Cross-Image-Attention for Conditional Embeddings in Deep Metric Learning

Dmytro Kotovenko Pingchuan Ma Timo Milbich Björn Ommer
LMU Munich, IWR Heidelberg, MCML

Abstract

Learning compact image embeddings that yield semantic similarities between images and that generalize to unseen test classes, is at the core of deep metric learning (DML). Finding a mapping from a rich, localized image feature map onto a compact embedding vector is challenging: Although similarity emerges between tuples of images, DML approaches marginalize out information in an individual image before considering another image to which similarity is to be computed.

Instead, we propose during training to condition the embedding of an image on the image we want to compare it to. Rather than embedding by a simple pooling as in standard DML, we use cross-attention so that one image can identify relevant features in the other image. Consequently, the attention mechanism establishes a hierarchy of conditional embeddings that gradually incorporates information about the tuple to steer the representation of an individual image. The cross-attention layers bridge the gap between the original unconditional embedding and the final similarity and allow backpropagtion to update encodings more directly than through a lossy pooling layer. At test time we use the resulting improved unconditional embeddings, thus requiring no additional parameters or computational overhead. Experiments on established DML benchmarks show that our cross-attention conditional embedding during training improves the underlying standard DML pipeline significantly so that it outperforms the state-of-the-art.

1. Introduction

Deep metric learning (DML) seeks embeddings that allow a predefined distance metric to not only express semantic similarities between training samples, but to also transfers to unseen classes. The ability to learn compact image representations that generalize well and transfer in zero-shot manner to unseen test data distributions is crucial for a wide range of visual perception tasks such as visual retrieval [51, 63], image classification [44, 80, 88], clustering [7, 28], or person (re-)identification [11, 27, 69].

DML research has investigated important questions like the effective mining of training samples [4, 54, 57, 65, 68], the training loss function [51, 53, 68, 77, 81, 82], and ensemble strategies [18, 22, 55, 63, 86]. However, learning powerful embeddings is by definition a challenging problem: we seek a mapping from a rich local feature encoding that projects this tensor with all its comprehensive spatial information and local details onto a compact vector that acts as a holistic embedding for an entire image. Local details have to be aggregated and all the important spatial interrelations in an image, e.g., the spatial composition of a scene or the relative configuration of different body parts to another, have to be summed up in a mere vector. However, image similarity is multi-modal in nature—two images can be similar with respect to one characteristic but different in light of another. The challenge is consequently to learn which local details to marginalize out and which to preserve when the embedding function only sees one image and not also the one we want to compute its similarity to. However, during training we have access to all images and powerful loss functions such as multi-similarity loss [82] already compare all image tuples. Thus, we could significantly simplify learning the embedding by conditioning it on another image that we then compute the similarity to.

Contributions: During training we therefore compute similarities using *conditional* embeddings of the image we want to represent conditioned on another image we want to compare against. Thus, the second image focuses the attention of the embedding function on characteristics that are meaningful for a subsequent comparison. Rather then applying a mere pooling operation, we utilize cross-attention to project standard image feature encodings (such as a ResNet convolutional feature map [26]) onto an embedding vector while conditioning it on an embedding of the other image. Repeating these cross-attention blocks then creates a hierarchy of conditional embeddings by successively adding the conditioning information and gradually transitioning from the challenging unconditional embedding to the more accessible conditional one. The hierarchy therefore divides the difficult problem of learning an embedding into several smaller steps. Moreover, due to cross-attention error backpropagation from the similarity measure can now directly update the image encoding and the embeddings

rather than having to optimize the encoding only through the pooling operation of the embedding, which risks attenuated gradients. Consequently the encodings and unconditional embeddings improve over their counterparts in classical DML training. During inference, we therefore employ these unconditional representations so that the approach after training works just like standard DML with no additional parameters and no extra computational costs, simply encoding individual images using a ResNet feature encoder followed by a standard embedding network that outputs the usual pooled feature vector. Our experimental evaluation shows that through cross-attention, our conditional as well as the underlying unconditional embeddings significantly improve over the embeddings obtained by DML so far. Moreover, the computational overhead during training is negligible compared to the costs of current DML training.

2. Related Work

Deep Metric Learning: Deep Metric Learning (DML) [49, 52, 62] is one of the leading lines of research on similarity learning and related applications, such as image retrieval and search [30, 59, 68, 84] or face recognition [10,29,43,65], and even influenced the advance of self-supervised, contrastive representation learning [8, 25, 50]. With the goal of optimizing individual image projections into an expressive embedding space such that similarity relations between the images are reflected by a given distance metric, a multitude of different approaches for learning have been proposed. The main problem formulation of DML are surrogate ranking tasks over tuples of images, ranging from simple pairs [23] and triplets [65, 84] to higher-order quadruplets [9] and more generic n-tuples [27, 54, 68, 82]. These ranking tasks sometimes include geometrical constraints [10, 81]. To make learning feasible despite the exponential complexity of tuple combinations, such methods are often combined with tuple sampling strategies following either manually defined [65, 84, 85] or learned heuristics [19, 24, 60]. Often, this issue is also successfully alleviated by class proxies representing entire sets of training images such as NCA formulations [20, 37, 51, 57, 74] or classification-based approaches [10, 87]. Finally, extensions of these basic formulations further improved the out-of-distribution generalization capabilities of the learned embedding spaces, e.g by leveraging multi-task and ensemble learning [38,48,55,56,59,63], generating synthetic training samples [13,21,40,42,91], diverse, complementary feature semantics [47, 48], self-distillation [61] or sample memory banks [83].

The works above follow the predominant paradigm of determining image similarity by comparing mutually independent, holistic image projections in the embedding space. Thereby, the correspondence between images and spatial structures of them are missing. In our work, we break with this paradigm and learn a cross-attention module [32] that explicitly identifies and links holistic embeddings and local features for estimating similarity during training and refining the final embeddings.

Transformers and Attention Mechanisms: The attention mechanism allows neural networks to explicitly focus on dedicated parts of the model input [31], feature representations [78] and even output [32]. Introduced as hard attention, Spatial Transformers [31] proposed a differentiable input sampler. The powerful formulation of soft (self-)attention was pioneered by transformers [78] which revolutionized the field of natural language processing and also has been gaining much more influence in the vision domain [12]. Recently, cross-attention has been shown to be a flexible concept for relating two arbitrary data representations [32, 33], e.g. for effectively scaling Vision Transformers [12] to large input images. Models purely based on transformer layers have shown competitive performance on the tasks of classification and image retrieval ([15, 76]). In particular, [15, 36] proposed to train a Vision Transformer (ViT) with deep metric learning objectives and gained significant improvement over other architectures using conventional backbones as feature extractors. ViT layers are also deployed to perform extra tasks in DML. In [66], a message-passing network (essentially a ViT) was proposed to exchange information between holistic image representations. Despite the lack of spatial information on individual instances due to the holistic view, this process incorporates the global structure in a mini-batch of samples and refines the final embedding vectors. Similar work [14] utilizes second-order attention blocks to jointly enhance features from different layers of the backbone for individual images. Whereas in our work, we propose a framework to incorporate the information across different images and the spatial structure at the same time.

Local Feature Matching: Establishing correspondences between local image features is a long-standing problem in Computer Vision [17, 58, 67]. Generally, it entails the detection of local interest points in the images and a dedicated stage of 'Local feature matching' [3,5,64,71,89] based on a collection of local descriptors. To address the problem of image search on a large scale, VLAD [35] proposed a way of aggregating local descriptors into a limited dimensional vector. Follow-up work NetVLAD [1] designed a trainable generalized VLAD layer to replace pooling operation and to maintain more local information for later vectors comparison. Another common paradigm is to split this process into two stages, first indexing candidates with global descriptors (holistic embedding) and then re-ranking them with local descriptors [6, 73, 75].

Similar to above mentioned works, DIML [90] proposed a structural matching strategy to explicitly align the richer spatial feature maps by solving an expensive optimal trans-

port between pairs of images. However, solving this for each pair of images imposes costly computation overhead. In contrast, we cross attend the global and local features solely during training. As this is trained in an end-to-end fashion with the backbone network, the feature maps and holistic embeddings are all refined together. Hence, the conventional holistic embeddings are sufficient for similarity computation during inference time (see 4.1.1).

Explainability in Deep Learning: Deep Metric Learning methods typically are difficult to interpret due to the holistic nature of the optimized latent embedding spaces. ABE [38] uses a self-attention mechanism for learning an ensemble of global learners to implicitly focus on different parts of the input image. However, *(i)* attention is not performed between images, thus only masked image regions that are captured by a particular learner can be visualized and *(ii)* those image regions are only consistent for very attention channels. In contrast, our approach explicitly establishes local correspondences between images, which are used to determine individual similarities between object parts. These correspondences naturally allow to visualize fine-grained relations between objects that the model considers crucial for similarity assessment. Similarly, DIML [90] aims at finding local object correspondences, which, however, are limited to coarse object parts only, due to computational restrictions limiting the number of independent image regions to be represented. A widely used visualization in DML are UMAP [46] or tSNE [45] projections of the holistic image embeddings. While such visualizations help to show which images are overall similar and dissimilar, they only implicitly provide insights into why a model puts two images next to each other on the embedding manifold.

3. Approach

Typically, DML approaches have two stages. First an encoder function E maps an image $I \in \mathbb{R}^{H\times W\times 3}$ to a tensor $E(I) \in \mathbb{R}^{h\times w\times c}$ with lower spatial dimensionality $h \times w$ but with c channels, thus aggregating different visual patterns in different channels. Thereafter, this encoding is mapped on a d-dimensional embedding $\phi(E(I)) \in \mathbb{R}^d$, such that some similarity measure $s(\phi(\cdot), \phi(\cdot))$ in the embedding space, e.g. scalar product, corresponds to semantic similarities in the image space. Therefore, the embedding function ϕ typically marginalizes out the spatial dimensions (usually by simple average pooling across spatial dimensions) and projects onto a higher dimensional unit sphere.

The challenge is consequently to marginalize out only irrelevant local details and to retain all meaningful characteristics. Computing such a mapping is aggravated by the fact that we have to compute $\phi(E(I_i))$ only based the image I_i without knowing what other image I_j we are comparing to. In contrast, having some information about I_j such as its encoding $E(I_j)$ or its final embedding $\phi(E(I_j))$ would significantly simplify estimating $\phi(E(I_i)|\phi(E(I_j)))$, since the conditioning helps to focus on meaningful characteristics for a subsequent comparison $s(\phi(I_i|I_j)); \phi(I_j|I_i))$. We have dropped the image representations here to shorten notation. Our experimental evaluation in Sec.4 shows that conditioning on $\phi(E(I_j))$ during training significantly improves upon representations learned unconditionally with a negligible computational overhead during training compared to conditioning on $E(I_j)$ and no overhead during inference with the trained model.

Let us now use the embedding of I_j to attend to meaningful sites in $E(I_i) \in \mathbb{R}^{h\times w\times c}$ and compute the relevance of individual local encodings. Please note that for the sake of simplicity of notation we flatten first two dimensions of the representation to have $E(I_i) \in \mathbb{R}^{hw\times c}$. After applying linear layers [32] $Q : \mathbb{R}^d \to \mathbb{R}^d$ and $K, V : \mathbb{R}^c \to \mathbb{R}^d$, we can measure similarities $Q(\phi(E(I_j)))\, K^\top(E(I_i))$ to obtain weights for a subsequent weighting of the local encodings. Since only the relative similarity matters, we additionally utilize a softmax normalization.

$$\begin{aligned}&\text{Attn}(\phi(E(I_j)), E(I_i)) :=\\ &\quad \text{softmax}\left(\frac{Q(\phi(E(I_j)))\, K^\top(E(I_i))}{\sqrt{d}}\right) \in \mathbb{R}^{1\times hw}.\end{aligned} \tag{1}$$

Using Eq. 1 to weight a linear layer output $V(E(I_i))$ of the encodings then yields a cross-attention block [32],

$$CA(q, k, v) := \text{softmax}\left(\frac{Q(q)\, K^\top(k)}{\sqrt{d}}\right) V(v), \tag{2}$$

so that

$$\begin{aligned}&\phi(E(I_i)|\phi(E(I_j)) =\\ &\qquad CA(\phi(E(I_j)), E(I_i), E(I_i)) \in \mathbb{R}^d\end{aligned} \tag{3}$$

Eq. 2 maps $E(I_i)$ to the embedding space by focusing on those local features that also occur in I_j and that are therefore relevant for computing similarities afterwards.

To amplify meaningful local characteristics more, we apply the cross-attention in Eq. 3 repeatedly,

$$\phi^n(I_i|I_j) := \begin{cases}\phi^0(I_i|-) \equiv \phi(E(I_i)), & n = 0\\ CA(\phi^{(n-1)}(I_j|I_i), E(I_i), E(I_i)), & n > 0\end{cases} \tag{4}$$

After some N update steps, final similarities between images I_i and I_j become

$$s^N(I_i, I_j) := \frac{\phi^N(I_i|I_j)\, \left(\phi^N(I_j|I_i)\right)^\top}{\|\phi^N(I_i|I_j)\|\|\phi^N(I_j|I_i)\|}. \tag{5}$$

Both E and ϕ can then be trained in end-to-end fashion by backpropagating the error between the predicted similarities $s^N(I_i, I_j)$ and the ground-truth. We here employ

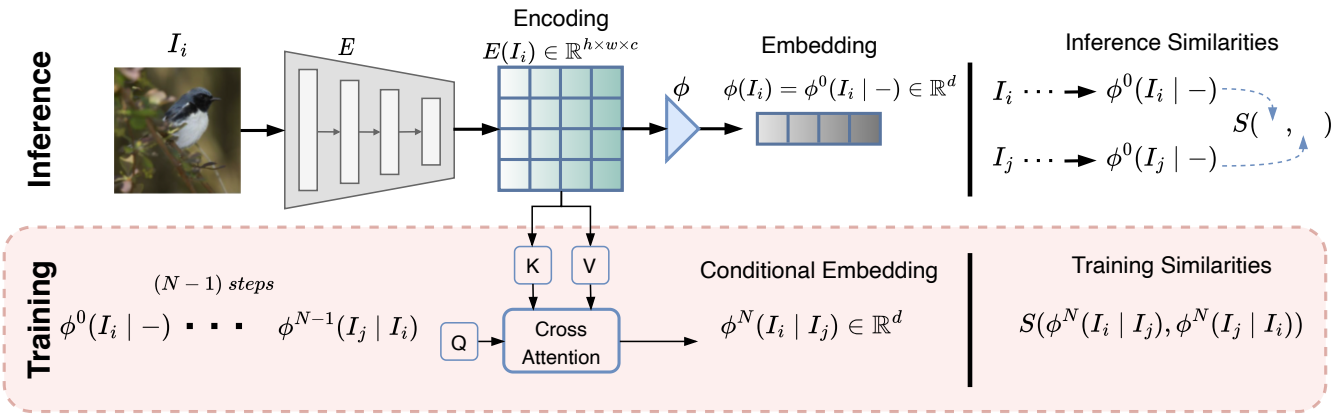


Figure 1. Approach overview. Our method proceeds differently in the inference and training stage. During inference, input images I_i are first processed by the encoder E yielding feature maps $E(I_i)$. Then $E(I_i)$ is mapped into a d-dimensional embedding space using the projection function $\phi(E_i) \equiv \phi^0(E_i|-)$ to represent the image I_i. During training, we want to learn a conditional embedding of I_i that facilitates the subsequent computation of similarity to some other image I_j. Therefore, we seek a mapping from $E(I_i)$ into the embedding space, conditioned on the embedding $\phi(E(I_j))$ of I_j. We utilize a cross-attention block to aggregate features from E_i relevant for I_j: $\phi^1(I_i|I_j) = CA(\phi^0(E(I_j)), E(I_i), E(I_i))$. We repeat this process N times to obtain the final conditional embedding $\phi^N(I_i|I_j)$. This refined embedding is then used to compute the loss and to train the model weights in E, ϕ, and all the cross-attention blocks.

multi-similarity loss [82] for all predictions within a batch:

$$\mathcal{L} := \frac{1}{b}\sum_{i=1}^{b}\left(\frac{1}{\alpha}\log\left[\sum_{k\in\mathcal{P}_i}\exp^{-\alpha(s^N(I_i,I_k)-\lambda)}\right] + \frac{1}{\beta}\log\left[\sum_{k\in\mathcal{N}_i}\exp^{\beta(s^N(I_i,I_k)-\lambda)}\right]\right). \tag{6}$$

Alg.1 summarizes the training procedure of our proposed approach.

Observations on Training: The successive cross-attention blocks of Eq. 4 establish a hierarchy of conditional embeddings that gradually feed information from the conditioning image into the $\phi^n(I_i|I_j)$. After some N update steps, the successive layers of this hierarchy are bridging the gap between the challenging unconditional embedding $\phi^0(I_i|-)$, which is however generally applicable (for similarities of I_i to arbitrary other images), and the simpler conditional embedding ϕ^N, which is specific for the tuple i, j. So rather than having to estimate ϕ^0 directly, we can gradually get there by backpropagating through the hierarchy: end-to-end training of Eq. 4 updates all ϕ^n and in particular also $\phi(E(I_i))$. In essence, backpropagation through the subsequent cross-attention blocks, iteratively distributes information between image encodings and embeddings. Moreover, the encoder receives gradients directly in the cross-attention and not attenuated through the pooling layer of ϕ, as is common in DML. More precisely, each cross-attention block backpropagates the gradient from the loss $\mathcal{L}$ in Eq. 6 through $\phi^{n+1} = CA(\phi^n(I_j|I_i), E(I_i), E(I_i)$ directly to the embedding $E(I_i)$ and to the weights of E. That way we bypass the lossy $\phi^0 \equiv \phi$ function that involves spatial pooling inside and obtain significantly improved $\phi(E(I))$ compared to standard DML training without Eq. 4. Fig. 2 visualizes how conditional embeddings are wired with each other and the encoder E.

Algorithm 1 Training

Require: E - pretrained ResNet-50,
X - dataset with images and class labels,
b - batch size
Initialize E
Initialize weights of initial embedding layer ϕ and weight of the projection heads Q, K, V in the cross-attention blocks
while not converged **do**
  Sample b Images with labels $(I_i, l_i) \in X$
  for $\forall i \in \{1, .., b\}$ **do**
    Compute backbone outputs $E(I_i)$
    Compute $\phi^0(I_i|-) = \phi(E(I_i))$
  end for
  for $\forall n \in \{1, .., N\}$ **do**
    $\forall i, j \in \{1, .., b\}$ compute
    $\phi^n(I_i|I_j) = CA(\phi^{(n-1)}(I_j|I_i), E(I_i), E(I_i))$
  end for
  Compute $s^N(I_i, I_j)$ with Eq.5
  Compute loss $\mathcal{L}$ specified in Eq.6
  Backpropagate gradients of $\mathcal{L}$ into weights $\theta_\phi, \theta_Q, \theta_K, \theta_V$.
end while

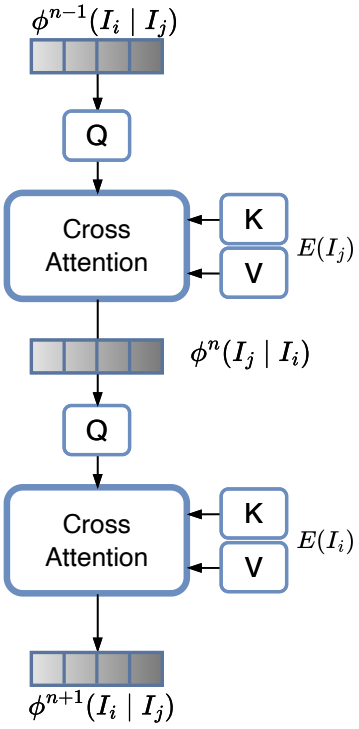


Figure 2. We iteratively refine the conditional embedding of image I_i conditioned on an image I_j. The main computation block enabling refinement of features is the cross-attention between the conditional embedding from the previous step, $\phi^n(I_j|I_i)$, of image I_j conditioned on the I_i and the rich image encoding $E(I_i)$.

Inference with the Trained Model: Embedding all test data using $\phi^N(I_i, I_j)$ to subsequently compute similarities has a quadratic complexity in the size of the test set as opposed to the linear complexity of $\phi(I_i)$. However, the end-to-end training of Eq. 4 has updated all ϕ^n and thereby also $\phi(E(I_i))$, which is now significantly improved compared to standard DML training. Therefore, we simply employ $s^0(I_i, I_j)$, which is just the usual DML inference step leading to $\frac{\phi(E(I_i))\phi(E(I_i))^\top}{\|\phi(E(I_i))\|\|\phi(E(I_i))^\top\|}$, according to Eq. 5 and Eq. 4. Hence, computing similarities for all pairs of images during image retrieval boils down to a conventional nearest neighbour computation as it is done in the usual DML settings. We further discuss and analyze the effects of using $s^n(I_i, I_j)|n > 0$ on the retrieval performance in Sec. 4.4.

4. Experiments

Subsequently, we first discuss the experimental setup, including the implementation details and the used benchmark datasets. Then we compare our model to the current state-of-the-art approaches in DML and ablate certain parts of our model. Finally, we conduct additional experiments to investigate and visually explain attention maps of different cross-attention blocks.

Implementation details We follow the common training protocol [59, 63, 84] for DML and utilize a ResNet50 [26] encoder pretrained on the ImageNet dataset. The model is implemented in the Tensorflow2 framework. All the experiments are conducted on a single RTX 8000 or a single RTX 6000 GPU.

For training, we use the Adam [39] optimizer with a fixed learning rate of 10^{-5} and default β_1, β_2 parameters with no learning rate scheduling being applied. A default batch size of 128 is used unless stated otherwise. We choose the popular multi-similarity loss [82] as our DML objective function using default parameters stated in the original paper. For all the experiments we resize input images to the size 256×256px following standard practice [52, 60]. At training we crop a random patch, resize it to 224×224px and randomly flip image horizontally. At inference time, to further follow standarad protocol, we apply center cropping to size 224×224px after the initial resize to 256×256px and feed it to the network. Our embeddings lie in the $d = 512$-dimensional space, thus the output of ϕ, $\phi^n(\cdot,\cdot)\forall n$ is in 512-dimensional space. We use $N = 6$ cross-attention blocks for all our experiments and for all datasets. Below we discuss how the number of blocks affects the retrieval performance. In every cross-attention block, keys and values are first processed with layer norm [2] and queries are also normalized to have a unit norm.

Datasets. We evaluate the performance on three standard DML benchmark datasets using the default train-test splits:

- *CARS196* [41], which contains 16,185 images from 196 car classes. The first 98 classes containing 8054 images are used for training, while the remaining 98 classes with 8131 images are used for testing.

- *CUB200-2011* [79] with 11,788 bird images from 200 classes. Training/test sets contain the first/last 100 classes with 5864/5924 images respectively.

- *Stanford Online Products (SOP)* [54] provides 120,053 images divided in 22,634 product classes. 11318 classes with 59551 images are used for training, while the remaining 11316 classes with 60502 images are used for testing.

4.1. Ablations

4.1.1 Cross-Attention Blocks

We observe an improvement when using multiple CA attention blocks and using later conditional embeddings ϕ^N. This indicates that having more steps to refine the conditional embedding computation is beneficial for the perfor-

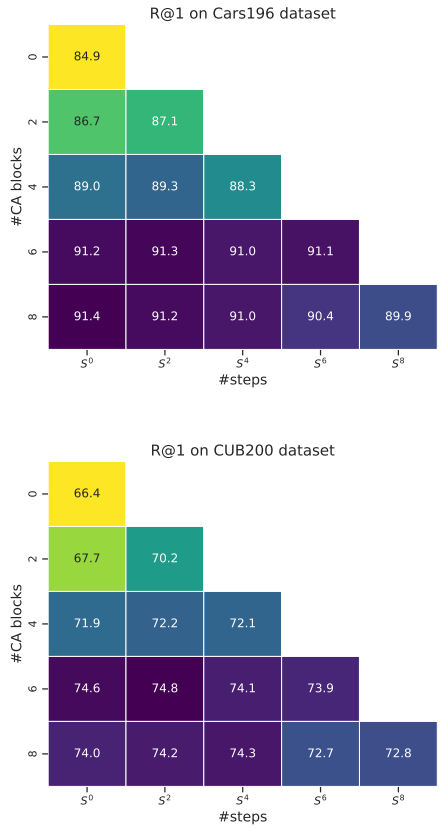


Figure 3. Retrieval performance depends on the number of cross-attention blocks in the model architecture. Moreover for a model with N cross-attention blocks we obtain similarities s^n at different levels $n < N$ and each of those can be used to compute retrieval scores at testing phase. This figure summarizes R@1 scores on the test sets of the Cars196 and CUB200 datasets. We observe improvements in performance when using multiple CA blocks (in rows) but this improvement saturates around $6-8$ blocks. We also see that using higher similarities s^n(in rows) for inference degrades the performance, thus indicating that during the test phase it is reasonable to use s^0. That means that unconditional embeddings $\phi^0(\cdot|-)$ already outperform other state-of-the-art models.

mance. In Fig. 3 we summarize the effect of different numbers of cross-attention blocks on the $R@1$ scores for two different datasets. The optimal number of attention blocks required for computation is $N = 6$. We additionally visualize how $R@1$ scores change when we use different conditional embeddings ϕ^n at different cross-attention blocks. As it has been specified above we only use $s^0(I_i, I_j)$ to compute the similarities in Tab. 1. We do not need to compute all pairs of embeddings $\phi^n(I_i|I_j)$. However, using the later similarities $s^n(I_i, I_j)$ for $n > 0$ we get a slight improvement in performance.

4.1.2 Number of Network Parameters

In order to verify that our improvements in recall performance R@1 come from better matching local regions of one image to the holistic conditional embedding of another image and not from simply adding extra parameters, we run our model with a larger ResNet-101 backbone encoder E and without cross-attention similarity learning, i.e. a multi-similarity loss [82] baseline model. While our model based on the conventional ResNet-50 backbone and with 6 cross-attention similarity blocks only has 27.7M parameters, the baseline ResNet-101 has 45.3M. Nevertheless, our model reaches 74.6% and 91.2% R@1 on CUB200 and Cars196 respectively, compared to the ResNet-101 baseline reaching only scores 67.4% and 83.7% respectively.

4.2. Comparison to State-of-the-Art

In this section, we evaluate our approach on the standard benchmark sets in DML, i.e. CUB200 [79], CARS196 [41] and SOP [54], and compare it to the state-of-the-art methods with the widely used Recall@k score [34], measuring image retrieval performance. Tab. 1 demonstrates that our framework significantly outperforms all approaches, especially when increasing the spatial resolution of the spatial feature map output of the encoder E, and thus the number of image regions that can be individually and independently represented. Our approach improves over holistic embedding-based state-of-the-art performance up to 4.3% on the CUB200 dataset, 1.7% on CARS196 and 0.9% on SOP. This clearly demonstrates the importance of emphasizing local features conditioned on the holistic representation of another image. To ensure the fairness and coherence of the comparison we do not specify methods utilizing bigger backbones, e.g. ViT-16 [12] pretrained on much bigger internal JFT [70] dataset.

4.3. Emerging properties of Cross-Attention

4.3.1 Cross-Attention Maps

Let us now check the attention matrix $\text{Attn}(\phi(E_i), \phi(E_j))$ for different cross-attention blocks and interpret the results. We visualize our results in Fig. 4. The figure shows that deeper layers tend to focus on fewer details compared to the attention maps of the earlier cross-attention blocks. Moreover, we see that attention for the same image changes depending on which embedding we use as a query. We observe from the Fig. 4 that though attention maps look similar they still have different activations as visualized in the middle row.

4.3.2 Local Parts Discovery

Our approach is based on computing the attention between the holistic representation of one image $\phi^n(I_j|I_i)$ and the

		CUB200-2011 [79]			CARS196 [41]			SOP [54]		
Method	BB	R@1	R@2	NMI	R@1	R@2	NMI	R@1	R@10	NMI
Margin^{128} [84]	R50	63.6	74.4	69.0	79.6	86.5	69.1	72.7	86.2	90.7
Multi-Sim^{512} [82]	BNI	65.7	77.0	-	84.1	90.4	-	78.2	90.5	-
MIC^{128} [59]	R50	66.1	76.8	69.7	82.6	89.1	68.4	77.2	89.4	90.0
HORDE^{512} [30]	BNI	66.3	76.7	-	83.9	90.3	-	80.1	91.3	-
Softtriple^{512} [57]	BNI	65.4	76.4	69.3	84.5	90.7	70.1	78.3	90.3	92.0
XBM^{128} [83]	BNI	65.8	75.9	-	82.0	88.7	-	80.6	91.6	-
PADS^{128} [60]	R50	67.3	78.0	69.9	83.5	89.7	68.8	76.5	89.0	89.9
GroupLoss^{1024} [16]	BNI	65.5	77.0	69.0	85.6	91.2	72.7	75.1	87.5	90.8
DIML^{512} [90]	R50	67.9	-	-	87.0	-	-	78.5	-	-
ProxyAnchor^{512} [37]	BNI	68.4	79.2	-	86.1	91.7	-	79.1	90.8	-
D&C^{512} [63]	R50	68.2	-	69.5	87.8	-	70.7	79.8	-	89.7
SynProxy^{512} [21]	R50	69.2	79.5	-	86.9	92.4	-	79.8	90.9	-
DiVA^{512} [47]	R50	69.2	79.3	71.4	87.6	92.9	72.2	79.6	91.2	90.6
S2D2^{512} [61]	R50	70.1	79.7	71.6	89.5	93.9	72.9	80.0	91.4	90.8
Intra-Batch^{512} [66]	R50	70.3	80.3	74.0	88.1	93.3	74.8	81.4	91.3	92.6
MH-DML^{512} [14]	R50	70.6	80.9	-	90.1	94.2	-	81.7	92.0	-
Ours^{512}	R50	**74.6**	**83.7**	**76.9**	**91.2**	**94.4**	**77.3**	**82.3**	**92.2**	**93.1**

Table 1. *Comparison to the state-of-the-art methods on CUB200-2011* [79], *CARS196* [41] *and SOP* [54]. 'BB' denote the backbone architecture being used ('R50'=ResNet50 [26], 'BNI'=BN-InceptionNet [72]).)

rich spatial representation of another image $E(I_i)$. These two entities are combined together when computing the attention matrix in the cross-attention block evaluation as defined in Eq. 2. For an image I we compute the encoding $E(I)$ and sample a location x, y, now the element $E(I)_{x,y}$ encodes visual information in the corresponding part of an image. We compute now the most similar parts from $E(I')$ for all the other images I' in the dataset. In Fig. 5 we visualize these by showing cropped out patches corresponding to the most similar parts. All the retrievals are semantically similar and also share similar appearance. Thus we observe that our model has learned semantic parts given only image level labels. Moreover, we denote with color whether a retrieved image region is from an image with the same label (green) as the query image or not (red).

4.4. Computational Complexity

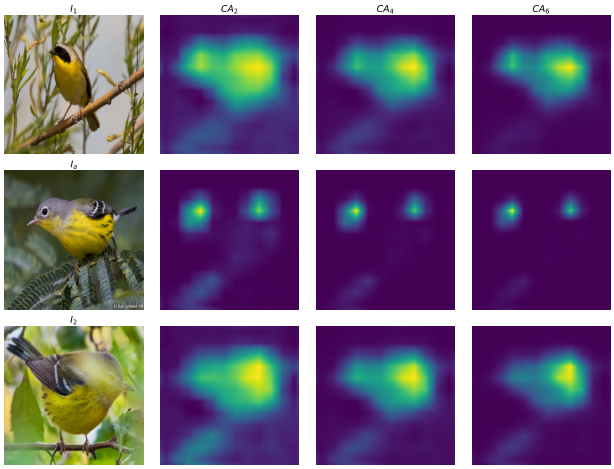
Our approach has two specific implications on computational complexity that we subsequently discuss.

First, to obtain attention scores we need to compute the conditional embeddings $\phi^n(I_i|I_j)$ with queries, keys and values of size $1 \times d$, $t \times d$ and $t \times d$ respectively, where $t = hw$ denotes the number of tokens and d denotes the number of feature dimensions. Hence, computation of a single cross-attention block requires an extra $2 \cdot d \cdot t$ multiplications. In our case $d \gg t$, thus the computation overhead is negligible. This would not be the case if we were to compute the attention between $E(I_i)$ and $E(I_j)$. Our loss is computed given $\phi^N(I_i, I_j) \forall i, j \in 1, .., b$ for a batch of b images. This additionally multiplies the number of calculations by b^2, which results in $2 \cdot b^2 \cdot d \cdot t$ which depends quadratically on the batch size only. We train our models with batches of 128 samples and we observe only 11% increase in computation time per batch compared to the baseline model without any cross-attention blocks.

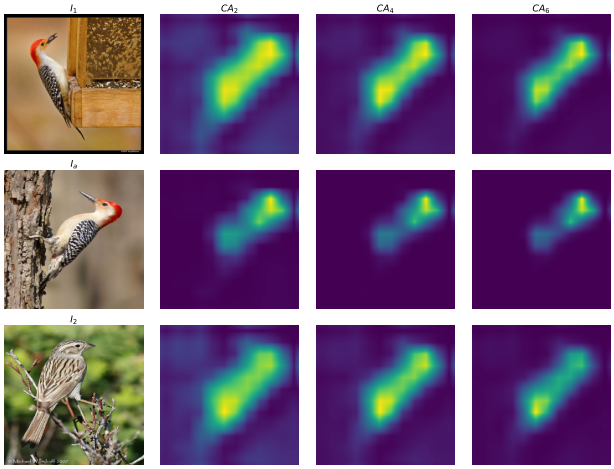
Second, computing conditional embeddings $\phi^N(I_i, I_j)$ would imply high computational cost at inference stage for large scale image retrieval. The DML setup demands estimating similarities for all pairs of images exhaustively. This would require M^2 computations for a dataset of M images. But from Fig. 3 we see that it is sufficient to use ϕ^0 instead of ϕ^N to obtain high quality retrieval. This proves that we need the conditional embeddings first and foremost during training to improve the gradients that update the weights of E and ϕ, which we use during inference.

5. Conclusion

Embedding rich image encodings into compact vectors is a main challenge in DML. If the embedding function had access to a tuple that is to be compared rather than only an individual image, it would be significantly easier to focus on the meaningful features for this particular comparison. Therefore, we have utilized conditional embeddings, where the representation of one image of the tuple is learned conditioned on its partner. Using cross-attention, we have established a hierarchy of conditional embeddings that gradu-



(a) Example 1.



(b) Example 2.

Figure 4. Visualizing the attention maps of different cross-attention blocks for two exemplary image triples. For the triples in (a) and (b) we compute the attention $\text{Attn}(\phi(E(I_a)), E(I_1))$ and $\text{Attn}(\phi(E(I_a)), E(I_2))$ in the top and bottom rows respectively. Different columns stand for different cross-attention blocks, we visualize here only layers 2, 4 and 6. In the middle row we show the difference between the upper and the lower row to amplify locations with different attention. We observe that later cross-attention blocks focus less on the background clutter and more on the distinctive features of a bird. All birds in the first triple share the same breast color, hence the attention is focused more around the head area which is helpful to distinguish I_a from I_1. In the second triple, attention is concentrated around the head area. This is the most prominent feature relating images I_a and I_1.

ally incorporates information about the tuple into the original unconditional embedding. Our experimental evaluation has shown that this hierarchy significantly improves image encodings and embeddings, particularly also the unconditional embeddings of standard DML that we use during inference. Our approach only augments DML during training and with only negligible computational overhead. There, is no change to the DML architecture during inference and no additional parameters or computational cost.

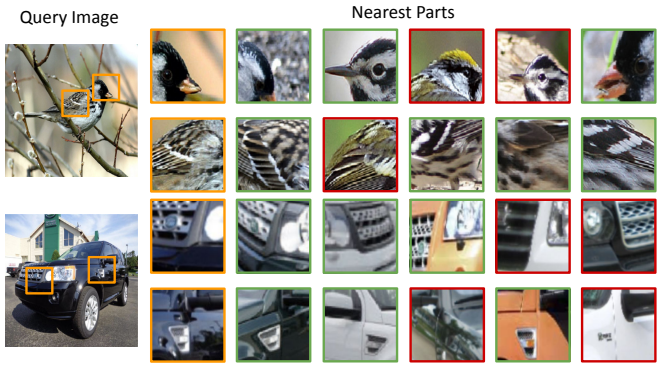


Figure 5. The encoder representation $E(I)$ learns semantic parts specific to the dataset without any extra labels provided. For each of the two images on the left we pick two locations (indicated with orange rectangles). For each of these locations we find the most similar parts across all the other images in the dataset. With a green frame we denote a crop from an image having the same label as the query image and with a red frame a crop from an image with a different label.

Acknowledgement

This work has been funded in part by Bayer AG, the German Federal Ministry for Economic Affairs and Climate Action project CLINIC 5.1 - Comprehensive Lifesciences Neural Information Computing, the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) project 421703927, and the bidt project KLIMA-MEMES.

References

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

[2] Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *ArXiv*, abs/1607.06450, 2016. 5

[3] Dániel Baráth and Jiri Matas. Magsac: Marginalizing sample consensus. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[4] Miguel Ángel Bautista, Artsiom Sanakoyeu, and Björn Ommer. Deep unsupervised similarity learning using partially ordered sets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1923–1932, 2017. 1

[5] Eric Brachmann and Carsten Rother. Neural- Guided RANSAC: Learning where to sample model hypotheses. In *ICCV*, 2019. 2

[6] Bingyi Cao, Andre Araujo, and Jack Sim. Unifying deep local and global features for image search. In *European Conference on Computer Vision*, pages 726–743. Springer, 2020. 2

[7] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. *European Conference on Computer Vision*, 2018. 1

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh, editors, *International Conference on Machine Learning*, volume 119, 2020. 2

[9] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[10] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[11] Shengyong Ding, Liang Lin, Guangrun Wang, and Hongyang Chao. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 48(10):2993–3003, 2015. 1

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 2, 6

[13] Yueqi Duan, Wenzhao Zheng, Xudong Lin, Jiwen Lu, and Jie Zhou. Deep adversarial metric learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[14] Mohammad K Ebrahimpour, Gang Qian, and Allison Beach. Multi-head deep metric learning using global and local representations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3031–3040, 2022. 2, 7

[15] Alaaeldin El-Nouby, Natalia Neverova, Ivan Laptev, and Hervé Jégou. Training vision transformers for image retrieval. *CoRR*, 2021. 2

[16] Ismail Elezi, Sebastiano Vascon, Alessandro Torcinovich, Marcello Pelillo, and Laura Leal-Taixe. The group loss for deep metric learning. In *European Conference on Computer Vision (ECCV)*, 2020. 7

[17] Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, 1981. 2

[18] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997. 1

[19] Weifeng Ge. Deep metric learning with hierarchical triplet loss. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 269–285, 2018. 2

[20] Jacob Goldberger, Geoffrey E Hinton, Sam Roweis, and Russ R Salakhutdinov. Neighbourhood components analysis. In L. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 17. MIT Press, 2005. 2

[21] Geonmo Gu, Byungsoo Ko, , and Han-Gyu Kim. Proxy synthesis: Learning with synthetic classes for deep metric learning. In *AAAI Conference on Artificial Intelligence*, 2021. 2, 7

[22] Jian Guo and Stephen Gould. Deep cnn ensemble with data augmentation for object detection. *arXiv preprint arXiv:1506.07224*, 2015. 1

[23] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2006. 2

[24] Ben Harwood, BG Kumar, Gustavo Carneiro, Ian Reid, Tom Drummond, et al. Smart mining for deep metric learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2821–2829, 2017. 2

[25] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[26] 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. 1, 5, 7

[27] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In Defense of the Triplet Loss for Person Re-Identification. *arXiv e-prints*, page arXiv:1703.07737, Mar. 2017. 1, 2

[28] John R Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe. Deep clustering: Discriminative embeddings for segmentation and separation. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 31–35. IEEE, 2016. 1

[29] J. Hu, J. Lu, and Y. Tan. Discriminative deep metric learning for face verification in the wild. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 2

[30] Pierre Jacob, David Picard, Aymeric Histace, and Edouard Klein. Metric learning with horde: High-order regularizer for deep embeddings. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 7

[31] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and koray kavukcuoglu. Spatial transformer networks. In *Advances in Neural Information Processing Systems 28*. 2015. 2

[32] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier J. Hénaff, Matthew M. Botvinick, Andrew Zisserman, Oriol Vinyals, and João Carreira. Perceiver IO: A general architecture for structured inputs & outputs. *CoRR*, abs/2107.14795, 2021. 2, 3

[33] Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and João Carreira. Perceiver: General perception with iterative attention. In *ICML*, 2021. 2

[34] Herve Jegou, Matthijs Douze, and Cordelia Schmid. Product quantization for nearest neighbor search. *IEEE transactions on pattern analysis and machine intelligence*, 33(1):117–128, 2011. 6

[35] Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez. Aggregating local descriptors into a compact image representation. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3304–3311, 2010. 2

[36] SHICHAO KAN, Yixiong Liang, Min Li, Yigang Cen, Jianxin Wang, and Zhihai He. Coded residual transform for generalizable deep metric learning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. 2

[37] Sungyeon Kim, Dongwon Kim, Minsu Cho, and Suha Kwak. Proxy anchor loss for deep metric learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 7

[38] Wonsik Kim, Bhavya Goyal, Kunal Chawla, Jungmin Lee, and Keunjoo Kwon. Attention-based ensemble for deep metric learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2, 3

[39] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. 2015. 5

[40] Byungsoo Ko and Geonmo Gu. Embedding expansion: Augmentation in embedding space for deep metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

[41] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 554–561, 2013. 5, 6, 7

[42] Xudong Lin, Yueqi Duan, Qiyuan Dong, Jiwen Lu, and Jie Zhou. Deep variational metric learning. In *The European Conference on Computer Vision (ECCV)*, September 2018. 2

[43] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[44] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. Large-margin softmax loss for convolutional neural networks. In *ICML*, volume 2, page 7, 2016. 1

[45] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. 3

[46] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861, 2018. 3

[47] Timo Milbich, Karsten Roth, Homanga Bharadhwaj, Samarth Sinha, Yoshua Bengio, Björn Ommer, and Joseph Paul Cohen. Diva: Diverse visual feature aggregation for deep metric learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2, 7

[48] Timo Milbich, Karsten Roth, Biagio Brattoli, and Björn Ommer. Sharing matters for generalization in deep metric learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 2

[49] T. Milbich, K. Roth, S. Sinha, L. Schmidt, M. Ghassemi, and B. Ommer. Characterizing generalization under out-of-distribution shifts in deep metric learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2

[50] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[51] Yair Movshovitz-Attias, Alexander Toshev, Thomas K Leung, Sergey Ioffe, and Saurabh Singh. No fuss distance metric learning using proxies. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 360–368, 2017. 1, 2

[52] Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2, 5

[53] Hyun Oh Song, Stefanie Jegelka, Vivek Rathod, and Kevin Murphy. Deep metric learning via facility location. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5382–5390, 2017. 1

[54] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4004–4012, 2016. 1, 2, 5, 6, 7

[55] Michael Opitz, Georg Waltner, Horst Possegger, and Horst Bischof. Bier-boosting independent embeddings robustly. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5189–5198, 2017. 1, 2

[56] Michael Opitz, Georg Waltner, Horst Possegger, and Horst Bischof. Deep metric learning with bier: Boosting independent embeddings robustly. *IEEE transactions on pattern analysis and machine intelligence*, 2018. 2

[57] Qi Qian, Lei Shang, Baigui Sun, Juhua Hu, Hao Li, and Rong Jin. Softtriple loss: Deep metric learning without triplet sampling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 1, 2, 7

[58] Rahul Raguram, Jan-Michael Frahm, and Marc Pollefeys. A comparative analysis of RANSAC techniques leading to adaptive real-time random sample consensus. In *ECCV*, 2008. 2

[59] Karsten Roth, Biagio Brattoli, and Bjorn Ommer. Mic: Mining interclass characteristics for improved metric learning. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 2, 5, 7

[60] Karsten Roth, Timo Milbich, and Bjorn Ommer. Pads: Policy-adapted sampling for visual similarity learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2, 5, 7

[61] Karsten Roth, Timo Milbich, Bjorn Ommer, Joseph Paul Cohen, and Marzyeh Ghassemi. Simultaneous similarity-based

self-distillation for deep metric learning. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 9095–9106. PMLR, 18–24 Jul 2021. 2, 7

[62] Karsten Roth, Timo Milbich, Samarth Sinha, Prateek Gupta, Bjorn Ommer, and Joseph Paul Cohen. Revisiting training strategies and generalization performance in deep metric learning. In *International Conference on Machine Learning*, 2020. 2

[63] Artsiom Sanakoyeu, Pingchuan Ma, V. Tschernezki, and Björn Ommer. Improving deep metric learning by divide and conquer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021. 1, 2, 5, 7

[64] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *CVPR*, 2020. 2

[65] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 1, 2

[66] Jenny Seidenschwarz, Ismail Elezi, and Laura Leal-Taixé. Learning intra-batch connections for deep metric learning, 2021. 2, 7

[67] Sivic and Zisserman. Video google: a text retrieval approach to object matching in videos. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 1470–1477 vol.2, 2003. 2

[68] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Information Processing Systems*, pages 1857–1865, 2016. 1, 2

[69] Chi Su, Fan Yang, Shiliang Zhang, Qi Tian, Larry S Davis, and Wen Gao. Multi-task learning with low rank attribute embedding for person re-identification. In *Proceedings of the IEEE international conference on computer vision*, pages 3739–3747, 2015. 1

[70] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Kumar Gupta. Revisiting unreasonable effectiveness of data in deep learning era. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 843–852, 2017. 6

[71] Weiwei Sun, Wei Jiang, Eduard Trulls, Andrea Tagliasacchi, and Kwang Yi. Acne: Attentive context normalization for robust permutation-equivariant learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[72] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 7

[73] Fuwen Tan, Jiangbo Yuan, and Vicente Ordonez. Instance-level image retrieval using reranking transformers. In *International Conference on Computer Vision (ICCV)*, 2021. 2

[74] Eu Wern Teh, Terrance DeVries, and Graham W. Taylor. Proxynca++: Revisiting and revitalizing proxy neighborhood component analysis. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[75] Giorgos Tolias, Tomas Jenicek, and Ondřej Chum. Learning and aggregating deep local descriptors for instance-level recognition. In *European Conference on Computer Vision*, pages 460–477. Springer, 2020. 2

[76] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *CoRR*, abs/2012.12877, 2020. 2

[77] Evgeniya Ustinova and Victor Lempitsky. Learning deep embeddings with histogram loss. In *Advances in Neural Information Processing Systems*, 2016. 1

[78] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017. 2

[79] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 5, 6, 7

[80] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition, 2018. 1

[81] Jian Wang, Feng Zhou, Shilei Wen, Xiao Liu, and Yuanqing Lin. Deep metric learning with angular loss. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2593–2601, 2017. 1, 2

[82] Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R. Scott. Multi-similarity loss with general pair weighting for deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1, 2, 4, 5, 6, 7

[83] Xun Wang, Haozhi Zhang, Weilin Huang, and Matthew R Scott. Cross-batch memory for embedding learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 7

[84] 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. 2, 5, 7

[85] Hong Xuan, Abby Stylianou, and Robert Pless. Improved embeddings with easy positive triplet mining. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020. 2

[86] Yuhui Yuan, Kuiyuan Yang, and Chao Zhang. Hard-aware deeply cascaded embedding. In *Proceedings of the IEEE international conference on computer vision*, pages 814–823, 2017. 1

[87] Andrew Zhai and Hao-Yu Wu. Classification is a strong baseline for deep metric learning, 2018. 2

[88] Andrew Zhai and Hao-Yu Wu. Making classification competitive for deep metric learning. *arXiv preprint arXiv:1811.12649*, 2018. 1

[89] Jiahui Zhang, Dawei Sun, Zixin Luo, Anbang Yao, Lei Zhou, Tianwei Shen, Yurong Chen, Long Quan, and Hongen Liao. Learning two-view correspondences and geometry using order-aware network. *International Conference on Computer Vision (ICCV)*, 2019. 2

[90] Wenliang Zhao, Yongming Rao, Ziyi Wang, Jiwen Lu, and Jie Zhou. Towards interpretable deep metric learning with structural matching. In *ICCV*, 2021. 2, 3, 7

[91] Wenzhao Zheng, Zhaodong Chen, Jiwen Lu, and Jie Zhou. Hardness-aware deep metric learning. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2