Deep Metric Learning Beyond Binary Supervision

Sungyeon Kim[†] Minkyo Seo[†] Ivan Laptev[‡] Minsu Cho[†] Suha Kwak[†]
POSTECH, Pohang, Korea[†] Inria / École Normale Supérieure, Paris, France[‡]
{tjddus9597, mkseo, mscho, suha.kwak}@postech.ac.kr, ivan.laptev@inria.fr

Abstract

Metric Learning for visual similarity has mostly adopted binary supervision indicating whether a pair of images are of the same class or not. Such a binary indicator covers only a limited subset of image relations, and is not sufficient to represent semantic similarity between images described by continuous and/or structured labels such as object poses, image captions, and scene graphs. Motivated by this, we present a novel method for deep metric learning using continuous labels. First, we propose a new triplet loss that allows distance ratios in the label space to be preserved in the learned metric space. The proposed loss thus enables our model to learn the degree of similarity rather than just the order. Furthermore, we design a triplet mining strategy adapted to metric learning with continuous labels. We address three different image retrieval tasks with continuous labels in terms of human poses, room layouts and image captions, and demonstrate the superior performance of our approach compared to previous methods.

1. Introduction

The sense of similarity has been known as the most basic component of human reasoning [36]. Likewise, understanding similarity between images has played essential roles in many areas of computer vision including image retrieval [19, 43, 44, 50], face identification [12, 39, 46], place recognition [4], pose estimation [45], person re-identification [10, 40], video object tracking [42, 47], local feature descriptor learning [25, 58], zero-shot learning [7, 57], and self-supervised representation learning [52]. Also, the perception of similarity has been achieved by learning similarity metrics from labeled images, which is called *metric learning*.

Recent approaches in metric learning have improved performance dramatically by adopting deep Convolutional Neural Networks (CNNs) as their embedding functions. Specifically, such methods train CNNs to project images onto a manifold where two examples are close to each other if they are semantically similar and far apart otherwise. While in principle such a metric can be learned using any type of semantic similarity labels, previous approaches typically rely on binary labels over image pairs indicating whether the image pairs are similar or not. In this aspect, only a small subset of real-world image relations has been addressed by previous approaches. Indeed, binary similarity labels are not sufficient to represent sophisticated relations between images with structured and continuous labels, such as image captions [30, 35, 56], human poses [3, 21], camera poses [5, 13], and scene graphs [24, 31]. Met-

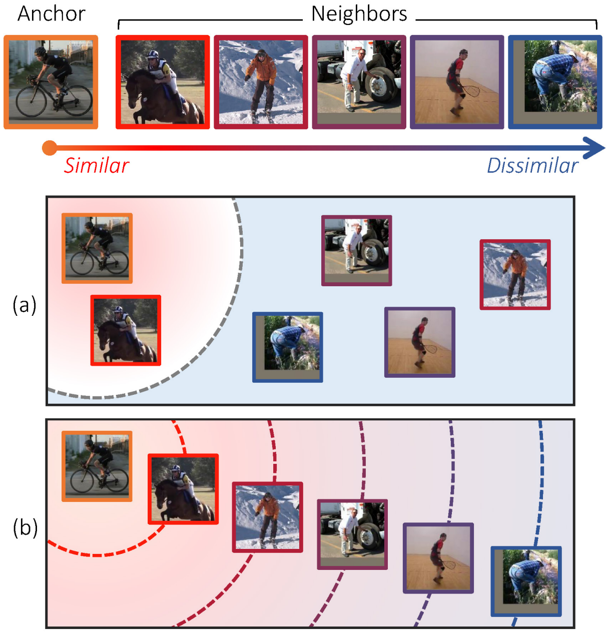


Figure 1. A conceptual illustration for comparing existing methods [4, 16, 27, 32, 45] and ours. Each image is labeled by human pose, and colored in red if its pose similarity to the anchor is high. (a) Existing methods categorize neighbors into positive and negative classes, and learn a metric space where positive images are close to the anchor and negative ones far apart. In such a space, the distance between a pair of images is not necessarily related to their semantic similarity since the order and degrees of similarities between them are disregarded. (b) Our approach allows distance ratios in the label space to be preserved in the learned metric space so as to overcome the aforementioned limitation.

ric learning with continuous labels has been addressed in [4, 16, 27, 32, 45]. Such methods, however, reduce the problem by quantizing continuous similarity into binary labels (*i.e.*, *similar* or *dissimilar*) and applying the existing metric learning techniques. Therefore, they do not fully exploit rich similarity information in images with continuous labels as illustrated in Figure 1(a) and require a careful tuning of parameters for the quantization.

In this paper, we propose a novel method for deep metric learning to overcome the aforementioned limitations. We first design a new triplet loss function that takes full advantage of continuous labels in metric learning. Unlike existing triplet losses [39, 53, 54] that are interested only in the equality of class labels or the order of label distances, our loss aims to preserve ratios of label distances in the learned embedding space. This allows our model to consider degrees of similarities as well as their order and to capture richer similarity information between images as illustrated in Figure 1(b).

Current methods construct triplets by sampling a positive (*similar*) and a negative (*dissimilar*) examples to obtain the binary supervision. Here we propose a new strategy for triplet sampling. Given a minibatch composed of an anchor and its neighbors, our method samples every triplet including the anchor by choosing every pair of neighbors in the minibatch. Unlike the conventional approaches, our method does not need to introduce quantization parameters to categorize neighbors into the two classes and can utilize more triplets given the same minibatch.

Our approach can be applied to various problems with continuous and structured labels. We demonstrate the efficacy of the proposed method on three different image retrieval tasks using human poses, room layouts, and image captions, respectively, as continuous and structured labels. In all the tasks, our method outperforms the state of the art, and our new loss and the triplet mining strategy both contribute to the performance boost. Moreover, we find that our approach learns a better metric space even with a significantly lower embedding dimensionality compared to previous ones. Finally, we show that a CNN trained by our method with caption similarity can serve as an effective visual feature for image captioning, and it outperforms an ImageNet pre-trained counterpart in the task.

2. Related Work

In this section, we first review loss functions and tuple mining techniques for deep metric learning, then discuss previous work on metric learning with continuous labels.

2.1. Loss Functions for Deep Metric Learning

Contrastive loss [6, 12, 17] and triplet loss [39, 50, 54] are standard loss functions for deep metric learning. Given an image pair, the contrastive loss minimizes their distance in the embedding space if their classes are the same, and separates them a fixed margin away otherwise. The triplet loss takes triplets of anchor, positive, and negative images, and enforces the distance between the anchor and the positive to be smaller than that between the anchor and the negative. One of their extensions is quadruple loss [10, 42], which considers relations between a quadruple of images and is formulated as a combination of two triplet losses. A natural way to generalize the above losses is to use a higher order relations. For example, n-tuplet loss [41] takes as its input an anchor, a positive, and $n-2$ negative images, and jointly optimizes their embedding vectors. Similarly, lifted structured loss [44] considers all positive and negative pairs in a minibatch at once by incorporating hard-negative mining functionality within itself. For the same purpose, in [48] the area of intersection between similarity distributions of positive and negative pairs are minimized, and in [28, 43] clustering objectives are adopted for metric learning.

All the aforementioned losses utilize image-level class labels or their equivalent as supervision. Thus, unlike ours, it is not straightforward for them to take relations between continuous and/or structured labels of images into account.

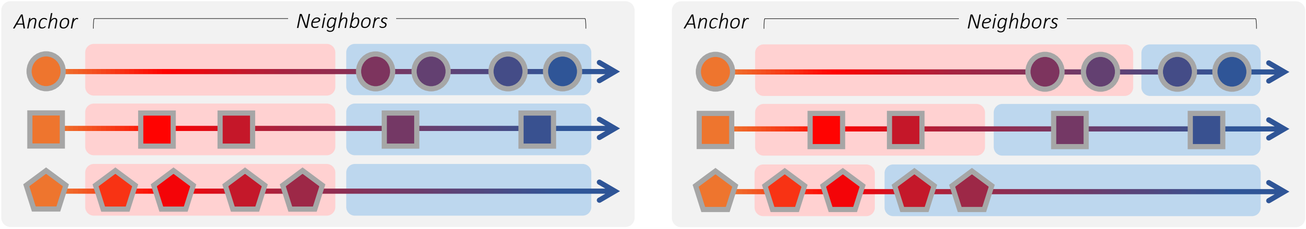
2.2. Techniques for Mining Training Tuples

Since tuples of k images are used in training, the number of possible tuples increases exponentially with k. The motivation of mining techniques is that some of such a large number of tuples do not contribute to training or can even result in decreased performance. A representative example is semi-hard triplet mining [39], which utilizes only semi-hard triplets for training since easy triplets do not update the network and hardest ones may have been corrupted due to labeling errors. It also matters how to measure the hardness. A common strategy [39, 44] is to utilize pairwise Euclidean distances in embedding space, *e.g.*, negative pairs with small Euclidean distances are considered hard. In [19, 20, 55], an underlying manifold of embedding vectors, which is ignored in Euclidean distances, is taken into account to improve the effectiveness of mining techniques. Also, in [57] multiple levels of hardness are captured by a set of embedding models with different complexities.

Although the above techniques substantially improve the quality of learned embedding space, they are commonly based on binary relations between image pairs, thus they are not directly applicable for metric learning with continuous labels.

2.3. Metric Learning Using Continuous Labels

There have been several metric learning methods using data with continuous labels. For example, similarities between human pose annotations have been used to learn an image embedding CNN [27, 32, 45]. This pose-aware CNN then extracts pose information of given image efficiently



(a) Thresholding

(b) Nearest neighbor search

Figure 2. The binary quantization strategies and their limitations. The orange circle indicates a rare example dissimilar to most of the others, and the orange pentagon is a common example similar with a large number of samples. (a) If the quantization is done by a single distance threshold, populations of positive and negative examples would be significantly imbalanced. (b) In the case of nearest neighbor search, positive neighbors of a rare example would be dissimilar and negative neighbors of a common example would be too similar.

without explicit pose estimation, which can be transferred to other tasks relying on pose understanding like action recognition. Also, in [16] caption similarities between image pairs are used as labels for metric learning, and the learned embedding space enables image retrieval based on more comprehensive understanding of image content. Other examples of continuous labels that have been utilized for metric learning include GPS data for place recognition [4] and camera frusta for camera relocalization [5].

However, it is hard for the above methods to take full advantage of continuous labels because they all use conventional metric learning losses based on binary relations. Due to their loss functions, they quantize continuous similarities into binary levels through distance thresholding [4, 32, 45] or nearest neighbor search [16, 27]. Unfortunately, both strategies are unnatural for continuous metric learning and have clear limitations as illustrated in Figure 2. Furthermore, it is not straightforward to find a proper value for their quantization parameters since there is no clear boundary between positive and negative examples whose distances to the anchors are continuous. To the best of our knowledge, our work is the first attempt to *directly* use continuous labels for metric learning.

3. Our Framework

To address limitations of existing methods described above, we propose a new triplet loss called *log-ratio loss*. Our loss directly utilizes continuous similarities without quantization. Moreover, it considers degrees of similarities as well as their rank so that the resulting model can infer sophisticated similarity relations between continuous labels. In addition, we present a new, simple yet effective triplet mining strategy supporting our log-ratio loss since the existing mining techniques in Section 2.2 cannot be used together with our loss.

In the following sections, we briefly review the conventional triplet loss [39] for a clear comparison, then present details of our log-ratio loss and the new triplet mining technique.

3.1. Review of Conventional Triplet Loss

The triplet loss takes a triplet of an anchor, a positive, and a negative image as input. It is designed to penalize triplets violating the rank constraint, namely, that the distance between the anchor and the positive must be smaller than that between the anchor and the negative in the embedding space. The loss is formulated as

$$\ell_{\text{tri}}(a,p,n) = \Big[D(f_a,f_p) - D(f_a,f_n) + \delta\Big]_+, \tag{1}$$

where f indicates an embedding vector, $D(\cdot)$ means the squared Euclidean distance, δ is a margin, and $[\cdot]_+$ denotes the hinge function. Note that the embedding vectors should be L_2 normalized since, without such a normalization, their magnitudes tend to diverge and the margin becomes trivial. For training, gradients of ℓ_{tri} with respect to the embedding vectors are computed by

$$\frac{\partial \ell_{\text{tri}}(a,p,n)}{\partial f_p} = 2(f_p - f_a) \cdot \mathbb{1}\big(\ell_{\text{tri}}(a,p,n) > 0\big), \tag{2}$$

$$\frac{\partial \ell_{\text{tri}}(a,p,n)}{\partial f_n} = 2(f_a - f_n) \cdot \mathbb{1}\big(\ell_{\text{tri}}(a,p,n) > 0\big), \tag{3}$$

$$\frac{\partial \ell_{\text{tri}}(a,p,n)}{\partial f_a} = -\frac{\partial \ell_{\text{tri}}(a,p,n)}{\partial f_p} - \frac{\partial \ell_{\text{tri}}(a,p,n)}{\partial f_n}, \tag{4}$$

where $\mathbb{1}$ is the indicator function. One may notice that the gradients only consider the directions between the embedding vectors and the rank constraint violation indicator. If the rank constraint is satisfied, all the gradients are zero.

3.2. Log-ratio Loss

Given a triplet with samples, we propose a log-ratio loss that aims to approximate the ratio of label distances by the ratio of distances in the learned embedding space. Specifically, we define the loss function as

$$\ell_{\text{lr}}(a,i,j) = \left\{ \log \frac{D(f_a,f_i)}{D(f_a,f_j)} - \log \frac{D(y_a,y_i)}{D(y_a,y_j)} \right\}^2, \tag{5}$$

where f indicates an embedding vector, y is a continuous label, and $D(\cdot)$ denotes the squared Euclidean distance. Also,

(a, i, j) is a triplet of an anchor a and its two neighbors i and j without positive-negative separation, unlike p and n in Eq. (1). By approximating ratios between label distances instead of the distances themselves, the proposed loss enables to learn a metric space more flexibly regardless of the scale of the labels.

The main advantage of the log-ratio loss is that it allows a learned metric space to reflect degrees of label similarities as well as the rank of them. Ideally, the distance between two images in the learned metric space will be proportional to their distance in the label space. Hence, an embedding network trained with our loss can represent continuous similarities between images more thoroughly than those focusing only on the rank of similarities like the triplet loss. This property of the log-ratio loss can be also explained through its gradients, which are given by

$$\frac{\partial \ell_{\text{lr}}(a,i,j)}{\partial f_i} = \frac{(f_i - f_a)}{D(f_a, f_i)} \cdot \ell'_{\text{lr}}(a,i,j), \tag{6}$$

$$\frac{\partial \ell_{\text{lr}}(a,i,j)}{\partial f_j} = \frac{(f_a - f_j)}{D(f_a, f_j)} \cdot \ell'_{\text{lr}}(a,i,j), \tag{7}$$

$$\frac{\partial \ell_{\text{lr}}(a,i,j)}{\partial f_a} = -\frac{\partial \ell_{\text{lr}}(a,i,j)}{\partial f_i} - \frac{\partial \ell_{\text{lr}}(a,i,j)}{\partial f_j}, \tag{8}$$

where $\ell'_{\text{lr}}(a,i,j)$ is a scalar value computed by

$$\ell'_{\text{lr}}(a,i,j) = 4\left\{\log\frac{D(f_a, f_i)}{D(f_a, f_j)} - \log\frac{D(y_a, y_i)}{D(y_a, y_j)}\right\}. \tag{9}$$

As shown in Eq. (6) and (7), the gradients of the log-ratio loss are determined not only by the directions between the embedding vectors but also by $\ell'_{\text{lr}}(a,i,j)$ that quantifies the discrepancy between the distance ratio in the label space and that in the embedding space. Thus, even when the rank constraint is satisfied, the magnitudes of the gradients could be significant if $\ell'_{\text{lr}}(a,i,j)$ is large. In contrast, the gradients of the triplet loss in Eq. (2) and (3) become zero under the same condition.

Another advantage of the log-ratio loss is that it is parameter-free. Unlike ours, the triplet loss requires the margin, which is a hyper-parameter tuned manually and forces embedding vectors to be L_2 normalized. Last but not least, we empirically find that the log-ratio loss can outperform the triplet loss even with embeddings of a significantly lower dimensionality, which enables a more efficient and effective image retrieval.

3.3. Dense Triplet Mining

The existing triplet mining methods in Section 2.2 cannot be used in our framework since they are specialized to handle images annotated by discrete and categorical labels. Hence, we design our own triplet mining method that is well matched with the log-ratio loss.

First of all, we construct a minibatch B of training samples with an anchor, k nearest neighbors of the anchor in terms of label distance, and other neighbors randomly sampled from the remaining ones. Note that including nearest neighbors helps speed up training. Since the label distance between an anchor and its nearest neighbor is relatively small, triplets with a nearest neighbor sample in general induce large log-ratios of label distances in Eq. (9), which may increase the magnitudes of the associated gradients consequently.

Given a minibatch, we aim to exploit all triplets sharing the anchor so that our embedding network can observe the greatest variety of triplets during training. To this end, we sample triplets by choosing every pair of neighbors (i, j) in the minibatch and combining them with the anchor a. Furthermore, since (a, i, j) and (a, j, i) have no difference in our loss, we choose only (a, i, j) and disregard (a, j, i) when $D(y_a, y_i) < D(y_a, y_j)$ to avoid duplication. We call the above procedure *dense triplet mining*. The set of triplets densely sampled from the minibatch B is then given by

$$\begin{aligned}\mathcal{T}(B) = \big\{(a,i,j) \mid D(y_a, y_i) < D(y_a, y_j),& \\ i \in B \setminus \{a\},\ j \in B \setminus \{a\}\big\}.&\end{aligned} \tag{10}$$

Note that our dense triplet mining strategy can be combined also with the triplet loss, which is re-formulated as

$$\begin{aligned}\ell^{\text{dense}}_{\text{tri}}(a,i,j) &= \Big[D(f_a, f_i) - D(f_a, f_j) + \delta\Big]_+ \\ \text{subject to } & (a,i,j) \in \mathcal{T}(B).\end{aligned} \tag{11}$$

where the margin δ is set small compared to that of ℓ_{tri} in Eq. (1) since the label distance between i and j could be quite small when they are densely sampled. This dense triplet loss is a strong baseline of our log-ratio loss. However, it still requires L_2 normalization of embedding vectors and ignores degrees of similarities as the conventional triplet loss does. Hence, it can be regarded as an intermediary between the existing approaches in Section 2.3 and our whole framework, and will be empirically analyzed for ablation study in the next section.

4. Experiments

The effectiveness of the proposed framework is validated on three different image retrieval tasks based on continuous similarities: human pose retrieval on the MPII human pose dataset [3], room layout retrieval on the LSUN dataset [59], and caption-aware image retrieval on the MS-COCO dataset [30]. We also demonstrate that an image embedding CNN trained with caption similarities through our framework can be transferred to image captioning as an effective visual representation.

In the rest of this section, we first define evaluation metric and describe implementation details, then present qualitative and quantitative analysis of our approach on the retrieval and representation learning tasks.

4.1. Evaluation: Measures and Baselines

Evaluation metrics. Since image labels are continuous and/or structured in our retrieval tasks, it is not appropriate to evaluate performance based on standard metrics like Recall@k. Instead, following the protocol in [27], we adopt two evaluation metrics, mean label distance and a modified version of nDCG [8, 27]. The mean label distance is the average of distances between queries and retrieved images in the label space, and a smaller means a better retrieval quality. The modified nDCG considers the rank of retrieved images as well as their relevance scores, and is defined as

$$\text{nDCG}_K(q) = \frac{1}{Z_K} \sum_{i=1}^{K} \frac{2^{r_i}}{\log_2{(i+1)}}, \tag{12}$$

where K is the number of top retrievals of our interest and Z_K is a normalization factor to guarantee that the maximum value of nDCG$_K$ is 1. Also, $r_i = -\log_2{(\|y_q - y_i\|_2 + 1)}$ denotes the relevance between query q and the i^{th} retrieval, which is discounted by $\log_2{(i+1)}$ to place a greater emphasis on one returned at a higher rank. A higher nDCG means a better retrieval quality.

Common baselines. In the three retrieval tasks, our method is compared with its variants for ablation study. These approaches are denoted by combinations of loss function L and triplet mining strategy M, where Log-ratio is our log-ratio loss, Triplet means the triplet loss, Dense denotes the dense triplet mining, and Binary indicates the triplet mining based on binary quantization. Specifically, M(Binary) is implemented by nearest neighbor search, where 30 neighbors closest to anchor are regarded as positive. Our model is then represented as L(Log-ratio)+M(Dense). We also compare our model with the same network trained with the margin based loss and distance weighted sampling [55], a state-of-the-art approach in conventional metric learning. Finally, we present scores of Oracle and ImageNet pretrained ResNet-34 as upper and lower performance bounds. Note that nDCG of Oracle is always 1.

4.2. Implementation Details

Datasets. For the human pose retrieval, we directly adopt the dataset and setting of [27]. Among in total 22,285 full-body pose images, 12,366 images are used for training and 9,919 for testing, while 1,919 images among the test set are used as queries for retrieval. For the room layout retrieval, we adopt the LSUN room layout dataset [59] that contains 4,000 training images and 394 validation images of 11 layout classes. Since we are interested in continuous and fine-grained labels only, we use only 1,996 images of the 5th layout class, which is the class with the largest number of images. Among them 1,808 images are used for training and 188 for testing, in which 30 images are employed as queries. Finally, for the caption-aware image retrieval, the MS-COCO 2014 caption dataset [30] is used. We follow the Karpathy split [22], where 113,287 images are prepared for training and 5,000 images for validation and testing, respectively. The retrieval test is conducted only on the testing set, where 500 images are used as queries.

Preprocessing and data augmentation. For the human pose retrieval, we directly adopt the data augmentation techniques used in [27]. For the room layout retrieval, the images are resized to 224 × 224 for both training and testing, and flipped horizontally at random during training. For the caption-aware retrieval, images are jittered in both scale and location, cropped to 224 × 224, and flipped horizontally at random during training. Meanwhile, test images are simply resized to 256 × 256 and cropped at center to 224 × 224.

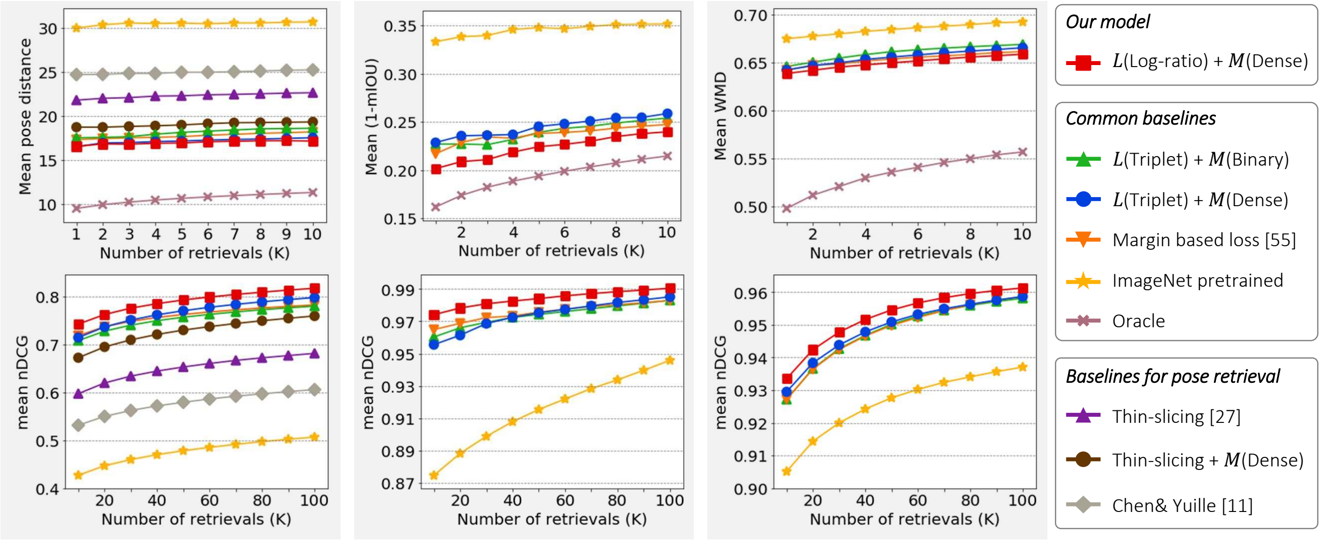
Embedding networks and their training. For the human pose and room layout retrieval, we choose ResNet-34 [18] as our backbone network and append a 128-D FC layer on top for embedding. They are optimized by the SGD with learning rate 10^{-2} and exponential decay for 15 epochs. For the caption-aware image retrieval, ResNet-101 [18] with a 1,024 dimensional embedding layer is adopted since captions usually contain more comprehensive information than human poses and room layouts. This network is optimized by the ADAM [23] with learning rate $5 \cdot 10^{-6}$ for 5 epochs. All the networks are implemented in PyTorch [34] and pretrained on ImageNet [38] before being finetuned.

Hyper-parameters. The size of minibatch is set to 150 for the human pose, 100 for the room layout, and 50 for the caption-aware image retrieval, respectively. On the other hand, k, the number of nearest neighbors in the minibatch for the dense triplet mining, is set to 5 for all experiments. For the common baselines, the margin δ of the conventional triplet loss is set to 0.2 and that of the dense triplet loss 0.03.

4.3. Human Pose Retrieval

The goal of human pose retrieval is to search for images similar with query in terms of human poses they exhibit. Following [27], the distance between two poses is defined as the sum of Euclidean distances between body-joint locations. Our model is compared with the previous pose retrieval model called thin-slicing [27] and a CNN for explicit pose estimation [11] as well as the common baselines.

Quantitative evaluation results of these approaches are summarized in Figure 3(a), where our model clearly outperforms all the others. In addition, through comparisons between ours and its two variants L(Triplet)+M(Dense) and L(Triplet)+M(Binary), it is demonstrated that both of our log-ratio loss and the dense triplet mining contribute to the improvement. Qualitative examples of human pose retrieval are presented in Figure 4. Our model and thin-slicing over-



(a) Human pose retrieval (b) Room layout retrieval (c) Caption-aware image retrieval

Figure 3. Quantitative evaluation of the three retrieval tasks in terms of mean label distance (*top*) and mean nDCG (*bottom*).

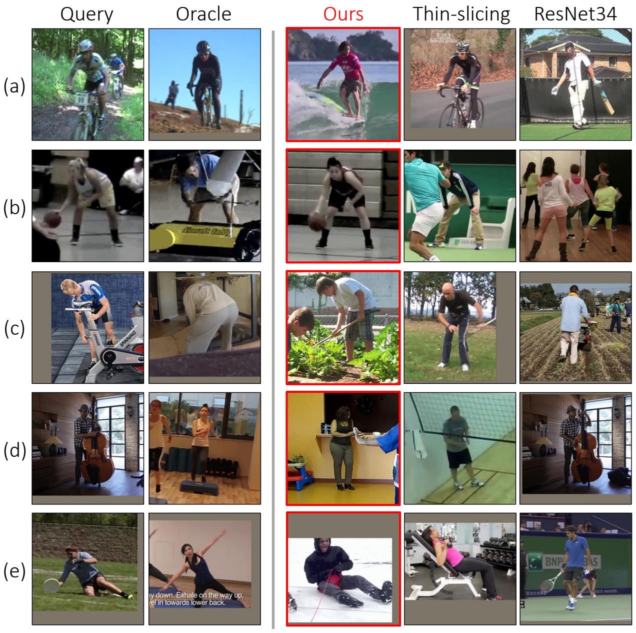


Figure 4. Qualitative results of human pose retrieval.

all successfully retrieve images exhibiting similar human poses with queries, while ResNet-34 focuses mostly on object classes and background components. Moreover, ours tends to capture subtle characteristics of human poses (*e.g.*, bending left-arms in Figure 4(b)) and handle rare queries (*e.g.*, Figure 4(e)) better than thin-slicing.

Finally, we evaluate the human pose retrieval performance by varying embedding dimensionality to show how much effective our embedding space is. As illustrated in Figure 5, when decreasing the embedding dimensionality to 16, the performance of our model drops marginally while that of L(Triplet)+M(Dense) is reduced significantly. Consequently, the 16-D embedding of our model outperforms 128-D embedding of L(Triplet)+M(Dense). This result demonstrates the superior quality of the embedding space learned by our log-ratio loss.

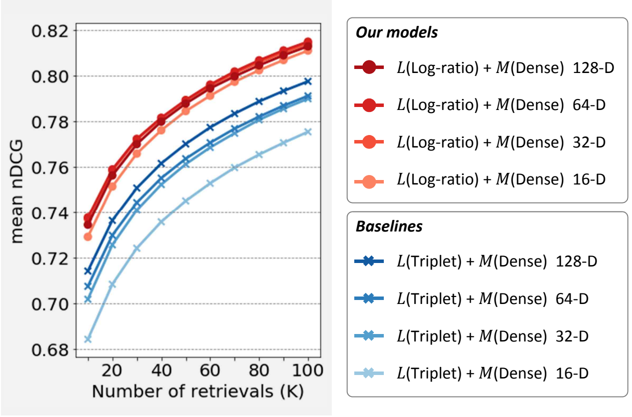


Figure 5. Performance versus embedding dimensionality.

4.4. Room Layout Retrieval

The goal of this task is to retrieve images whose 3-D room layouts are most similar with that of query image, with no explicit layout estimation in test time. We define the distance between two rooms i and j in terms of their room layouts as $1 - \text{mIoU}(R_i, R_j)$, where R denotes the groundtruth room segmentation map and mIoU denotes mean Intersection-over-Union.

Since this paper is the first attempt to tackle the room layout retrieval task, we compare our approach only with the common baselines. As shown quantitatively in Figure 3(b), the advantage of the dense triplet mining is not significant in

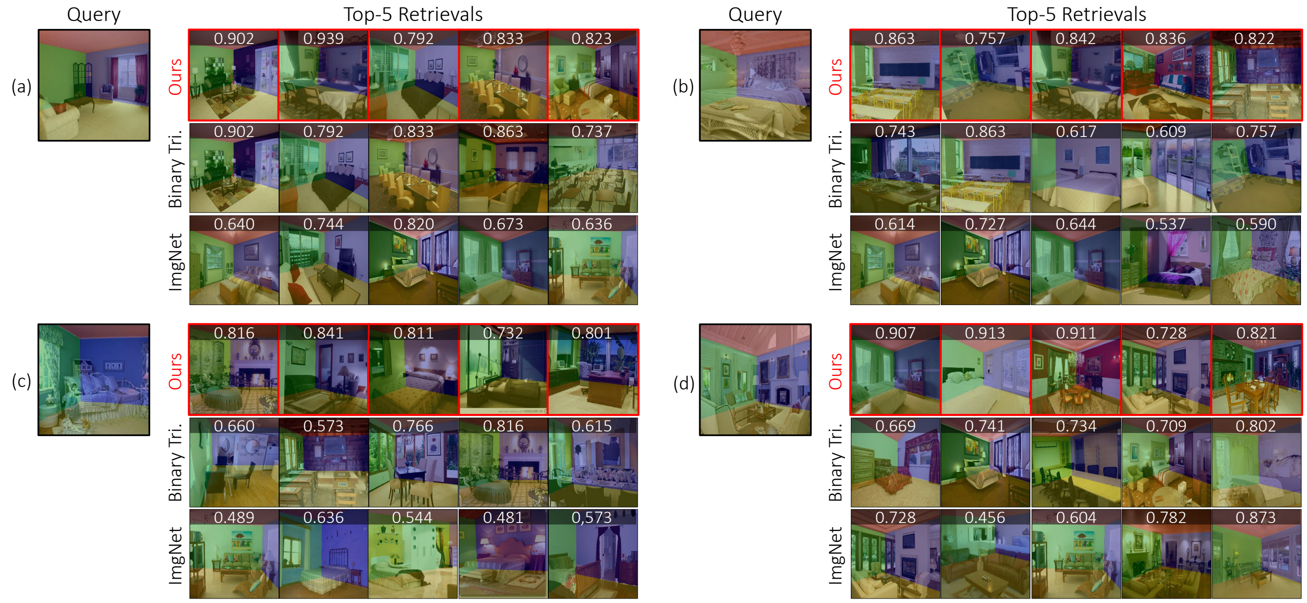


Figure 6. Qualitative results of room layout retrieval. For an easier evaluation, the retrieved images are blended with their groundtruth masks, and their mIoU scores are reported together. Binary Tri.: L(Triplet)+M(Binary). ImgNet: ImageNet pretraiend ResNet101.

this task, probably because room layout labels of the training images are diverse and sparse so that it is not straightforward to sample triplets densely. Nevertheless, our model outperforms all the baselines by a noticeable margin thanks to the effectiveness of our log-ratio loss.

Qualitative results of the room layout retrieval are illustrated in Figure 6. As in the case of the pose retrieval, results of the ImageNet pretrained model are frequently affected by object classes irrelevant to room layouts (*e.g.*, bed in Figure 6(b) and sofa in Figure 6(d)), while those of our approach are accurate and robust against such distractors.

4.5. Caption-aware Image Retrieval

An image caption describes image content thoroughly. It is not a simple combination of object classes, but involves richer information including their numbers, actions, interactions, relative locations. Thus, using caption similarities as supervision allows our model to learn image relations based on comprehensive image understanding.

Motivated by this, we address the caption-aware image retrieval task, which aims to retrieve images described by most similar captions with query. To define a caption-aware image distance, we adopt a sentence distance metric called Word Mover's Distance (WMD) [26]. Let $W(x, y)$ be the WMD between two captions x and y. As each image in our target dataset [30] has 5 captions, we compute the distance between two caption sets X and Y through WMD by

$$W(X,Y) = \sum_{x \in X} \min_{y \in Y} W(x,y) + \sum_{y \in Y} \min_{x \in X} W(x,y). \quad (13)$$

We train our model and the common baselines with the WMD labels. As shown in Figure 3(c), our model outperforms all the baselines, and both of the log-ratio loss and the dense triplet mining clearly contribute to the performance boost, while the improvement is moderate due to the difficulty of the task itself. As illustrated in Figure 7, our model successfully retrieves images that contain high-level image content described by queries like object-object interactions (*e.g.*, person-umbrella in Figure 7(a)), object actions (*e.g.*,holding something in Figure 7(b,d)), and specific objects of interest (*e.g.*, hydrant in Figure 7(c)). In contrast, the two baselines in Figure 7 often fail to retrieve relevant images, especially those for actions and interactions.

4.6. Representation Learning for Image Captioning

An ImageNet pretrained CNN has been widely adopted as an initial or fixed visual feature extractor in many image captioning models [9, 14, 37, 51]. As shown in Figure 7, however, similarities between image pairs in the ImageNet feature space do not guarantee their caption similarities. One way to further improve image captioning quality would be exploiting caption labels for learning a visual representation specialized to image captioning.

We are motivated by the above observation, and believe that a CNN learned with caption similarities through our continuous metric learning framework can be a way to implement the idea. To this end, we adopt our caption-aware retrieval model described in Section 4.5 as an initial, caption-aware visual feature extractor of two image captioning networks: Att2all2 [37] and Topdown [2]. Specifically, our caption-aware feature extractor is compared with the ImageNet pretrained baseline of ours, and (14 × 14 ×

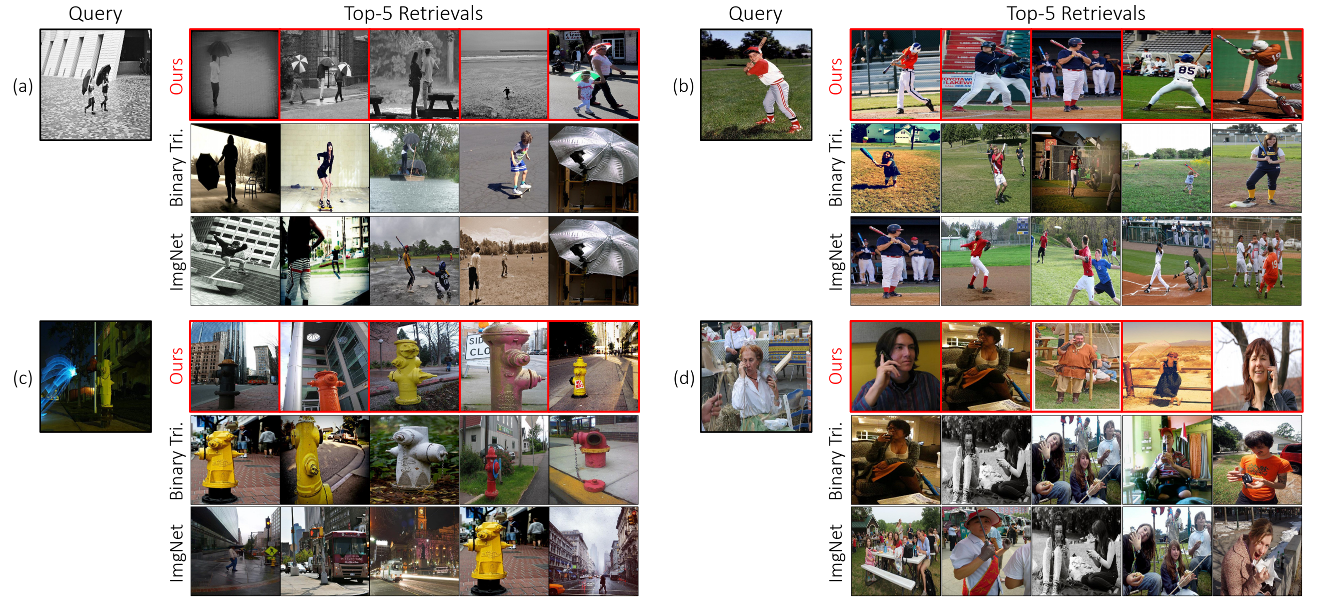


Figure 7. Qualitative results of caption-aware image retrieval. Binary Tri.: L(Triplet)+M(Binary). ImgNet: ImageNet pretraiend ResNet101.

Model	Train		B4	C	M	R	S
ATT	Img	XE	0.3302	1.029	0.2585	0.5456	0.192
		RL	0.3348	1.131	0.2630	0.5565	0.1965
	Cap	XE	0.3402	1.052	0.2608	0.5504	0.1942
		RL	**0.3465**	**1.159**	**0.2673**	**0.5613**	**0.2010**
TD	Img	XE	0.3421	1.087	0.2691	0.5543	0.2011
		RL	0.3573	1.201	0.2739	0.5699	0.2085
	Cap	XE	0.3479	1.097	0.2707	0.5573	0.2012
		RL	**0.3623**	**1.213**	**0.2758**	**0.5718**	**0.2107**

Table 1. Captioning performance on the Karpathy test split [22]. We report scores obtained by a single model with the beam search algorithm (beam size = 2). ATT: Att2all2 [37]. TD: Topdown [2]. Img: ImageNet pretrained feature. Cap: Caption-aware feature. XE: Pretrained with cross-entropy. RL: Finetuned by reinforcement learning. B4: BLEU-4 [33]. C: CIDEr-D [49]. M: METEOR [15]. R: ROGUE-L [29]. S: SPICE [1].

2048) average pooled outputs of their last convolution layers are utilized as caption-aware and ImageNet pretrained features. For training the two captioning networks, we directly follow the training scheme proposed in [37], which first pretrains the networks with cross-entropy (XE) loss then finetunes them using reinforcement learning (RL) with the CIDEr-D [49] metric.

Table 1 quantitatively summarizes captioning performance of the ImageNet pretrained feature and our caption-aware feature. The scores of reproduced baseline are similar or higher than those reported in its original paper. Nonetheless, our caption-aware feature consistently outperforms the baseline in all evaluation metrics and for both of two captioning models. Also, qualitative examples of captions generated by the models in Table 1 are presented in Figure 8, where baselines generate incorrect captions while the models based on our caption-aware feature avoid choosing the wrong word and generate better captions.

GT1: a tennis player swinging the rackets towards the ball
GT2: a man swings his acket to hit a tennis ball

Img XE: a tennis player in a red shirt is playing tennis
Cap XE: a tennis player swinging a racket at a ball
Img RL: a man holding a tennis ball on a tennis court
Cap RL: a man hitting a tennis ball with a tennis racket

Figure 8. Captions generated by the Topdown attention [2]. GT: groundtruth caption. Img: ImageNet pretrained feature. Cap: Caption-aware feature. XE: Pretrained with cross-entropy. RL: Finetuned by reinforcement learning.

5. Conclusion

We have presented a novel loss and tuple mining strategy for deep metric learning using continuous labels. Our approach has achieved impressive performance on three different image retrieval tasks with continuous labels using human poses, room layouts and image captions. Moreover, we have shown that our framework can be used to learn visual representation with continuous labels. In the future, we will explore the effect of label distance metrics and a hard tuple mining technique for continuous metric learning to further improve the quality of learned metric space.

Acknowledgement: This work was supported by Basic Science Research Program and R&D program for Advanced Integrated-intelligence for IDentification through the National Research Foundation of Korea funded by the Ministry of Science, ICT (NRF-2018R1C1B6001223, NRF-2018R1A5A1060031, NRF-2018M3E3A1057306, NRF-2017R1E1A1A01077999), and by the Louis Vuitton - ENS Chair on Artificial Intelligence.

References

[1] P. Anderson, B. Fernando, M. Johnson, and S. Gould. Spice: Semantic propositional image caption evaluation. In *Proc. European Conference on Computer Vision (ECCV)*, pages 382–398. Springer, 2016. 8

[2] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 7, 8

[3] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 1, 4

[4] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 2, 3

[5] V. Balntas, S. Li, and V. Prisacariu. Relocnet: Continuous metric learning relocalisation using neural nets. In *Proc. European Conference on Computer Vision (ECCV)*, 2018. 1, 3

[6] J. Bromley, I. Guyon, Y. Lecun, E. Sckinger, and R. Shah. Signature verification using a "siamese" time delay neural network. In *Proc. Neural Information Processing Systems (NIPS)*, 1994. 2

[7] M. Bucher, S. Herbin, and F. Jurie. Improving semantic embedding consistency by metric learning for zero-shot classification. In *Proc. European Conference on Computer Vision (ECCV)*, 2016. 1

[8] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. Learning to rank using gradient descent. In *Proc. International Conference on Machine Learning (ICML)*, 2005. 5

[9] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6298–6306. IEEE, 2017. 7

[10] W. Chen, X. Chen, J. Zhang, and K. Huang. Beyond triplet loss: A deep quadruplet network for person re-identification. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2

[11] X. Chen and A. Yuille. Articulated pose estimation by a graphical model with image dependent pairwise relations. In *Proc. Neural Information Processing Systems (NIPS)*, 2014. 5

[12] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005. 1, 2

[13] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Niessner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[14] B. Dai and D. Lin. Contrastive learning for image captioning. In *Advances in Neural Information Processing Systems*, pages 898–907, 2017. 7

[15] M. Denkowski and A. Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, 2014. 8

[16] A. Gordo and D. Larlus. Beyond instance-level image retrieval: Leveraging captions to learn a global visual representation for semantic retrieval. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2, 3

[17] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006. 2

[18] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 5

[19] C. Huang, C. C. Loy, and X. Tang. Local similarity-aware deep feature embedding. In *Proc. Neural Information Processing Systems (NIPS)*, 2016. 1, 2

[20] A. Iscen, G. Tolias, Y. Avrithis, and O. Chum. Mining on manifolds: Metric learning without labels. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[21] S. Johnson and M. Everingham. Learning effective human pose estimation from inaccurate annotation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011. 1

[22] A. Karpathy. Neuraltalk2. https://github.com/karpathy/neuraltalk2. 5, 8

[23] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proc. International Conference on Learning Representations (ICLR)*, 2015. 5

[24] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision (IJCV)*, 2017. 1

[25] V. Kumar B G, G. Carneiro, and I. Reid. Learning local image descriptors with deep siamese and triplet convolutional networks by minimising global loss functions. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[26] M. J. Kusner, Y. Sun, N. I. Kolkin, and K. Q. Weinberger. From word embeddings to document distances. In *Proc. International Conference on Machine Learning (ICML)*, 2015. 7

[27] S. Kwak, M. Cho, and I. Laptev. Thin-slicing for pose: Learning to understand pose without explicit pose estimation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 2, 3, 5

[28] M. T. Law, R. Urtasun, and R. S. Zemel. Deep spectral clustering learning. In *Proc. International Conference on Machine Learning (ICML)*, 2017. 2

[29] C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, 2004. 8

[30] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. In *Proc. European Conference on Computer Vision (ECCV)*, 2014. 1, 4, 5, 7

[31] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei. Visual rela-

tionship detection with language priors. In *Proc. European Conference on Computer Vision (ECCV)*, 2016. 1

[32] G. Mori, C. Pantofaru, N. Kothari, T. Leung, G. Toderici, A. Toshev, and W. Yang. Pose embeddings: A deep architecture for learning to match human poses. *arXiv preprint arXiv:1507.00302*, 2015. 1, 2, 3

[33] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, 2002. 8

[34] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. In *AutoDiff, NIPS Workshop*, 2017. 5

[35] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2015. 1

[36] W. V. Quine. *Ontological relativity, and other essays*. Columbia University Press, New York, 1969. 1

[37] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel. Self-critical sequence training for image captioning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 7, 8

[38] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, pages 1–42, April 2015. 5

[39] F. Schroff, D. Kalenichenko, and J. Philbin. FaceNet: A unified embedding for face recognition and clustering. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1, 2, 3

[40] H. Shi, Y. Yang, X. Zhu, S. Liao, Z. Lei, W. Zheng, and S. Z. Li. Embedding deep metric for person re-identification: A study against large variations. In *Proc. European Conference on Computer Vision (ECCV)*, 2016. 1

[41] K. Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Proc. Neural Information Processing Systems (NIPS)*, 2016. 2

[42] J. Son, M. Baek, M. Cho, and B. Han. Multi-object tracking with quadruplet convolutional neural networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2

[43] H. O. Song, S. Jegelka, V. Rathod, and K. Murphy. Deep metric learning via facility location. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2

[44] H. O. Song, Y. Xiang, S. Jegelka, and S. Savarese. Deep metric learning via lifted structured feature embedding. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 2

[45] O. Sumer, T. Dencker, and B. Ommer. Self-supervised learning of pose embeddings from spatiotemporal relations in videos. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2017. 1, 2, 3

[46] O. Tadmor, T. Rosenwein, S. Shalev-Shwartz, Y. Wexler, and A. Shashua. Learning a metric embedding for face recognition using the multibatch method. In *Proc. Neural Information Processing Systems (NIPS)*, 2016. 1

[47] R. Tao, E. Gavves, and A. W. M. Smeulders. Siamese instance search for tracking. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[48] E. Ustinova and V. Lempitsky. Learning deep embeddings with histogram loss. In *Proc. Neural Information Processing Systems (NIPS)*, 2016. 2

[49] R. Vedantam, C. Lawrence Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 8

[50] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu. Learning fine-grained image similarity with deep ranking. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 1, 2

[51] L. Wang, A. Schwing, and S. Lazebnik. Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space. In *Advances in Neural Information Processing Systems*, pages 5756–5766, 2017. 7

[52] X. Wang and A. Gupta. Unsupervised learning of visual representations using videos. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2015. 1

[53] K. Q. Weinberger, J. Blitzer, and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. In *Proc. Neural Information Processing Systems (NIPS)*, 2006. 2

[54] P. Wohlhart and V. Lepetit. Learning descriptors for object recognition and 3d pose estimation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2

[55] C.-Y. Wu, R. Manmatha, A. J. Smola, and P. Krahenbuhl. Sampling matters in deep embedding learning. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2017. 2, 5

[56] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2014. 1

[57] Y. Yuan, K. Yang, and C. Zhang. Hard-aware deeply cascaded embedding. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2017. 1, 2

[58] S. Zagoruyko and N. Komodakis. Learning to compare image patches via convolutional neural networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1

[59] Y. Zhang, F. Yu, S. Song, P. Xu, A. Seff, and J. Xiao. Large-scale scene understanding challenge: Room layout estimation. 4, 5