on tasks like object classification and detection (*e.g*. [9, 27, 40, 2, 12]), where the goal is to recognize an object (*i.e*., its shape) regardless of its color or texture (so *a black dog* and *a white dog* should both be classified as *a dog*). In fact, many self-supervised approaches are explicitly designed to learn color invariant features [2, 12]. Thus, as we illustrate in Figure 1, prior work in SSL often does not generalize to tasks where colors and textures are important.

In this paper, we propose Self-supervised Tasks for Visual Attribute (S-VAL) to learn visual attributes while generating *shape invariant* features for fashion compatibility, where a system recommends fashion items compatible and complement each other when worn together in an outfit. Motivated by the observation that similar color or texture items are likely to be compatible [30], S-VAL is designed to learn embedding images with similar colors and texture patterns are embedded nearby each other. To be specific, our approach consists of three major components. First, we propose a new self-supervised pretext task where a model predicts color histograms of input images to understand dominant colors of an image. Second, we introduce shapeless local patch discrimination, where we perform Instance Discrimination (ID) [40] on very small image patches of an image. This helps ensure that little shape information is present in an image and the model must focus on recognizing color and texture information instead. Finally, we obtain texture features using a Gram matrix [7, 20, 19] computed over the whole image, and then encourage ID to learn discriminative texture representations. Our approach uses no labels in training (*i.e*., unsupervised), but, as our experiments will show, we get comparable performance to some fully-supervised methods. Figure 2 provides an overview of our approach.

The work that is the closest in spirit to ours is Hsiao *et al*. [16], which automatically identifies individual clothing items from full-body photos of people and then uses the parsed outfits as labels for fashion compatibility. This is reminiscent of the part-based methods used in tasks like object classification [6], where the goal is to learn how to identify the parts (or individual clothing items) in order to recognize the object (or to recognize compatible items). However, this still requires having weak-labels and is a task-specific application (*i.e*., it only is applicable to fashion compatibility). Another significant drawback is that the images used



(a) Object Recognition (b) Fashion Compatibility (c) Non-transferability

Figure 1: Differences between the (a) object recognition and (b) fashion compatibility tasks. (a) Object recognition needs *color invariant* but *shape sensitive* features. (b) Tasks like fashion compatibility needs *color sensitive* but *shape invariant* features in order to match different category fashion items, in which items of the same object category can be embedded far under different visual attributes. In (c), we show that a model trained on object category labels hurts performance on the fashion compatibility task and vice versa, which helps motivate us to propose a new form of SSL pretext tasks.

for training were from a different domain (full-body images of people) than the images they are evaluated on (images containing a single product on a white background). Thus, as our experiments will show, our approach significantly outperforms the weakly-supervised approach of Hsiao *et al*. [16] despite our approach lacking any supervision and without making task-specific assumptions. Specifically, in addition to comparing to Hsiao *et al*. on fashion compatibility, we also evaluate our approach on In-Shop Clothing Retrieval [23], demonstrating that our approach generalizes.

Our contributions are summarized below:

- We propose Self-supervised Tasks for Visual Attribute (S-VAL) to learn *colors and textures of images* while generating *shape invariant* features. To the best of our knowledge, ours is the first work to propose SSL methods for capturing color and texture information.
- We obtain a 9.5-16% gain in fill-in-the-blank outfit completion using Polyvore Outfits [35] and on In-Shop Retrieval using DeepFashion [23] over prior SSL methods. Notably, our approach outperforms some supervised methods on Polyvore Outfits despite using no labels.
- We show our approach creates powerful features that transfer across datasets. Specifically, we train on Polyvore Outfits and test on Capsule Wardrobes [16], and train on the Fashion-Gen dataset [31] and test on Polyvore Outfits, reporting a 6-8% gain over prior work.
- We demonstrate that self-supervised learning should consider *different characteristics of downstream tasks* by highlighting the difference between object recognition and tasks like fashion compatibility and image retrieval, which we hope inspires future work in SSL.

2. Related Work

Self-supervised Learning (SSL). Self-supervised learning [9, 27, 40, 12, 2, 26] generates self-supervisory signals for a pretext task from an input. By solving a pretext task, a model can learn semantically meaningful features from raw data. Handcrafted pretext tasks such as predicting rotations [9] and solving jigsaw puzzles [27] provide useful features for object recognition and detection tasks. Wu *et al*. [40] proposes an Instance Discrimination (ID) pretext task with contrastive loss [10]. ID learns visual similarity in different images by treating an image as its own class (*i.e*., positive pair) but all other images as negative pairs. While ID is effective at learning strong visual representations, ID can be biased to texture or colors of an object which is harmful to objection recognition. In later work, ID with strong data augmentation techniques like color distortion (*e.g*., color jittering and gray-scale images) [2, 3] significantly improved the recognition or detection performance by providing color and texture invariant features. While these perform well for a task like object recognition or detection, they focus on learning an object's shape. For example, the images of "black dog" and "white dog" should be classified as the same class "dog" as shown in Fig. 1(a), so that "black" and "white" attribute should ideally be ignored. However, many tasks require reasoning about multiple similarity notions such as color, texture and style these methods ignore. Specifically, in tasks like fashion compatibility, where two items are considered compatible if they would complement each other when worn in the same outfit, items of different categories (*e.g*., a shirt and pants) can be compatible with each other. Thus, learning object categories can be harmful to performance (Fig. 1(c)). Instead, we propose an SSL framework that learns visual attributes for tasks where color and texture are important.

Fashion Compatibility. Other than the weakly-supervised approach of Hsiao *et al*. [16] we discussed in the Introduction, much of the recent work on fashion compatibility has assumed labels are available during training [4, 11, 25, 35, 41, 32, 21, 37]. Many of these approaches aim to decompose the fashion compatibility task into similarity conditions that may be learned automatically [32, 21] or could be explicitly defined [25, 35, 41, 36]. All of these methods require many labels of positive pairs and arbitrarily choose negative samples, since datasets are not annotated with incompatible items, which can result in poor constraints [39]. Also, as our experiments will show, we outperform some supervised fashion compatibility methods without using any supervision.

Visual Attribute Learning. Visual attributes such as colors (*e.g*., red, blue), texture (*e.g*., palm, colorblock), or fabric (*e.g*., leather, tweed) provide natural visual patterns of fashion items. In order to learn these visual attributes in items, some methods [38, 1] leverage visual attribute labels such as color or style extracted from text descriptions. However, these attribute labels can be very sparse and highly non-curated. Plummer *et al*. [30] introduce an attribute explanation model to find salient attributes for fashion item matching and find that colors are the one of the most salient attributes. Our SSL learns colors of fashion items and embed them near each other to build better representations for the task of fashion compatibility.

3. S-VAL: Self-supervised Tasks for Visual Attribute Learning

We explore image similarity learning under an unsupervised setting where we have only unlabeled images $\mathcal{D} = \{(\mathbf{x}_i)\}_{i=1}^{N}$. These items include items of different categories such as pants, tops, and shoes. Compared to prior work in self-supervised learning (SSL), our approach aims to learn visual attributes without encoding any shape clues which could hurt downstream task performance (*i.e*., shape-invariant features). Our SSL approach consists of three sub-tasks: (1) predicting color histograms, (2) shapeless local patch discrimination (SLPD), and (3) texture discrimination (TD). We train a model with three sub-tasks jointly. Our model consists of a CNN feature extractor $F(\cdot) \in \mathbb{R}^n$ and separate projection heads $C(\cdot)$ for each sub-tasks. Figure 2 contains an overview of our method.

3.1. Predicting Color Histogram

Colors are salient attributes in tasks fashion compatibility [30, 33, 41] or microscopy image classification [28]. Thus, a color histogram of an item can provide useful properties of an image including its colors, contrast, and brightness of an item. In contrast to previous color reconstruction methods such as AutoEncoders [15], we learn to predict an RGB color histogram, which is an *orderless* visual representation and therefore does not encode shape information [22]. This means that objects from different categories (*e.g*., black top and black pants) can be embedded closely in the color embedding space. Given an image $\mathbf{x}$ with width w and height h, we first compute the normalized histogram of n bins for each $R, G,$ and$B$ channels, for example,

$$h_r(l) = \frac{|\{i,j\} : e_l \leq \mathbf{x_r}(i,j) < e_{l+1}|}{w \times h} \tag{1}$$

where h_r represents the histogram of the R channel of the image (*i.e*., $\mathbf{x_r}$) and e_l is the l-th bin edge. h_g and h_b are defined similarly. In the case we are learning a presentation for product images commonly found in e-commerce websites, we exclude any white background pixel values.

From the image feature from a CNN (*i.e*., $\mathbf{f} = F(\mathbf{x})$), we compute predictions of histograms for the R channel $C_r(f) \in \mathbb{R}^n$, G channel $C_g(f) \in \mathbb{R}^n$, and B channel $C_b(f) \in \mathbb{R}^n$. In order to obtain the probability distributions of each channel (*i.e*. $p_r, p_g,$ and, p_c), we apply the softmax function. Then, we minimize the KL divergence between predicted distribution and the ground-truth histogram,

$$\mathcal{L}_{rgb} = D_{KL}\left[p_r \| h_r\right] + D_{KL}\left[p_g \| h_g\right] + D_{KL}\left[p_b \| h_b\right] \tag{2}$$

3.2. Shapeless Local Patch Discrimination (SLPD)

While predicting histogram captures the dominant colors in images, it lacks in learning detailed color patterns such as the spatial organization of colors and textons in fashion items. In this section, we aim to learn discriminative color or texture representations by using shapeless local patches. In previous SSL methods, strong augmentation techniques with color distortion with Instance Discrimination (ID) [40, 2, 3] can be used together to become invariant to color or texture information so they learn to better identify shapes. While this may be appropriate for tasks like object recognition, as shown in Fig. 1(c), learning shape information harms performance on tasks like fashion compatibility where image similarity is not determined completely by an item's shape.

To avoid focusing on shape, we perform ID on shapeless small local patches (SLP) that contain little or no shape information. Figure 2 shows examples of the SLPs. While random cropping has been used in prior work [2, 40], they often use relatively large cropping ratios r (*i.e*., [0.2, 1.0]) to maximize the consensus between local-to-global views. However, these will often contain shape information, whereas SLP use very small ratio values of r (*e.g*., $r = 0.05$) to lose such information. As such, a model must learn to discriminate between color and texture information rather than shape, which we found often performs better.

To perform the shapeless local patch discrimination, we first initialize the memory bank $\boldsymbol{V}$ to store features of all

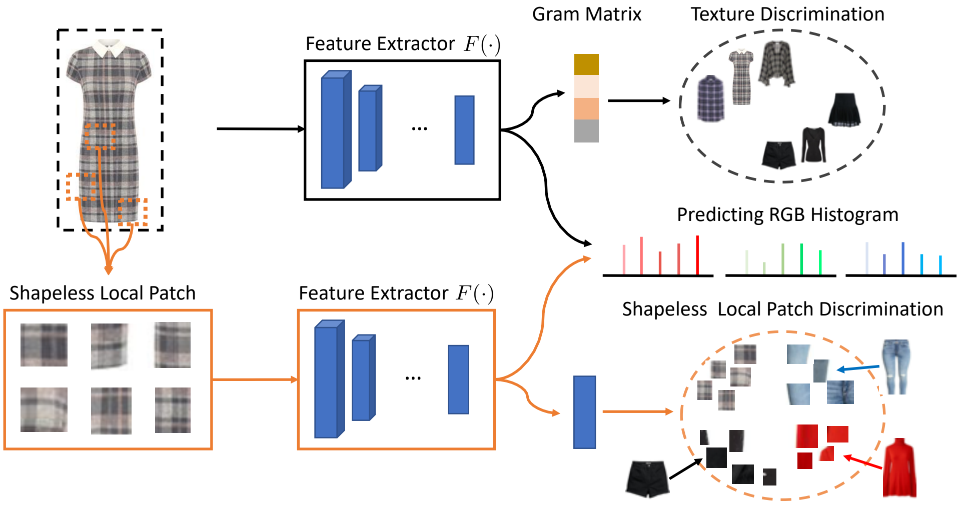

Gram Matrix
Texture Discrimination
Feature Extractor $F(\cdot)$
...
Predicting RGB Histogram
Shapeless Local Patch
Feature Extractor $F(\cdot)$
Shapeless Local Patch Discrimination

Figure 2: An overview of our Self-supervised Tasks for Visual Attribute (S-VAL), where we aim to learn discriminative features in colors and textures without encoding shape information. To achieve this goal, we propose thee sub-tasks (1) predicting RGB histogram, (2) shapeless local patch discrimination, and (3) texture discrimination.

training images,

$$\boldsymbol{V} = [\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_N] \tag{3}$$

where $\mathbf{v_i}$ is the feature of the shapeless local patch $\mathbf{x}'_i$ from the i-th original image $\mathbf{x}_i$ (*i.e.*, $\mathbf{v_i} = C_{SLPD}(F(\mathbf{x}'_i))$ and N is the total number of images. We randomly choose a square SLP $\mathbf{x}'_i$ out of the whole image (*e.g.*, a random region cropped with $r = 0.05$ of the whole area). Then, given an image $\mathbf{x}'_j$ in a minibatch , we compute the feature $\mathbf{f}_j = C_{SLPD}(F(\mathbf{x}'_j)$ minimize the contrastive loss [40] to discriminate the shapeless local patch,

$$\mathcal{L}_{SLPD} = -\log \frac{\exp((\mathbf{v}_j)^\top \mathbf{f}_j / \tau)}{\sum_{k=1}^{N} \exp((\mathbf{v}_k)^\top \mathbf{f}_j / \tau)}, \tag{4}$$

where the temperature parameter τ is the concentration level [14].

3.3. Texture Discrimination (TD)

Unlike the SLPD, texture discrimination (TD) uses the whole image to learn global texture patterns. Inspired by [20, 7], we use a gram matrix (also called bilinear features) to obtain a texture representation for an image. Then, similar to the SLPD, we perform ID so items with similar textures embed nearby each other. First, we compute the feature map $\mathbf{g}_i = C_{TD}(F(x_i))$ of an input image $\mathbf{x_i}$ and a Gram matrix for texture representation [20, 7],

$$\mathbf{G}_i(j,k) = \mathbf{g}_i(j)\mathbf{g}_i(k) \tag{5}$$

where $\mathbf{G}(j,k)$ is the inner product between the vectorized features of j-th and k-th channels in the feature map $\mathbf{g}_i$. In order to perform texture discrimination, we initialize the memory bank $\boldsymbol{T}$ to store texture representation of all training images.

$$\boldsymbol{T} = [\mathbf{t}_1, \mathbf{t}_2, \cdots, \mathbf{t}_N] \tag{6}$$

where $\mathbf{T}_i$ is the texture representation of i-th image (*i.e.*, $\mathbf{T}_i = \mathbf{G}_i(j,k)$). During training, similar to above, we compute the texture representation $\mathbf{G}_j$ of x_j in a minibatch and minimize the contrastive loss [40] to discriminate texture representations between images,

$$\mathcal{L}_{TD} = -\log \frac{\exp((\mathbf{t}_j)^\top \mathbf{G}_j / \tau)}{\sum_{k=1}^{N} \exp((\mathbf{t}_k)^\top \mathbf{G}_j / \tau)}, \tag{7}$$

Finally, the overall learning objective for S-VAL is,

$$\hat{\theta} = \lambda_{rgb}\mathcal{L}_{rgb} + \lambda_{SLPD}\mathcal{L}_{SLPD} + \lambda_{TD}\mathcal{L}_{TD} \tag{8}$$

where $\lambda_{rgb}, \lambda_{SLPD}, \lambda_{TD}$ are the hyper-parameters for each loss. SLPD takes only shapeless local patches as input and TD takes the whole image to understand the global textures. Predicting the RGB histogram takes both types of input.

After updating the network parameters with each minibatch B, we also update the memory features in the memory banks $\boldsymbol{V}$ and $\boldsymbol{T}$ with a momentum $\eta = 0.5$ following [40]:

$$\begin{aligned} \forall i \in B, \quad \mathbf{v}_i &= (1-\eta)\mathbf{v}_i + \eta C_{SLPD}(\mathbf{f}_i), \\ \mathbf{t}_i &= (1-\eta)\mathbf{t}_i + \eta C_{TD}(\mathbf{f}_i), \end{aligned} \tag{9}$$

4. Experiments

Following Han *et al.* [11], we evaluate on the fashion compatibility and fill-in-the-blank (FITB) tasks as described below. We denote the feature of an image $\mathbf{x}_i$ as $\mathbf{f}_i = F(\mathbf{x}_i)$.

Fashion Compatibility. In this task the goal is to discriminate between compatible and incompatible outfits. Following Han *et al.* [11], we report the area under a receiver operating characteristic curve (AUC) and average precision (AP) from the compatibility scores. Given N fashion items in an outfit, we compute the compatibility scores by computing the average pair-wise cosine similarities: $\frac{2}{N(N-1)} \sum_{i=0}^{N-1} \sum_{j=i+1}^{N-1} cos\_sim(\mathbf{f}_i, \mathbf{f}_j)$.

Fill in the Black (FITB). In this task the goal is to complete a partial outfit by selecting from a set of options. Similar to above, we compute the average similarity between each option and the partial outfit and select the one that gets the highest average compatibility. Performance is measured based on how often the choice was correct.

Fashion Retrieval. We also explore the fashion retrieval task. In fashion retrieval, the goal is to find the same item from a database given a query item that may be in a different view than those in the database. Similar to the fashion compatibility task, this task also needs some understanding of colors and textures, but shape also plays a factor since we are looking for exactly the same object. However, the shape of fashion items can still be changed significantly, as the items can appear in different poses, illumination, and camera angles. We report recall@k as our metric.

Implementation details. We use a ResNet-50 [13] which is pre-trained on ImageNet [18] for our feature extractor $F(\cdot)$ and all baselines. For each sub-tasks in Sec. 3, we attach the separate projection heads after the feature extractor. Following [2], these heads consist of two fully connected layers with ReLU activations followed by a ℓ_2 normalization layer. All three self-supervised sub-tasks are trained jointly. We use each validation set to tune hyper-parameters for each sub-task and report averaged results over three runs. We randomly sample shapeless local patches with $r \in$ [0.05, 0.15] of the original image area. We use a Adam optimizer optimizer [17] with a learning rate $5e^{-5}$. We train a model for 150 epochs and set the number of bins for each RGB channel as 10 and hyper-parameters $\lambda_{rgb} = 1, \lambda_{SLPD} = 1e^{-2}, \lambda_{TD} = 1e^{-5}$ in Eq. 8 using the validation set [35]. We set $\tau = 0.07$ in Eqs. 4 and 7 following [40].

We also provide the following self-supervised baselines for comparison: AutoEncoder [15], colorization [42], sovling jigsaw puzzles [27], predicting rotation [9], Instance Discrimination (ID) [40], and Local Aggregation [43]. Please note that all methods finetune the same ResNet-50 initialized with ImageNet pretrained weights as our approach.

4.1. Datasets

Polyvore Outfits [35] has 53,306 outfits from 204K images for training, 10K outfits from 47K images for testing and 5K outfits from 25K images for validation. We use the provided fashion compatibility and FITB questions, where items in ground truth outfits were replaced with random items of the same type for fashion compatibility, or 3 random items of the same type were selected as incorrect answers for FITB (resulting in 4 choices). We also use the Polyvore-D split that contains 71K images. In this split no item that appears in the training outfits also appears in the testing outfits.

Capsule Wardrobes [16] contains 15K fashion compatibility questions from 6K images, which are all used for testing. We train on the Polyvore Outfits dataset when evaluating on Capsule Wardrobes.

Fashion-Gen [31] has 260K images of luxury fashion items with descriptions. We only train on this dataset and evaluate on Polyvore Outfits since no outfit information is publicly available.

In-Shop Clothing Retrieval benchmark in DeepFashion [23] contains 52K images of 8K clothing items from web data containing large poses and scale variations. This benchmark splits its test data into a query and gallery set, where no items in either of these sets are shared with those seen during training.

4.2. Unsupervised Evaluation Results

Table 1 contains results on the Polyvore [35] and Capsule Wardrobe test set [16]. In Table 1(a), we report the performance of supervised models with trained compatibility labels or attribute labels in Polyvore as a reference. In Table 1(b), we report the performance of the self-supervised learning baselines fine-tuned on Polyvore from the ImageNet pretrained model. We see that existing self-supervised learning methods including reconstruction based methods [42, 15] and handcrafted sub-tasks [9, 27] actually harm performance compared to the ImageNet pre-trained model. We also observe that ID and Local Aggregation with color distortion underperform the ImageNet pre-trained model. When we remove the color distortion augmentation in their methods, these methods outperform the ImageNet pre-trained model. These results suggest that directly applying the existing self-supervised learning methods does not help on the fashion compatibility task. From now on, we remove the color distortion augmentation in ID and Local Aggregation for all other comparisons.

We show the performance of our method in Table 1(c) including an ablation analysis. We observe that each subtask predicting RGB histograms (Sec. 3.1), shapeless local patch discrimination (Sec. 3.2), and texture discrimination (Sec. 3.3), improves the performance over the ImageNet pretrained network. Combing all three components gets the best performance, resulting in a 9.5-10 points improvement on Polyvore Outfits over prior SSL baselines, and 4 points better on Capsule Wardrobes. In addition to outperforming the SSL baselines, our full model without any labels outperforms simple the Simaese Network trained with compatibility labels,

	Method	Label?	Polyvore Outfits		Capsule
			Comp. AUC	FITB acc.	Comp. AP
(a) With Label	Bi-LSTM [11]	Comp.	0.65	39.7	18.4
	SiameseNet [35]	Comp.	0.81	52.9	-
	Type-Aware Network [35]	Comp.	0.86	55.3	-
	SCE-Net [32]	Comp.	**0.91**	**61.6**	-
	Attribute Classifier	Attributes	0.73	46.3	25.0
(b) Self-sup. Baselines	ImageNet pre-trained	✗	0.66	39.1	21.1
	Capsule Network (weakly-sup.) [16]	✗	-	-	19.9
	AutoEncoder [15]	✗	0.58	34.0	19.8
	Colorization [42]	✗	0.63	34.1	18.6
	Jigsaw [27]	✗	0.52	27.9	18.6
	Rotation [9]	✗	0.53	29.4	18.5
	ID [40] w/ color distortion	✗	0.57	30.8	18.9
	ID [40] w/o color distortion	✗	0.74	45.9	23.3
	LA [43] w/ color distortion	✗	0.56	30.4	19.1
	LA [43] w/o color distortion	✗	0.74	46.3	24.0
(c) S-VAL (Ours)	Predicting RGB histogram (RGB)	✗	0.77	47.2	23.3
	Shapeless Local Patch Disc. (SLPD)	✗	0.83	54.6	27.7
	Texture Disc (TD)	✗	0.77	50.3	25.2
	RGB + SLPD	✗	0.83	55.4	27.7
	RGB + SLPD + TD	✗	**0.84**	**55.8**	**27.9**

Table 1: Comparison of (a) supervised models with compatibility or attribute labels and (b,c) unsupervised models on the Polyvore Outfits [35] and Capsule [16] datasets. *All methods use ImageNet pre-trained weights and finetuned on Polyvore Outfits.* We report the performance of existing self-supervised learning baselines in (b) and our proposed approach in (c).

	Polyvore-D	
Method	Comp. AUC	FITB acc.
ID [40]	0.69	43.2
LA [43]	0.73	46.2
RGB	0.74	45.7
RGB+SLPD	**0.81**	53.9
RGB+SLPD+TD	**0.81**	**54.3**

Table 2: Fashion compatibility evaluation on the Polyvore-D Split. The Polyvore-D split containg less training data than Polyvore. Our method outperforms the baselines.

	Fashion-Gen → Polyvore	
Method	Comp. AUC	FITB acc.
ID [40]	0.71	45.5
LA [43]	0.73	46.5
RGB	0.76	48.1
RGB+SLPD	0.80	52.9
RGB+SLPD+TD	**0.81**	**53.3**

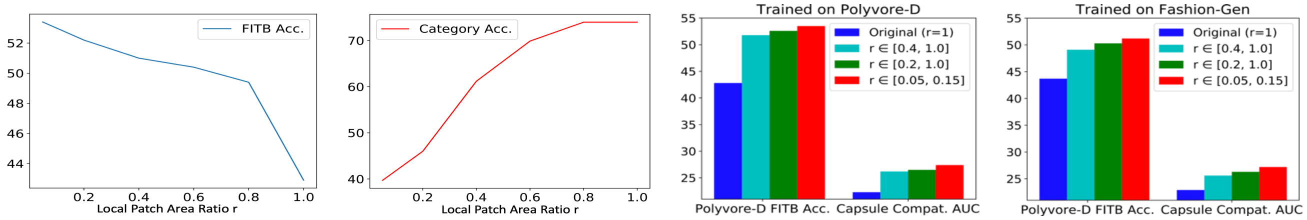
Table 3: Cross dataset evaluation on the fashion compatibility task. We train a model on the Fashion-Gen dataset and test it on the Polyvore dataset. We report the number of self-supervised learning baselines and ours. Our method is able to generalize across different datasets.

as well as Bi-LSTM [11], while also being comparable to the fully-supervised Type-Aware Network.

4.3. Additional Analysis

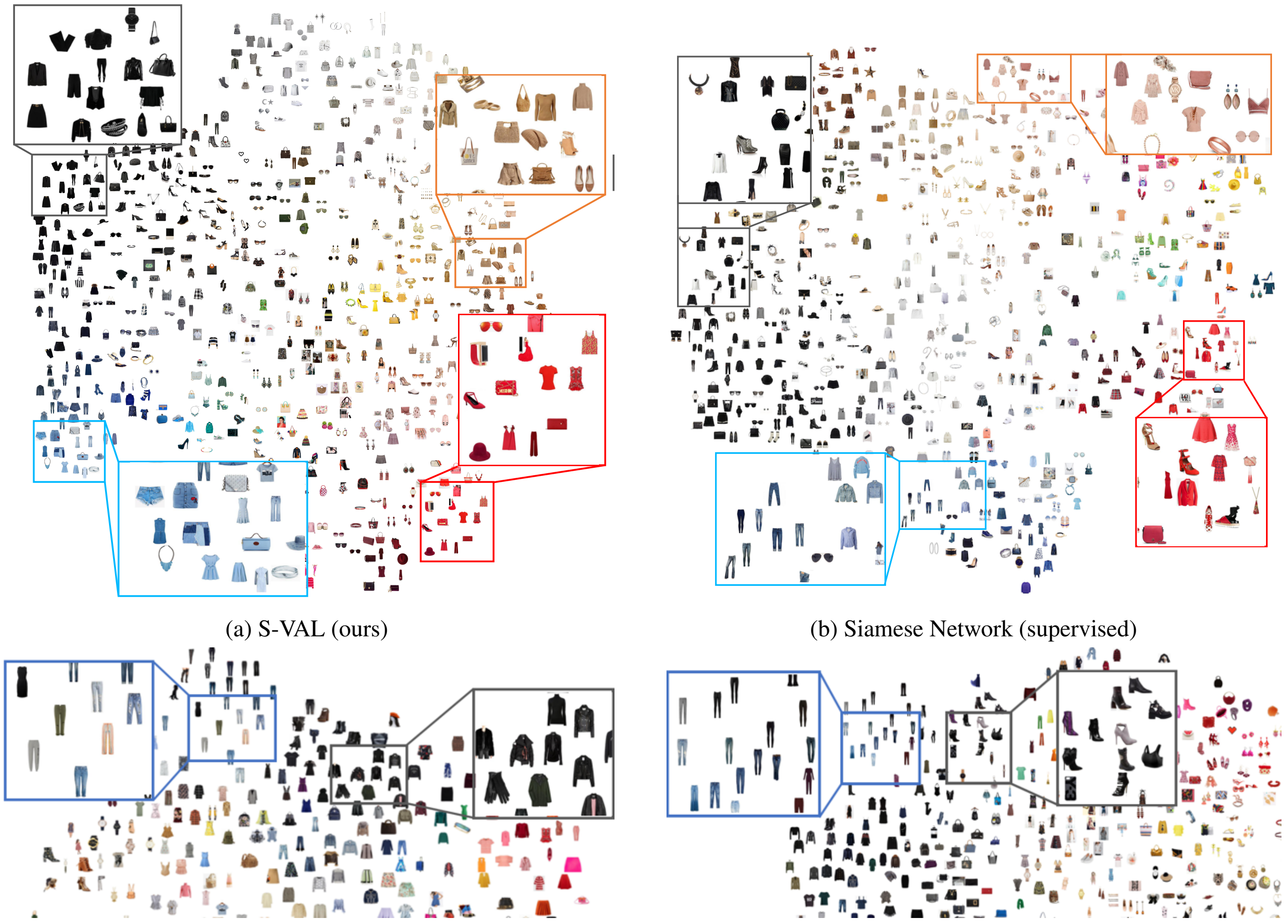
Polyvore-D and Cross Dataset Evaluation. Table 2 shows the comparison on Polyvore-D containing three times fewer training images than Polyvore Outfits. Table 3 explores a cross dataset evaluation scenario, where a model is trained on Fashion-Gen but evaluated on Polyvore Outfits. In both cases, our approach outperforms the best SSL baseline, Local Aggregation, by 8-9 points on both tasks.

Ablation Study on Patch Area Ratio. In this section, we analyze how the different area ratios affect the performance on both fashion compatibility and object recognition (denoted by "Category Acc") in Fig. 3. We measure the object recognition accuracy with a kNN classifier [40] on image features. In Fig. 3(a), we report the FITB accuracy using different local patch sizes. It is clear that using a small local patch improves performance considerably over using a large local patch. Fig. 3(b) reports category recognition accuracy, which appears to have an inverse relationship with 3(a), demonstrating that addressing fashion compatibility requires different methods than typically used in prior work in SSL that mainly investigated methods for object recognition. Finally in Fig. 3(c, d), we compare models trained with ID using different area ratios r: original image only (*i.e.*, $r = 1.0$) and different random cropping ratios of $r \in [0.4, 1.0], [0.2, 1.0], [0.05, 0.15]$. We see that using larger patches harms the performance compared to using smaller patches only. These results also suggest that the performance gain mostly comes from the small patches. Thus,



(a) (b) (c) (d)

Figure 3: Ablation study on the effect of local patch area ratio r on Polyvore-D. In (a,b), we report the performances of the task of fashion compatibility and object recognition according to the different area ratios of the local patch. In (c,d), we provide the comparisons on original input size r=1 and random cropping with different ratios in the specified range during training. These results show that using smaller patches performs better while generating shape invariant features than using larger patches.



(a) S-VAL (ours)

(b) Siamese Network (supervised)

(c) ImageNet pre-trained Network

(d) ID

Figure 4: t-SNE visualizations. Similar to (b) the supervised model, (a) our unsupervised model learns a similar embedding which embeds items with similar visual attributes (*e.g.*, colors and texture) nearby regardless of object categories. While the ImageNet pre-trained network and ID generate features biased to object shapes, items with different visual attribute can be embedded nearby.

training with very small local patches losing shape clues is a key component in SSL for fashion compatibility.

Visualization. Figure 4 shows t-SNE visualizations [24] of features on Polyvore from each model. We also confirm the observation of [30] that the Siamese network trained on compatibility labels embeds similar color or texture items nearby ignoring fashion item categories (the third row in Table 1(a)). By comparing the Fig. 1(a, b), our model produces a very similar feature distribution as the Siamese network. Both models tend to cluster similar items nearby in terms of colors and texture regardless of object categories. However, Fig. 4(c) and (d) cluster items based on shape, so that items with different attributes from the same object class are embedded nearby, which could be harmful to the fashion compatibility task as discussed earlier.

Linear Classification Protocol. We evaluate our method on a linear classification protocol [40, 12, 2]. In this evaluation, we use fixed image features $\mathbf{f} \in \mathbb{R}^{2048}$ and train only a linear classifier $\mathbf{W} \in \mathbb{R}^{2048\times 64}$ on compatibility labels using triplet loss. To effectively evaluate the features learned from SSLs, we report performance when different numbers of training labels are available in Fig. 5. We compare ours with the ImageNet pre-trained network and Local Aggregation, which is the best performing self-supervised baseline. We observe that our method consistently outperforms other baselines and the benefit of our method is more significant when there are fewer labels.

Fashion Retrieval Evaluation. In Table 4, we report the accuracy of recall@k of the DeepFashion Inshop retrieval task [23]. We start with an ImageNet pre-trained model and use SSL methods without labels. As a reference, we show the accuracy of a fully supervised model with 200K pair annotations in the first row of Table 4. We use a standard triplet loss to train the supervised model similar to a Siamese Network in [35]. Different from the fashion compatibility task, the ID with color distortion augmentations improves performance compared to the ImageNet pre-trained model. By removing the color distortion augmentations, ID further improves the performance. As expected, this result shows that shape information is helpful to learn useful features for the fashion retrieval task. Then we perform each of the components in S-VAL. In this task, predicting RGB histogram does not help much. This could be because predicting RGB histogram does not consider item shapes and enforce a model to produce invariant features to object shapes. As learning shape is also important to retrieve the same category item in this task, predicting RGB histogram is not desirable. We see that SLPD and TD outperform ID by a large margin by learning color patterns in a local patch and global texture patterns from TD. These results suggest that directly applying the existing method on any downstream task is not the best option. We argue that self-supervised learning methods should consider the characteristics of a downstream task.

5. Conclusion

While prior self-supervised learning approaches have been successful, their downstream task is mostly related to object recognition which focuses on learning object shape variant and color invariant features. In this paper, we explore self-supervised methods for the fashion compatibility and retrieval task, where colors and texture are important. We propose a new Self-supervised Tasks for Visual Attribute Learning (S-VAL) which learns colors and texture patterns while generating shape-invariant features. Our method is built upon an observation that similar color or texture items are more likely compatible, but it is possible that different color items can be matched. We also show that prior work in self-supervised learning often fails to generalize to computer vision tasks that require a model that learns visual cues other than object shape. On the fashion compatibility task, S-VAL outperforms prior self-supervised learning approaches by 9.5-16% and by 16.5% in the fashion retrieval task. Notably, our approach obtained similar performance to some fully-supervised methods from prior work of fashion compatibility despite the fact our approach does not use any labels. We hope that our work will inspire research in self-supervised learning in additional application areas, as well as provide valuable insights to improve fashion recommendation systems in future work.

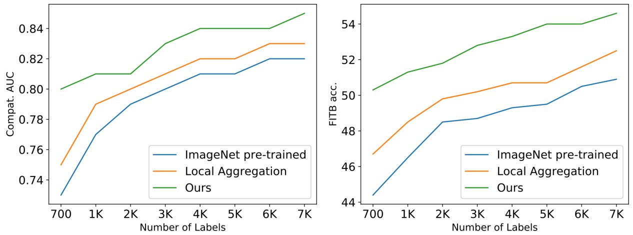


Figure 5: Comparison under linear classification protocol with fashion compatibility labels. “Ours” denotes our full method, RGB + SLPD + TD.

	DeepFashion In-shop Retrieval [23]		
Method	Recall@1	Recall@5	Recall@10
Triplet (supervised)	63.6	85.4	89.3
ImageNet pre-trained	16.8	36.4	42.5
ID w/ color distortion	25.5	51.5	60.1
ID w/o color distortion	30.0	56.1	67.6
RGB	25.0	48.5	55.6
SLPD	39.9	64.6	70.6
TD	29.7	54.9	64.2
SLPD+TD	**46.5**	74.8	**81.3**
SLPD+TD+RGB	46.2	**75.0**	81.2

Table 4: Evaluation on the in-shop fashion retrieval task [23]. The top row reports the accuracy of the supervised model with a triplet loss for reference. We report the unsupervised fashion retrieval accuracy using Recall@k.

6. Acknowledgments

This work was supported by the NSF and the DARPA LwLL program.

References

[1] Tamara L Berg, Alexander C Berg, and Jonathan Shih. Automatic attribute discovery and characterization from noisy web data. In *European Conference on Computer Vision*, pages 663–676. Springer, 2010. 3

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 1, 2, 3, 5, 8

[3] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2, 3

[4] Guillem Cucurull, Perouz Taslakian, and David Vazquez. Context-aware visual compatibility prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12617–12626, 2019. 1, 3

[5] Kun Duan, Devi Parikh, David Crandall, and Kristen Grauman. Discovering localized attributes for fine-grained recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 1

[6] Ian Endres, Kevin J. Shih, Johnston Jiaa, and Derek Hoiem. Learning collections of part models for object recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013. 1

[7] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015. 1, 4

[8] Timnit Gebru, Judy Hoffman, and Li Fei-Fei. Fine-grained recognition in the wild: A multi-task domain adaptation approach. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017. 1

[9] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018. 1, 2, 5, 6

[10] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006. 2

[11] Xintong Han, Zuxuan Wu, Yu-Gang Jiang, and Larry S Davis. Learning fashion compatibility with bidirectional lstms. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1078–1086, 2017. 1, 3, 4, 5, 6

[12] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019. 1, 2, 8

[13] 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. 5

[14] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 4

[15] Geoffrey E Hinton and Richard S Zemel. Autoencoders, minimum description length and helmholtz free energy. In *Advances in neural information processing systems*, pages 3–10, 1994. 3, 5, 6

[16] Wei-Lin Hsiao and Kristen Grauman. Creating capsule wardrobes from fashion images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7161–7170, 2018. 1, 2, 3, 5, 6

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

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 5

[19] Tsung-Yu Lin and Subhransu Maji. Visualizing and understanding deep texture representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2791–2799, 2016. 1

[20] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear cnn models for fine-grained visual recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 1449–1457, 2015. 1, 4

[21] Yen-Liang Lin, Son Tran, and Larry S. Davis. Fashion outfit complementary item retrieval. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[22] Li Liu, Jie Chen, Paul Fieguth, Guoying Zhao, Rama Chellappa, and Matti Pietikäinen. From bow to cnn: Two decades of texture representation for texture classification. *International Journal of Computer Vision*, 127(1):74–109, 2019. 3

[23] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016. 2, 5, 8

[24] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. 7

[25] Samarth Mishra, Zhongping Zhang, Yuan Shen, Ranjitha Kumar, Venkatesh Saligrama, and Bryan A. Plummer. Effectively leveraging attributes for visual similarity. In *The IEEE International Conference on Computer Vision (ICCV)*, 2021. 3

[26] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. *arXiv preprint arXiv:1912.01991*, 2019. 2

[27] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pages 69–84. Springer, 2016. 1, 2, 5, 6

[28] Wei Ouyang, Casper Winsnes, Martin Hjelmare, Anthony Cesnik, Lovisa Åkesson, Hao Xu, Devin Sullivan, Shubin Dai, Jun Lan, Park Jinmo, Shaikat Mahmood Galib, Christof Henkel, Kevin Hwang, Dmytro Poplavskiy, Bojan Tunguz, Russel Wolfinger, Yinzheng Gu, Chuanpeng Li, Jinbin Xie, and Emma Lundberg. Analysis of the human protein atlas image classification competition. *Nature Methods*, 16:1254–1261, 12 2019. 1, 3

[29] Christian S Perone, Pedro Ballester, Rodrigo C Barros, and Julien Cohen-Adad. Unsupervised domain adaptation for

medical imaging segmentation with self-ensembling. *NeuroImage*, 194:1–11, 2019. 1

[30] Bryan A. Plummer, Mariya I. Vasileva, Vitali Petsiuk, Kate Saenko, and David Forsyth. Why do these match? explaining the behavior of image similarity models. In *The European Conference on Computer Vision (ECCV)*, 2020. 1, 3, 7

[31] Negar Rostamzadeh, Seyedarian Hosseini, Thomas Boquet, Wojciech Stokowiec, Ying Zhang, Christian Jauvin, and Chris Pal. Fashion-gen: The generative fashion dataset and challenge. *arXiv preprint arXiv:1806.08317*, 2018. 2, 5

[32] Reuben Tan, Mariya I Vasileva, Kate Saenko, and Bryan A Plummer. Learning similarity conditions without explicit supervision. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10373–10382, 2019. 1, 3, 6

[33] Pongsate Tangseng and Takayuki Okatani. Toward explainable fashion recommendation. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 2153–2162, 2020. 3

[34] Ivona Tautkute, Aleksandra Możejko, Wojciech Stokowiec, Tomasz Trzciński, Łukasz Brocki, and Krzysztof Marasek. What looks good with my sofa: Multimodal search engine for interior design. In *Proceedings of the 2017 Federated Conference on Computer Science and Information Systems*, volume 11 of *Annals of Computer Science and Information Systems*, pages 1275–1282, 2017. 1

[35] Mariya I Vasileva, Bryan A Plummer, Krishna Dusad, Shreya Rajpal, Ranjitha Kumar, and David Forsyth. Learning type-aware embeddings for fashion compatibility. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 390–405, 2018. 1, 2, 3, 5, 6, 8

[36] Andreas Veit, Serge Belongie, and Theofanis Karaletsos. Conditional similarity networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 830–838, 2017. 3

[37] Andreas Veit, Balazs Kovacs, Sean Bell, Julian McAuley, Kavita Bala, and Serge Belongie. Learning visual clothing style with heterogeneous dyadic co-occurrences. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4642–4650, 2015. 1, 3

[38] Sirion Vittayakorn, Takayuki Umeda, Kazuhiko Murasaki, Kyoko Sudo, Takayuki Okatani, and Kota Yamaguchi. Automatic attribute discovery with neural activations. In *European Conference on Computer Vision*, pages 252–268. Springer, 2016. 3

[39] Chao-Yuan Wu, R Manmatha, Alexander J Smola, and Philipp Krahenbuhl. Sampling matters in deep embedding learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2840–2848, 2017. 3

[40] Zhirong Wu, Yuanjun Xiong, Stella Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance-level discrimination. *arXiv preprint arXiv:1805.01978*, 2018. 1, 2, 3, 4, 5, 6, 8

[41] Xun Yang, Xiangnan He, Xiang Wang, Yunshan Ma, Fuli Feng, Meng Wang, and Tat-Seng Chua. Interpretable fashion matching with rich attributes. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 775–784, 2019. 3

[42] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016. 5, 6

[43] Chengxu Zhuang, Alex Lin Zhai, and Daniel Yamins. Local aggregation for unsupervised learning of visual embeddings. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6002–6012, 2019. 5, 6