Integrating Language Guidance into Vision-based Deep Metric Learning

Karsten Roth[1], Oriol Vinyals[2], Zeynep Akata[1,3]

[1]University of Tübingen, [2]DeepMind, [3]MPI for Intelligent Systems

Abstract

Deep Metric Learning (DML) proposes to learn metric spaces which encode semantic similarities as embedding space distances. These spaces should be transferable to classes beyond those seen during training. Commonly, DML methods task networks to solve contrastive ranking tasks defined over binary class assignments. However, such approaches ignore higher-level semantic relations between the actual classes. This causes learned embedding spaces to encode incomplete semantic context and misrepresent the semantic relation between classes, impacting the generalizability of the learned metric space. To tackle this issue, we propose a language guidance objective for visual similarity learning. Leveraging language embeddings of expert- and pseudo-classnames, we contextualize and realign visual representation spaces corresponding to meaningful language semantics for better semantic consistency. Extensive experiments and ablations provide a strong motivation for our proposed approach and show language guidance offering significant, model-agnostic improvements for DML, achieving competitive and state-of-the-art results on all benchmarks. Code available at github.com/ExplainableML/LanguageGuidance_for_DML.

1. Introduction

Visual similarity learning with deep networks drives important applications such as image retrieval [94, 115], face verification [19, 63], clustering [6] or contrastive supervised [49] and unsupervised representation learning [13, 39]. Deep Metric Learning (DML) has proven a useful and widely adopted framework to contextualize visual similarities by learning (deep) metric representation/embedding spaces in which a predefined distance metric, such as the euclidean or cosine distance, has a strong connection to the actual underlying semantic similarity of two samples [72, 91].

In most visual similarity tasks, transfer beyond just the training distribution and classes is crucial, which requires learned representation spaces to encode meaningful semantic context that generalizes beyond relations seen during training. However, the majority of DML methods introduce
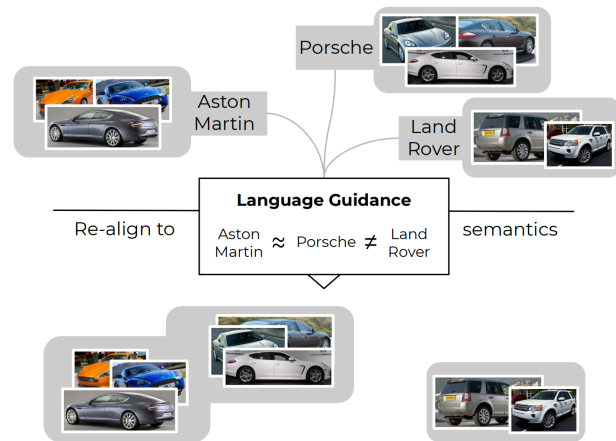

Figure 1. *Language-guidance for better semantic visual alignment.* We leverage language context to better align learned metric spaces with higher-level semantic relations and significantly improve generalization performance.

training paradigms based only around class labels provided in given datasets to define ranking tasks for networks to solve. This treats every class the same, with the arrangement of classes in embedding space solely derived from the class-label generated ranking tasks. In doing so, high-level semantic connections between different classes (e.g. sports cars vs. pickup trucks) can't be accounted for, even though semantic context beyond what can be derived from purely discriminative class labels facilitates stronger generalization especially to novel classes [60, 66, 88, 123]. And while contextualization via e.g. the definition of hierarchies [4, 10, 17, 22] can help refine classlabels, such approaches commonly rely on predefined rules or expert knowledge.

To address this problem, we propose to leverage the large corpora of readily available, large-scale pretrained natural language models [1] to provide task-independent contextualization for class labels and encourage DML models to learn semantically more consistent visual representation spaces. Using language-based pretraining for contextualization of visual similarities is long overdue - for vision-based DML, pretraining (on ImageNet) has already become standard

---

[1]S.a. transformer language models [107] with strong zero-/fewshot generalization [8, 23, 64, 83, 84] across a large variety of applications. Provided e.g. through public libraries such as huggingface [113].

[19, 66, 72, 75, 91, 92, 94, 95, 115] since it provides a strong and readily available starting point. This starting point transfers well to a multitude of downstream domains and ensures ranking tasks underlying most DML methods to be much better defined initially, improving training and generalization performance (e.g. [72, 91]). Such ImageNet pretraining is standard, and crucial, even for unsupervised DML [12, 45, 58, 120], and can similar be found in other areas of Deep Learning such as image detection [35, 36, 40, 59, 62, 85, 87]. Thus, there is little reason to bottleneck DML to only leverage visual pretraining while disregarding the potential benefits of language context.

To incorporate pretrained language models to facilitate visual similarity learning, we therefore propose *language guidance (ELG)* for DML. Given natural language class names, language embeddings and respective language similarities are computed, which are then used via distillation to re-arrange and correct visual embedding relations learned by standard DML methods. However, natural language class names require expert knowledge. To circumvent the need for such additional supervision, we further propose *pseudolabel-guidance (PLG)*. Leveraging the ubiquitously used ImageNet pretraining in DML pipelines, we define a quasi-unique collection of natural language ImageNet pseudolabels for samples and classes essentially "for free". Re-embedding these pseudolabels into pretrained language models gives access to a collection of less fine-grained, but generically applicable pseudolabel similarities which can then similarly be used for language guidance with little changes in generalization performance. Extensive experimentation and ablations support the validity of our proposed approach and showcase significant improvements to the generalization performance of DML models when using pretrained language models for additional visual semantic refinement, concurrently setting a new state-of-the-art with negligible overhead to training time.

2. Related Works

Deep Metric Learning. Visual Similarity Learning driven by advances in Deep Metric Learning has proven essential for zero-shot applications in image [88, 92, 94, 115] and video retrieval [7, 114], clustering [33, 97, 115, 119] or person re-identification [19, 63], but also for unsupervised and supervised representation learning methods relying on contrastive training [13, 39, 69, 104]. Research in Deep Metric Learning can be separated into various conceptual directions, such as the development of training surrogates via ranking losses, e.g. through pairwise [37], triplet [43], quadruplet [14], higher-order ranking constraints [73, 96, 110] or graph-based reweighting [95, 125], but also using regularizations and constraints on the incorporated distance functions [46, 103]. However, as the usage of tuples introduces an exponential increase in sample complex-

ity [94, 115], similar emphasis has also been placed in tuple selection heuristics to boost training speeds and generalization, either based on sample distances [94, 100, 110, 115], hierarchical arrangements [33] or adapted to the training process [38, 89]. Tuple complexity can also be addressed using proxies as stand-in replacements in the generation of tuples [20, 50, 71, 82, 102, 125]. However, while literature results suggests increasing generalization performance based on simple changes in re-ranking and tuple selection, recent work has instead highlighted a much stronger saturation in method performance [30, 72, 91], underlining the importance of fair and comparable training and evaluation protocols with fixed backbone network and pipeline parameter choices. For our experimental evaluation, we make sure to provide a thorough and fair comparison to allow for transferable insights. Fortunately, recent work have shown the benefit of orthogonal extensions to the common DML training approaches with strong relative improvements, such as through auxiliary feature learning [26, 66, 67, 88], artificial data generation [123], adversarial and variational extensions [25, 60], virtual class context [55], solving DML in metric subspaces [52, 75, 92, 117], knowledge distillation and transfer [15, 51, 90], distance regularization [70], mutual learning [76] or few-shot adaptation [68]. Our research into language-guided deep metric learning follows this line of thought and aims to extend the generalizability of learned metric spaces through the usage of dense language context incorporate beneficial semantics without the need for additional explicit supervision.

Crossmodal Similarity Learning. Our proposed approach, through the usage of implicit language supervision, ties into related works in crossmodal similarity learning, which finds applications in cross-modal retrieval and representation learning [11, 16, 28, 42, 44, 47, 74, 78, 105, 116, 122], where joint training can also support the generalization capabilities in single modalities, such as shown in [21, 32, 83]. For this work, our focus was placed primarily on language as a secondary modality to guide visual similarity learning. While hierarchical knowledge graphs (such as via WordNet [31] or more specific taxonomies) and color name priors have seen usage in representation learning tasks [2, 4, 9, 10, 22, 34, 81, 86, 106] and single-noun prior clustering [18], we propose to leverage dense relations between language representations generated by large, zero-shot capable language models [8, 83, 84] to re-arrange finegrained visual similarity spaces through relative matching (unlike e.g. language embedding prediction for image classification in [32]. In doing so, we find significant improvements in generalization performance across benchmarks while circumventing the need for task-specific semantic hierarchies and rules. In addition to that, we also show how our language guidance can be applied in settings where no expert class information is available.
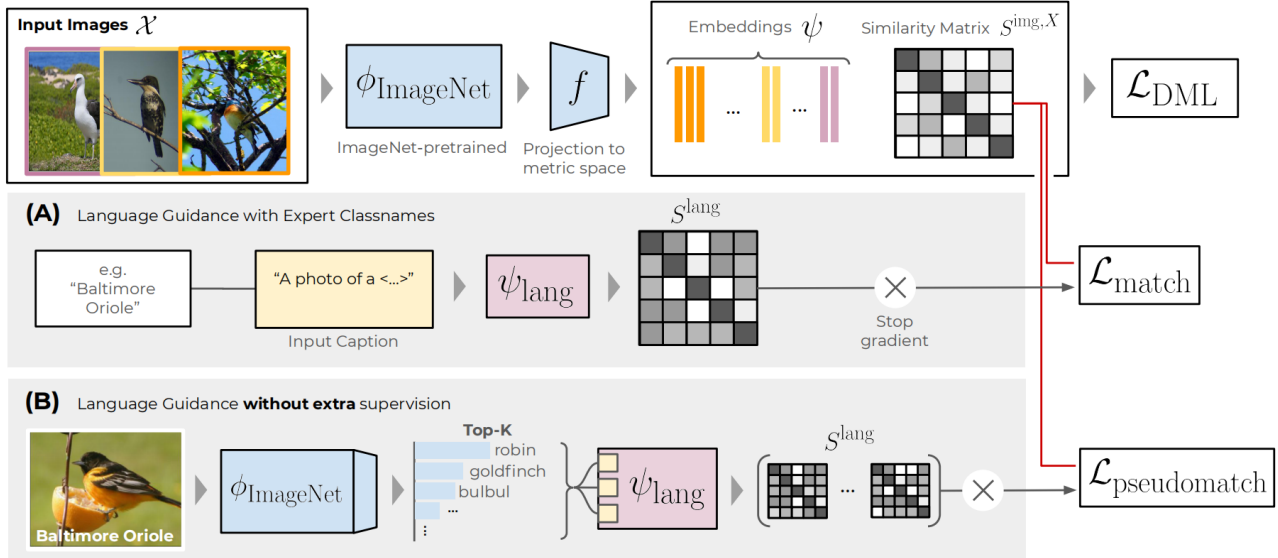
Figure 2. *Language guidance.* We extend the default DML pipeline for Visual Similarity Learning by embedding either **(A)** expert class names or **(B)** top-k ImageNet pseudolabels, which require no additional expert supervision, with a pretrained language model. This provides language similarity matrices S^{lang} which are used to guide the structuring of our finegrained visual similarity space generated by $f \circ \phi_{\text{ImageNet}}$ through distillation-based matching ($\mathcal{L}_{\text{match}}$ & $\mathcal{L}_{\text{pseudomatch}}$) between S^{lang} and image similarities $S^{\text{img, X}}$.

3. Language Guidance

This section addresses DML preliminaries, language guidance with expert class labels (§3.2) and language guidance without additional supervision (§3.3).

3.1. Preliminaries

Deep Metric Learning (DML) learns a distance metric $d_\psi(x_1, x_2)$ over images $x_i \in \mathcal{X}$ parametrized by a deep feature extraction model $\phi : \mathcal{X} \to \Phi$ (in contrast to hand-crafted feature extractors in setups predating the advent of Deep Learning, see e.g [99]) and a projection to the target metric space $f : \Phi \to \Psi \subset \mathbb{R}^d$, which defines a Mahalanobis (pseudo-)distance over features [99]. While non-end-to-end trainable methods primarily optimize for a parametrized metric over given features, in DML, both are trained jointly. This allows us to learn a projection from image space, $\psi = f \circ \phi$, that spans a metric (or embedding) space Ψ such that a predefined distance metric $d(\psi_1, \psi_2)$ on Ψ, usually the cosine or euclidean distance $d = \|\bullet, \bullet\|$, has a close connection to the true semantic similarity of input samples. Ψ is commonly normalized to the unit hypersphere $\Psi := \mathcal{S}_\Psi$ [19, 66, 91, 110, 115] for regularization [91, 109, 111, 115]. Such high-dimensional embedding spaces suitable for use with easy predefined, non-parametric metrics are attractive for fast, approximate similarity search methods [1, 48, 65]. In supervised DML, Ψ is commonly learned with ranking tasks defined using provided class label information, introducing contrastive objectives based on pairs, triplets or tuples of higher order. Taking for example

pairs (x_a, x_p) or (x_a, x_n) where $y_a = y_p \neq y_n$ with *anchor* x_a, *positive* x_p and *negative* x_n, one can define a training objective following e.g. [37] and [72] as

$$\mathcal{L} = \frac{1}{\mathcal{P}_\mathcal{B}} \sum_{(x_1, x_2) \in \mathcal{P}_\mathcal{B}} \mathbb{I}_{y_1 = y_2} \max[\gamma_p, d_\psi(x_1, x_2)] \\ - \mathbb{I}_{y_1 \neq y_2} \min[\gamma_n, d_\psi(x_1, x_2)] \tag{1}$$

with valid pairs $\mathcal{P}_\mathcal{B}$ in minibatch $\mathcal{B}$. Doing so embeds same-class samples closer (according to $d(\bullet, \bullet)$) while pushing different classes apart up to margins γ_p and γ_n. This can be easily extended to incorporate more complex relations by choice of higher-order tuples, tuple sampling heuristics [89, 94, 115] or proxy-representations [50, 71, 102].

3.2. Language Guidance with Expert Class Names

The reliance on ranking tasks defined solely using class labels however does not address high-level semantic relations between classes, even though such non-discriminative relations are crucial for strong downstream generalization [60, 66, 88, 123]. As such, we propose to leverage language semantics to better align visual representation spaces, and do so in two ways. This section will introduce the *Expert Language Guidance* (*ELG*) approach (see also Fig. 2a), which makes uses of expert class label names, while the next section covers *Pseudolabel Language Guidance* (*PLG*).

To incorporate language semantics into visual similarity learning via *ELG*, we make use of large pretrained language models ψ_{lang} like CLIP [83], BERT [23] or RoBERTa [64], which map an input sentence $c_i \in \mathcal{C}$, corresponding to the

images $x_i \in \mathcal{X}$ ground truth class label, to the respective language embedding space Ψ_{lang}. For brevity, we use $\psi_i^{\text{lang}} := \psi_{\text{lang}}(c_i)$. Given the minibatch of images $\mathcal{B}_{\mathcal{X}} \subset \mathcal{X}$ and labels $\mathcal{B}_{\mathcal{Y}} \subset \mathcal{Y}$, we generate input sequences $\mathcal{B}_{\mathcal{C}} \subset \mathcal{C}$ for the natural language model using some primer, e.g. $c_i =$ `"A photo of a $y_i$"`[2]. Let $\mathcal{B}_{\Psi_{\text{lang}}}$ be the natural language representation generated by ψ_{lang} from $\mathcal{B}_{\mathcal{C}}$ and S^{lang} the corresponding batch-wise cosine similarity matrix for all language representation in the minibatch. Similarly, let S^{img} be the corresponding visual similarity matrix for $\mathcal{B}$. We then define the language distillation loss

$$\mathcal{L}_{\text{match}}\left(\mathcal{S}^{\text{img}}, \mathcal{S}^{\text{lang}}\right) =$$

$$\frac{1}{|\mathcal{B}|} \sum_i^{|B|} \sigma\left(S_i^{\text{img},X}\right) \log\left(\frac{\sigma\left(S_i^{\text{img},X}\right)}{\sigma\left(S_i^{\text{lang}} + \gamma_{\text{lang}}\right)}\right) \quad (2)$$

as KL-Divergence matching between row-wise image- and language-level similarities $S_i^{\text{img},X}$ and S_i^{lang}, similar to contrastive distillation objectives utilized in e.g. [90,104]. Here $\sigma(\mathbf{x}_i) = \exp(\mathbf{x}_i)/\sum_j^{|\mathbf{x}|} \exp(\mathbf{x}_j)$ denotes a row-wise softmax with shift γ_{lang} to address the fidelity of the distribution matching between image and language similarity distributions. Finally, S^{lang} does not resolve similarities for samples within a class unlike S^{img} (for $x_i \neq x_j$ and $y_i = y_j$ we have $\psi_i \neq \psi_j$ and thus $S_{i,j}^{\text{img}} < 1$, but, since the classes are the same, $S^{\text{lang}} = 1$). To ensure that we do not lose intraclass resolution during distillation, we thus adapt S^{img}:

$$S_{i,j}^{\text{img},X} = \mathbb{I}_{y_i = y_j}\left[1 + \gamma_{\text{lang}}\right] + \mathbb{I}_{y_i \neq y_j}\left[S_{i,j}^{\text{img}}\right] \quad (3)$$

which masks out respective row entries of S^{img} where the classes are the same as the respective anchor class (plus the respective language offset γ_{lang}. This ignores matching within-class similarities for datasets where no sample-specific natural language description is available. Note that during training, backpropagation only occurs through S^{img}, as S^{lang} only provides the language target which the image representation space should be aligned towards. For the remainder of this work, we use "+*ELG*" to denote objectives that have been augmented with language guidance:

$$\mathcal{L}_{ELG} = \mathcal{L}_{\text{DML}} + \omega \cdot \mathcal{L}_{\text{match}} \quad (4)$$

3.3. Language Guidance without extra supervision

While *ELG* incorporates language context for visual similarity learning well, it requires expert class labelling. Fortunately, the ImageNet pretrained backbone (used in every DML pipeline) provides a suitable work-around. While previous work relies solely on the pretrained features to provide a starting point for the support over which the metric

---

[2]Special characters are adjusted to match the expected model input (e.g. `027.Shiny_Cowbird` → `Shiny Cowbird`.

space is spanned from, we go a step further and also leverage the classifier head to produce ImageNet pseudolabels. We then run both pretrained backbone and classifier head over all images within a class. For each sample, this allows us to produce softmax outputs corresponding to all ImageNet classes. These are then averaged for each class, and the top-k ImageNet pseudo-classnames $\mathcal{Y}_i^{\text{IN},k}$ are selected to represent the class (see Fig. 2(**B**) for examples).

While the pseudo classlabels are not as finegrained as expert labels, it is fair to assume that the generic object recognized in the target image x_i has, for the majority of cases, some relation with the true label y_i (e.g. `sports car` instead of `Aston Martin Coupe`). Running the language network ψ_{lang} on the pseudolabels then provides approximate semantic context. By repeating this process for all k of the top-k ImageNet-labels for each class, a more finegrained language resolution is then achieved, with the collection of top-k ImageNet-labels providing a much more unique semantic description than a single pseudolabel does. We then compute a collection of language-based similarity matrices $\left\{S^{\text{pseudolang},i}\right\}_{i \in [0,\ldots,k-1]}$, which respects the ordering of pseudolabels, i.e. $S^{\text{pseudolang},1}$ denotes the similarity matrix between language embeddings generated by embedding the respective highest matching pseudolabels, which we found to work best in practice (c.f. §4.4).

These language similarities are embedded into training by extending the matching objective in Eq. 2 to

$$\mathcal{L}_{\text{pseudomatch}}^k = \mathcal{L}_{\text{match}}\left(S^{\text{img}}, \frac{1}{k}\sum_j^k S^{\text{pseudolang,j}}\right) \quad (5)$$

by merging all $S^{\text{pseudolang}}$ (Eq. 5). In doing so, image representation are aligned to the semantics of multiple class concepts closely related to the underlying image. In addition, unlike matching in *ELG* (Eq. 2), we can now also address semantic differences between images of the same class, as pseudolabels can be extracted on a sample- instead of just the class level. This can be done by simply utilizing the unmasked image similarities S^{img} instead of its masked variant $S^{\text{img},X}$ (see Eq. 2 and 3). However, in practice we found no improvements on the evaluated benchmarks (cf. §4.4). For the remainder of this work, we mark language guidance using ImageNet-pseudolabels with "+*PLG*".

4. Experiments

This section lists experimental details (§4.1, with additional details in §A), highlights the generalization benefits of language guidance (§4.2), and experimentally motivates and ablates the proposed approach in §4.3 - §4.6.

4.1. Experimental Details

Datasets. We provide detailed evaluation of our work on the main DML benchmarks [25, 46, 60, 66, 72, 90, 91, 95,

Table 1. *State-of-the-art.* **Bold**: best results per literature setup. The results showcase competitive and state-of-the-art performance with little hyperparameter tuning. While a separation into backbones and embedding dimensions provides a fairer comparison, we note some pipeline changes that improve performance independently and should be taken into account when comparing: [a]: Better optimizer (RAdam [61] instead of Adam), [b]: Larger input images, [c]: Larger batchsizes on SOP, [d]: Combination of pooling operations in backbone.

BENCHMARKS →	CUB200 [108]			CARS196 [56]			SOP [73]		
METHODS ↓	R@1	R@2	NMI	R@1	R@2	NMI	R@1	R@10	NMI
ResNet50, 128 dim.									
Margin [115]	63.6	74.4	69.0	79.6	86.5	69.1	72.7	86.2	90.7
Div&Conq [92]	65.9	76.6	69.6	84.6	90.7	70.3	75.9	88.4	90.2
MIC [88]	66.1	76.8	69.7	82.6	89.1	68.4	77.2	89.4	90.0
PADS [89]	67.3	78.0	69.9	83.5	89.7	68.8	76.5	89.0	89.9
S2SD [90]	68.9 ± 0.3	79.0 ± 0.3	72.1 ± 0.4	87.6 ± 0.2	92.7 ± 0.2	72.3 ± 0.2	80.2 ± 0.2	91.5 ± 0.1	90.9 ± 0.1
Multisimilarity+*PLG*	67.8 ± 0.2	78.2 ± 0.2	70.1 ± 0.1	86.0 ± 0.3	91.4 ± 0.1	72.4 ± 0.2	77.9 ± 0.1	89.9 ± 0.2	90.2 ± 0.2
S2SD+*PLG*	**71.1 ± 0.1**	**80.6 ± 0.2**	**73.0 ± 0.2**	**89.1 ± 0.2**	**93.8 ± 0.2**	**73.1 ± 0.3**	**80.6 ± 0.1**	**91.8 ± 0.2**	**90.9 ± 0.1**
ResNet50, 512 dim.									
EPSHN [118]	64.9	75.3	-	82.7	89.3	-	78.3	90.7	-
NormSoft [121]	61.3	73.9	-	84.2	90.4	-	78.2	90.6	-
DiVA [66]	69.2	79.3	71.4	87.6	92.9	72.2	79.6	91.2	90.6
DCML-MDW [124]	68.4	77.9	71.8	85.2	91.8	73.9	79.8	90.8	90.8
IB-DML [95][a,b]	70.3	80.3	**74.0**	88.1	93.3	**74.8**	**81.4**	91.3	**92.6**
Multisimilarity+*PLG*	69.6 ± 0.4	79.5 ± 0.2	70.7 ± 0.1	87.1 ± 0.2	92.3 ± 0.3	73.0 ± 0.2	79.0 ± 0.1	91.0 ± 0.1	90.0 ± 0.1
S2SD+*PLG*	**71.4 ± 0.3**	**81.1 ± 0.2**	73.5 ± 0.3	**90.2 ± 0.3**	**94.4 ± 0.2**	72.4 ± 0.3	81.3 ± 0.2	**92.3 ± 0.2**	91.1 ± 0.2
Inception-BN, 512 dim.									
Group [29]	65.5	77.0	69.0	85.6	91.2	**72.7**	75.1	87.5	90.8
Multisimilarity[c] [110]	65.7	77.0	-	84.1	90.4	-	78.2	90.5	-
DR-MS [27]	66.1	77.0	-	85.0	90.5	-	-	-	-
ProxyGML [125]	66.6	77.6	69.8	85.5	91.8	72.4	78.0	90.6	90.2
ProxyAnchor [50][d]	68.4	79.2	-	86.8	91.6	-	79.1	90.8	-
Multisimilarity+*PLG*	69.2 ± 0.2	79.7 ± 0.1	70.6 ± 0.3	86.2 ± 0.2	91.5 ± 0.3	70.8 ± 0.3	78.6 ± 0.2	90.7 ± 0.1	90.0 ± 0.2
S2SD+*PLG*	**70.4 ± 0.2**	**80.5 ± 0.2**	**71.9 ± 0.3**	**88.1 ± 0.3**	**92.9 ± 0.1**	71.4 ± 0.3	**79.4 ± 0.1**	**91.2 ± 0.1**	**90.4 ± 0.2**

115, 123]: CUB200-2011 [108] (200 bird classes, 11,788 images), Cars196 [56] (196 car classes, 16,185 images) and Stanford Online Products [73] (22,634 finegrained product classes in 12 supergroups, 120,053 images). Training and test set contain different classes in all benchmarks.

Implementation Details. Our experiments use PyTorch [77] with backbones and training protocols adapted from previous research (e.g. [66, 72, 82, 90, 91, 110]) and codebases following [72, 91]. For our language backbone, we chose the language-part of CLIP [83] (*ViT-B/32*). However, any big language model can be used instead as well for significant performance improvements, as shown in §4.3. For the scaling ω of our language guidance (see Eq. 4), we found $\omega \in [1, 10]$ to work consistently for our experiments on CARS196 and CUB200-2011 and $\omega \in [0.1, 1]$ on SOP to account for the magnitude of the base loss $\mathcal{L}_{DML}$. For the state-of-the-art study in §4.2, we found these parameter values to transfer well to the other backbones and embedding dimensions. Additional details in Supp. A.

4.2. Benefits of language guidance

We have broken down the literature based on backbone architecture and embedding dimensionality, which are two of the main, DML-independent drivers for generalization performance [90, 91]. Stepwise LR-scheduling is performed at most twice based on the performance on a random validation subset (15%, see e.g. [50, 90]).

With little hyperparameter tuning, Tab. 1 shows competitive and state-of-the-art performance [3] of *PLG*-extended objectives across backbones and embedding dimensions, suggesting that language semantics are very beneficial for visual similarity learning without the need of expert labels. Note that we only compare using *PLG* as *ELG* does not offer any meaningful benefits on SOP due to the absence of finegrained expert classnames (with the exception of the twelve generic superclasses). Especially for lower-dimensional representation spaces (see ResNet50-128), significant improvements against previous state-of-the-art methods can be seen. In addition, language guidance boosts base objectives (*Multisimilarity* [110]) to match much more complex regularization approaches using e.g. RL-policies [89] or joint self-supervised training [66].

---

[2]We note that our implementation of the Multisimilarity loss follows [110] and [72], which differs in performance compared to [91].

[3]Note: NMI is less reliable, more error prone and dependent on the utilized implementation [72] than recall-based metrics, which is why beyond comparison to previous literature we primarily utilize Recall@1 and mAP@1000 (measured on recall [72, 91]).

Table 2. *Relative comparison.* We follow protocols proposed in [91][4], with no learning rate scheduling, to ensure exact comparability. The results show significant improvements when language-guidance is applied. (*) For SOP, only 12 superlabels are given for 11,318 training classes, with very few samples per class. This limits the benefits of language guidance.

Benchmarks→	CUB200	CARS196	SOP(*)
Approaches↓	R@1	R@1	R@1
Multisimilarity	62.8 ± 0.2	81.6 ± 0.3	76.0 ± 0.1
+ELG	67.3 ± 0.2	85.3 ± 0.1	76.0 ± 0.2
+PLG Top-5	67.1 ± 0.4	85.4 ± 0.2	76.4 ± 0.1
Margin, $\beta = 1.2$	62.7 ± 0.6	79.4 ± 0.5	78.0 ± 0.3
+ELG	65.3 ± 0.5	83.2 ± 0.5	77.8 ± 0.1
+PLG Top-5	65.2 ± 0.5	83.4 ± 0.4	78.3 ± 0.2
Multisimilarity + S2SD	67.7 ± 0.3	86.5 ± 0.1	77.7 ± 0.2
+ELG	68.9 ± 0.4	88.2 ± 0.2	77.8 ± 0.1
+PLG Top-5	69.0 ± 0.4	88.4 ± 0.3	78.0 ± 0.1

Table 3. *Model guidance quality.* Performance improves regardless of the exact language model and even with large-scale pretrained word embeddings s.a. FastText [5] and GloVe [79]. However, less transferable word hierarchies fall short in comparison.

Benchmarks→	CUB200-2011		CARS196	
Models↓	R@1	mAP @1000	R@1	mAP @1000
Baseline	62.8 ± 0.2	31.1 ± 0.3	81.6 ± 0.3	31.7 ± 0.1
+ CLIP-L [83]	67.3 ± 0.2	**34.8 ± 0.2**	85.3 ± 0.1	**32.7 ± 0.2**
(a) Language Models				
+ BERT [23]	66.9 ± 0.3	33.5 ± 0.2	84.9 ± 0.1	32.3 ± 0.1
+ Roberta-L [64]	67.3 ± 0.2	33.9 ± 0.3	85.1 ± 0.2	32.4 ± 0.2
+ Reformer [54]	66.7 ± 0.1	33.1 ± 0.1	85.5 ± 0.2	32.0 ± 0.2
+ GPT2 [84]	67.0 ± 0.3	33.7 ± 0.1	84.8 ± 0.4	32.4 ± 0.1
+ Top 3	**67.5 ± 0.2**	34.5 ± 0.3	**85.6 ± 0.3**	32.5 ± 0.3
(b) Word Embeddings				
+ FastText [5]	66.9 ± 0.3	33.7 ± 0.3	85.2 ± 0.1	**32.7 ± 0.1**
+ GloVe [79]	66.1 ± 0.2	33.1 ± 0.3	85.0 ± 0.2	32.1 ± 0.3
(c) Word Hierarchies				
+ WordNet [31]	63.7 ± 0.2	31.3 ± 0.2	82.5 ± 0.3	31.0 ± 0.2
+ HierMatch [2]	64.8 ± 0.3	32.2 ± 0.1	82.8 ± 0.2	31.8 ± 0.1

To ensure that benefits in performance do not solely stem from specific pipeline choices [72, 91], we provide additional experimental support following protocols in [91] with no learning rate scheduling. As baseline objectives, we select Margin loss with distance-based tuple mining [115] and the Multisimilarity loss [110], which are among the strongest baseline objectives investigates in [91] and [72]. We also apply language guidance alongside state-of-the-art dimensionality-regularization proposed in [90] to showcase benefits even for already heavily regularized objectives.

Results are highlighted in Tab. 2 (see supplementary for full table with Recall@1, NMI and mAP@1000) for all benchmark datasets. As can be seen, zero-shot generalization performance is improved significantly on CUB200-2011 and CARS196 for both the strong DML baseline objectives (Multisimiliarity [110] and Margin loss [115], see [91]) as well as state-of-the-art regularization methods such as S2SD [90], with improvements in parts of over 4% e.g. for Multisimilarity on CUB200-2011, and nearly 2% for state-of-the-art S2SD on e.g. Cars196 (86.5% → 88.4%). For SOP, we find benefits to be limited - although still significant - both for relative as well as state-of-the-art comparison(e.g. 76.0% → 76.4% relative changes). We attribute this to the much higher class-to-sample ratio in SOP (over 11k classes and only few samples per class), resulting in less specific, noisier class alignment. Unlike pseudo-label guidance however, where performance can be improved reliably, albeit only slightly, we find no benefits in using expert-label guidance. This is caused by the absence of finegrained expert class labels - instead, only twelve very coarse and generic superlabels are provided for over 11k classes, with only few samples per class. In *PLG*, at least access to the higher resolved 1k ImageNet pseudolabels is given.

4.3. Impact of different language context

This section studies the impact of different language context on the quality of *ELG*-visual representation spaces (see Table 3) with respect to different large language model architectures **(a)** as well as word embeddings (**(b)**, FastText [5] and GloVe [79][5]) trained on large text corpora.

The results show consistent and significant improvements over the vision-only baseline (Tab. 3, **(a)**) for different language model choices s.a. BERT [23] (62.8% to 66.9% for Recall@1 on CUB200) or Reformer [54] (62.8% to 66.7%), as well as for large-scale pretrained word-embeddings (e.g. FastText [5], 62.8% to 66.9%) shown in section **b**, with little computational overhead ($< 5\%$ in wall-time) and fast convergence (see Supp. B). This highlights independence of the language model choice, so long as sufficiently general knowledge about semantic class relations is available. We further provide improvements by averaging the top three language models (CLIP-L, Roberta-L and GPT2 - giving +0.2%/+0.3% on CUB200/CARS196, respectively), but choose CLIP-L for all other experiments to minimize computational overhead. Interestingly, the similar performance of large language models compared to simple, albeit large-scale pretrained, word embeddings indicates that the pretrained language context provided through large language models has yet to be fully utilized for language guidance. We leave this promising direction for future research to investigate further.

Finally, section **c** of Tab. 3 shows results for guidance based on predefined, but much more specific and less transferable word hierarchies. Specifically, we distill either from

---

[5]For classlabels with whitespaces we performed averaging of each respective word embedding.

Table 4. *Including language context.* Distillation of relative language embedding alignments offers the most benefits.

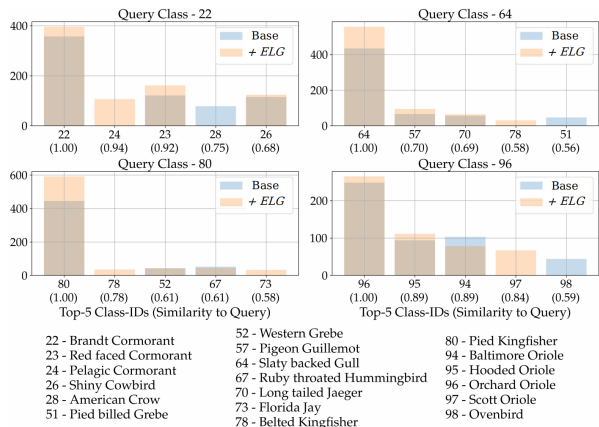| BENCHMARKS→	CUB200-2011		CARS196	
METHODS↓	R@1	mAP @1000	R@1	mAP @1000
Baseline	62.8 ± 0.2	31.1 ± 0.2	81.6 ± 0.3	31.7 ± 0.1
+ ELG	$\mathbf{67.3 \pm 0.2}$	$\mathbf{34.8 \pm 0.2}$	$\mathbf{85.3 \pm 0.1}$	$\mathbf{32.7 \pm 0.2}$
(a) Vision-Language representation matching				
DeVise-Style	64.2 ± 0.3	32.0 ± 0.1	83.9 ± 0.3	32.2 ± 0.2
ViT Prediction	65.1 ± 0.4	33.0 ± 0.1	83.2 ± 0.2	32.0 ± 0.3
CLIP-style	63.7 ± 0.4	31.9 ± 0.6	83.2 ± 0.4	32.2 ± 0.3
(b) Methods of incorporation				
Row-wise L2	65.2 ± 0.2	32.4 ± 0.2	84.4 ± 0.3	32.1 ± 0.1
Full Similarity	64.9 ± 0.3	32.3 ± 0.2	84.0 ± 0.1	32.0 ± 0.3
Mining mask	64.1 ± 0.3	32.6 ± 0.2	81.9 ± 0.2	32.3 ± 0.4
Weight scaling	59.2 ± 0.5	29.9 ± 0.4	74.6 ± 0.5	26.0 ± 0.8
No stop grad.	66.3 ± 0.4	34.1 ± 0.2	83.6 ± 0.2	30.8 ± 0.2

Figure 3. *Improved semantic retrieval.* Embedding test classes on CUB200-2011 for models with and without language guidance and evaluating the Top-5 retrieved classes (sorted by semantic similarity) to a query test classes highlights that language-guided models retrieve more samples from semantically related classes.

Wu-Palmer similarities between class names on WordNet [31] or follow Hierarchical Matching proposed in [2]. In the latter case, we leverage embeddings computed following [2] on hierarchy trees derived from wikispecies for CUB200-2011 [3] and WordNet otherwise. Training then involves minimizing the similarity between hierarchy and actual image similarities. In both cases, we find that while performance increases somewhat, benefits fall short compared to distillation-based guidance with more general dense language semantics, giving additional support for the benefits of general language knowledge as guidance signal.

4.4. Conceptual approaches to language inclusion

In this section, we motivate structural choice in *ELG* and *PLG* and compare against a range of method ablations as well as different approaches for language inclusion in Tab. 4. See Supp B.2 for more details.

More specifically, in section **a** of Tab. 4, we investigate different approaches to use language context to shape the visual representation space. For that, we evaluate training a MLP over image embeddings or a ViT [24] over sequences of feature vectors (flattened feature maps) to predict language representations by maximizing the cosine similarity between predicted and generated language embeddings (similar to prediction of separately trained word embeddings in DeVise [32] or caption predictions in multimodal settings [21]). Finally, we also look at CLIP-style training as regularizer against $\mathcal{L}_{\text{DML}}$ by directly contrasting between image and language presentation with no intermediate MLP. In all cases, we found performance to significantly lack behind our distillation-based objective (compare e.g. 63.7% on CUB200 using CLIP-style training to 67.3% using *ELG*), although separately learning a ViT model for language embedding prediction from feature sequence shows some promise for future research (65.1%). We attribute the difference in performance to the fact that our distillation-based matching in *ELG* matches the full similarity matrix of language embeddings without directly interfering with the finegrained visual similarity training.

Finally, section **(b)** of Tab. 4 looks at different distillation approaches, either matching rows via L2-Distance or the full similarity matrix via KL-Divergence. Additionally, we also include results in which we backpropagate into the language model, as well as leverage language context to directly manipulate the main DML objective $\mathcal{L}_{\text{DML}}$. For that, we adapt mining and contrastive operations in the Multisimilarity loss (see supp. B.2 for exact formulations). We find that for methods of distillation, relative alignment by row-wise KL-Divergence minimization following Eq. 2 works best, motivating the matching approach used in *ELG*. Furthermore, while some benefits where found directly adapting $\mathcal{L}_{\text{DML}}$ (62.8% to 64.1% by adjusting mining operations), more hyperparameter tuning is required to achieve reasonable results, while also being objective-specific.

We similarly also ablated architectural choices in *PLG*, looking at different number of pseudolabels, the application of hierarchical approaches to *PLG* and different methods of distilling from multiple language similarity matrices. However, as results where as expected and in line with experiments conducted in §4.4 for *ELG*, we provide the detailed results in Supp. B.3. In summary however, we found that firstly, even in the pseudolabel setting, hierarchical matching performed notably worse that distillation from large-scale pretrained language models. Secondly, we find a clear benefit of leveraging multiple pseudolabels. Finally, we find not major difference in distillation from a single, averaged pseudolabel similarity matrix versus from multiple pseudolabel-specific similarity matrix.

Table 5. *Do dense captions help?* We find class knowledge to be more valuable than generic image captions for generalization.

BENCHMARKS→	CUB200-2011		CARS196	
APPROACHES↓	R@1	mAP @1000	R@1	mAP @1000
Baseline	62.8 ± 0.2	31.1 ± 0.2	81.6 ± 0.3	31.7 ± 0.1
+ *ELG*	67.3 ± 0.2	34.8 ± 0.2	85.3 ± 0.1	32.7 ± 0.2
Caption	64.9 ± 0.4	32.2 ± 0.2	84.9 ± 0.2	31.9 ± 0.3
Caption + insertion	67.1 ± 0.2	34.3 ± 0.1	85.4 ± 0.2	32.7 ± 0.3
Classname only	67.3 ± 0.3	34.7 ± 0.1	85.1 ± 0.2	32.5 ± 0.1
Other primers	67.4 ± 0.1	34.7 ± 0.2	85.5 ± 0.3	32.7 ± 0.1

4.5. Retrieval based on semantic alignment

To understand the representation space impact of language guidance, we embed unseen test classes for models trained with and without. For every class, we use all samples as queries against the available gallery, collect the top twenty nearest retrieved samples and investigate the retrieved classes and respective counts. For both models, we sort the retrieved classes by semantic similarity to the query class (measured by a pretrained language model) and visualize the top five with their respective retrieved counts in Figure 3 for exemplary classes on CUB200-2011. Indeed, we find that language-guided models retrieve more samples from classes semantically related to the query class. For example, when querying samples from class `Brandt Cormorant`, methods not equipped with *ELG* fail completely to retrieve samples from the same bird type `Pelagic Cormorant`, which is heavily semantically related. This is fixed with language guidance and provides clear support that language guidance offers better semantic alignment for visual similarity learning. For methods equipped with *PLG*, we find very similar behaviour.

Finally, we look at the row-wise KL-divergences (see Eq. 2) between full embedding spaces, trained either with or without language guidance, and the actual language embeddings of the respective class names. After application of *ELG*-regularization, we find that alignment more than quadruples (base-to-language divergence of 1.49 ± 0.16 versus *ELG*-to-language divergence of 0.34 ± 0.09), showcasing the expected realignment based on language semantics.

4.6. Are dense captions better than class labels?

Finally, we check the need for primers (s.a. `"A photo of a `y_i`"`) and if full image captions can facilitate language guidance. We try a selection of primer sentences (including class labels only) as well as captions (e.g. `"A blue bird sitting on a branch."`) generated by a pretrained image caption network (OmniNet [80]). We also augment these captions with specific class labels by replacing dataset-dependent keywords (s.a. `bird`, giving e.g. `"A blue `y_i` sitting on a branch"`. The results in Table 5 show that the availability of class knowledge

is a stronger guidance proxy than dense image descriptions, which are often too generic ("*Caption*" vs. *+ELG*). This is supported by the boost in performance when replacing keywords with class labels ("*Caption + insertion*"). Finally, we find that the exact choice of primer sentence does not matter, with just the classname as input performing only slightly worse than the best primer for each benchmark.

5. Conclusion

In this work, we have shown that by matching the relative alignment of visual and language representations, dense language context can help significantly improve the generalization capabilities of visual similarity models. Extending this approach with "free" pseudo-labels, language guidance can be applied on top of arbitrary deep metric learning methods without the need of additional expert knowledge. Comprehensive ablation studies and benchmark experiments provide strong support for our proposed method and showcase that language guidance can help existing methods easily achieves competitive and state-of-the-art performance on all standard benchmarks.

Limitations. For language guidance to offer consistent improvements, either expert language classnames or access to suitable pseudolabels has to be available. While ImageNet pretraining is generally transferable to a wide range of downstream tasks, the degree can vary notably. This is exacerbated with high class counts, as seen in the performance on SOP. Additionally, if class separation becomes too finegrained, the likelihood of language models to provide facilitating language context can drop. Finally, for instance-based sample retrieval, where each class comprises samples for a particular class instance, language-guidance is not directly applicable, and requires additional adaptation.

Broader Impact. Improved generalization through language guidance without external expert labelling makes our method attractive as general extension to DML, and offers a new approach to visual metric representation space learning that can benefit applications s.a. image or video retrieval. We also note potential use in more controversial areas such as face ReID. However, while improvements can always be misused, we do not believe that the improvements gained substantially change the societal application. Rather, we argue that our approach offers a potential remedy to correct for possible visual biases by accounting to semantic context via respective guidance.

References

[1] Dmitry Baranchuk, Artem Babenko, and Yury Malkov. Revisiting the inverted indices for billion-scale approximate nearest neighbors. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 3

[2] Björn Barz and Joachim Denzler. Hierarchy-based image embeddings for semantic image retrieval. *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 638–647, 2019. 2, 6, 7

[3] Bjorn Barz and Joachim Denzler. Deep learning on small datasets without pre-training using cosine loss. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020. 7

[4] Luca Bertinetto, Romain Mueller, Konstantinos Tertikas, Sina Samangooei, and Nicholas A. Lord. Making better mistakes: Leveraging class hierarchies with deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 2

[5] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016. 6, 2

[6] Diane Bouchacourt, Ryota Tomioka, and Sebastian Nowozin. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In *AAAI 2018*, 2018. 1

[7] Biagio Brattoli, Joseph Tighe, Fedor Zhdanov, Pietro Perona, and Krzysztof Chalupka. Rethinking zero-shot video classification: End-to-end training for realistic applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. 1, 2

[9] Clemens-Alexander Brust, Björn Barz, and Joachim Denzler. Making every label count: Handling semantic imprecision by integrating domain knowledge. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 6866–6873, 2021. 2

[10] Clemens-Alexander Brust and Joachim Denzler. Integrating domain knowledge: Using hierarchies to improve deep classifiers. In *Asian Conference on Pattern Recognition (ACPR)*, 2019. 1, 2

[11] Wenming Cao, Qiubin Lin, Zhihai He, and Zhiquan He. Hybrid representation learning for cross-modal retrieval. *Neurocomputing*, 345:45–57, 2019. Deep Learning for Intelligent Sensing, Decision-Making and Control. 2

[12] Xuefei Cao, Bor-Chun Chen, and Ser-Nam Lim. Unsupervised deep metric learning via auxiliary rotation loss. *CoRR*, abs/1911.07072, 2019. 2

[13] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Everest Hinton. A simple framework for contrastive learning of visual representations. 2020. 1, 2

[14] 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

[15] Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Darkrank: Accelerating deep metric learning via cross sample similarities transfer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018. 2

[16] Yanbei Chen, Yongqin Xian, A. Sophia Koepke, Ying Shan, and Zeynep Akata. Distilling audio-visual knowledge by compositional contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7016–7025, June 2021. 2

[17] Ph Cimiano, Andreas Hotho, and Steffen Staab. Learning concept hierarchies from text corpora using formal concept analysis. *Journal of Artificial Intelligence Research*, 24, 09 2011. 1

[18] Niv Cohen and Yedid Hoshen. The single-noun prior for image clustering, 2021. 2

[19] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4685–4694, 2019. 1, 2, 3

[20] Jiankang Deng, Jia Guo, Jing Yang, Alexandros Lattas, and Stefanos Zafeiriou. Variational prototype learning for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11906–11915, June 2021. 2

[21] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. *CoRR*, abs/2006.06666, 2020. 2, 7

[22] Thomas Deselaers and Vittorio Ferrari. Visual and semantic similarity in imagenet. In *CVPR 2011*, pages 1777–1784, 2011. 1, 2

[23] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 1, 3, 6, 2

[24] 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. In *International Conference on Learning Representations*, 2021. 7, 2

[25] 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, 4

[26] Natalie Dullerud, Karsten Roth, Kimia Hamidieh, Nicolas Papernot, and Marzyeh Ghassemi. Is fairness only metric deep? evaluating and addressing subgroup gaps in deep metric learning. In *International Conference on Learning Representations*, 2022. 2

[27] Ujjal Kr Dutta, Mehrtash Harandi, and Chellu Chandra Sekhar. Unsupervised deep metric learning via orthogonality based probabilistic loss. *IEEE Transactions on Artificial Intelligence*, 1(1):74–84, 2020. 5

[28] Maksim Dzabraev, Maksim Kalashnikov, Stepan Komkov, and Aleksandr Petiushko. Mdmmt: Multidomain multimodal transformer for video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3354–3363, June 2021. 2

[29] Ismail Elezi, Sebastiano Vascon, Alessandro Torcinovich, Marcello Pelillo, and Laura Leal-Taixé. The group loss for deep metric learning. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 277–294, Cham, 2020. Springer International Publishing. 5

[30] Istvan Fehervari, Avinash Ravichandran, and Srikar Appalaraju. Unbiased evaluation of deep metric learning algorithms, 2019. 2

[31] Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. Bradford Books, 1998. 2, 6, 7

[32] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc' Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. 2, 7

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

[34] S. Ging, M. Zolfaghari, H. Pirsiavash, and T. Brox. Coot: Cooperative hierarchical transformer for video-text representation learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 22605–22618. Curran Associates, Inc., 2020. 2

[35] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015. 2

[36] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 2

[37] 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, 3

[38] 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

[39] 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)*, June 2020. 1, 2

[40] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 346–361, Cham, 2014. Springer International Publishing. 2

[41] 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

[42] Yonghao He, Shiming Xiang, Cuicui Kang, Jian Wang, and Chunhong Pan. Cross-modal retrieval via deep and bidirectional representation learning. *IEEE Transactions on Multimedia*, 18(7):1363–1377, 2016. 2

[43] 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

[44] Xin Huang and Yuxin Peng. Cross-modal deep metric learning with multi-task regularization. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pages 943–948, 2017. 2

[45] Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondřej Chum. Mining on manifolds: Metric learning without labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[46] 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, 4

[47] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. 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 4651–4664. PMLR, 18–24 Jul 2021. 2

[48] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*, 2017. 3

[49] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning, 2020. 1

[50] Sungyeon Kim, Dongwon Kim, Minsu Cho, and Suha Kwak. Proxy anchor loss for deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2, 3, 5

[51] Sungyeon Kim, Dongwon Kim, Minsu Cho, and Suha Kwak. Embedding transfer with label relaxation for improved metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3967–3976, June 2021. 2

[52] 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

[53] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 1

[54] Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. In *International Conference on Learning Representations*, 2020. 6, 2

[55] Byungsoo Ko, Geonmo Gu, and Han-Gyu Kim. Learning with memory-based virtual classes for deep metric learning, 2021. 2

[56] 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

[57] Anders Krogh and John A. Hertz. A simple weight decay can improve generalization. In *Advances in Neural Information Processing Systems*. 1992. 1

[58] Yang Li, Shichao Kan, and Zhihai He. Unsupervised deep metric learning with transformed attention consistency and contrastive clustering loss. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 141–157, Cham, 2020. Springer International Publishing. 2

[59] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2

[60] 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. 1, 2, 3, 4

[61] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. In *International Conference on Learning Representations*, 2020. 5

[62] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 21–37, Cham, 2016. Springer International Publishing. 2

[63] 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. 1, 2

[64] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692, 2019. 1, 3, 6, 2

[65] Yu A. Malkov and D. A. Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4):824–836, 2020. 3

[66] 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. *CoRR*, abs/2004.13458, 2020. 1, 2, 3, 4, 5

[67] T. Milbich, K. Roth, B. Brattoli, and B. Ommer. Sharing matters for generalization in deep metric learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020. 2

[68] Timo Milbich, Karsten Roth, Samarth Sinha, Ludwig Schmidt, Marzyeh Ghassemi, and Björn Ommer. Characterizing generalization under out-of-distribution shifts in deep metric learning. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. 2

[69] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 6706–6716. IEEE, 2020. 2

[70] Deen Dayal Mohan, Nishant Sankaran, Dennis Fedorishin, Srirangaraj Setlur, and Venu Govindaraju. Moving in the right direction: A regularization for deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[71] 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. 2, 3

[72] Kevin Musgrave, Serge J. Belongie, and Ser-Nam Lim. A metric learning reality check. *CoRR*, abs/2003.08505, 2020. 1, 2, 3, 4, 5, 6

[73] 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. 2, 5

[74] Andreea-Maria Oncescu, A. Sophia Koepke, João F. Henriques, Zeynep Akata, and Samuel Albanie. Audio retrieval with natural language queries, 2021. 2

[75] 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

[76] Wonpyo Park, Wonjae Kim, Kihyun You, and Minsu Cho. Diversified mutual learning for deep metric learning. 2020. 2

[77] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017. 5, 1

[78] Yuxin Peng and Jinwei Qi. Cm-gans: Cross-modal generative adversarial networks for common representation learning. *ACM Trans. Multimedia Comput. Commun. Appl.*, 15(1), Feb. 2019. 2

[79] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543, 2014. 6, 2

[80] Subhojeet Pramanik, Priyanka Agrawal, and Aman Hussain. Omninet: A unified architecture for multi-modal multi-task learning. *CoRR*, abs/1907.07804, 2019. 8

[81] Raphael C. Prates, Cristianne R. S. Dutra, and William Robson Schwartz. Predominant color name indexing structure for person re-identification. In *2016 IEEE International Conference on Image Processing, ICIP 2016, Phoenix, AZ, USA, September 25-28, 2016*, pages 779–783. IEEE, 2016. 2

[82] 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. 2, 5, 1

[83] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. 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 8748–8763. PMLR, 18–24 Jul 2021. 1, 2, 3, 5, 6

[84] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. 2019. 1, 2, 6

[85] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2

[86] Joseph Redmon and Ali Farhadi. Yolo9000: Better, faster, stronger. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6517–6525, 2017. 2

[87] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. 2

[88] 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*, pages 8000–8009, 2019. 1, 2, 3, 5

[89] 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, 3, 5

[90] 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, 4, 5, 6, 1

[91] 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 Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8242–8252. PMLR, 13–18 Jul 2020. 1, 2, 3, 4, 5, 6

[92] Artsiom Sanakoyeu, Vadim Tschernezki, Uta Buchler, and Bjorn Ommer. Divide and conquer the embedding space for metric learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 5

[93] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020. 2

[94] 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, 3

[95] Jenny Denise Seidenschwarz, Ismail Elezi, and Laura Leal-Taixé. Learning intra-batch connections 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 9410–9421. PMLR, 18–24 Jul 2021. 2, 4, 5

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

[97] Kihyuk Sohn, Wenling Shang, Xiang Yu, and Manmohan Chandraker. Unsupervised domain adaptation for distance metric learning. In *International Conference on Learning Representations*, 2019. 2

[98] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 16857–16867. Curran Associates, Inc., 2020. 2

[99] Juan-Luis Suárez, Salvador García, and Francisco Herrera. A tutorial on distance metric learning: Mathematical foundations, algorithms and software. *CoRR*, abs/1812.05944, 2018. 3

[100] Yumin Suh, Bohyung Han, Wonsik Kim, and Kyoung Mu Lee. Stochastic class-based hard example mining for deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[101] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Computer Vision and Pattern Recognition (CVPR)*, 2015. 1

[102] Eu Wern Teh, Terrance DeVries, and Graham W. Taylor. Proxynca++: Revisiting and revitalizing proxy neighborhood component analysis. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXIV*, volume 12369 of *Lecture Notes in Computer Science*, pages 448–464. Springer, 2020. 2, 3

[103] Yurun Tian, Axel Barroso Laguna, Tony Ng, Vassileios Balntas, and Krystian Mikolajczyk. Hynet: Learning local descriptor with hybrid similarity measure and triplet loss. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 7401–7412. Curran Associates, Inc., 2020. 2

[104] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *CoRR*, abs/1910.10699, 2019. 2, 4

[105] Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models, 2021. 2

[106] Joost van de Weijer and Fahad Shahbaz Khan. An overview of color name applications in computer vision. In Alain Trémeau, Raimondo Schettini, and Shoji Tominaga, editors, *Computational Color Imaging*, pages 16–22, Cham, 2015. Springer International Publishing. 2

[107] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 1

[108] 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

[109] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. *arXiv preprint arXiv:2005.10242*, 2020. 3

[110] 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. 2, 3, 5, 6, 1

[111] Eric W. Weisstein. Hypersphere, 2002. 3

[112] Ross Wightman. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019. 1

[113] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac,

Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing, 2020. 1, 2

[114] Michael Wray, Hazel Doughty, and Dima Damen. On semantic similarity in video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3650–3660, June 2021. 2

[115] 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. 1, 2, 3, 4, 5, 6

[116] Xing Xu, Li He, Huimin Lu, Lianli Gao, and Yanli Ji. Deep adversarial metric learning for cross-modal retrieval. *World Wide Web*, 22(2):657–672, Mar. 2019. 2

[117] Hong Xuan, Richard Souvenir, and Robert Pless. Deep randomized ensembles for metric learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 723–734, 2018. 2

[118] 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. 5

[119] Jiexi Yan, Lei Luo, Cheng Deng, and Heng Huang. Unsupervised hyperbolic metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12465–12474, June 2021. 2

[120] Mang Ye, Xu Zhang, Pong C. Yuen, and Shih-Fu Chang. Unsupervised embedding learning via invariant and spreading instance feature. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[121] Andrew Zhai and Hao-Yu Wu. Making classification competitive for deep metric learning. *CoRR*, abs/1811.12649, 2018. 5

[122] Liangli Zhen, Peng Hu, Xu Wang, and Dezhong Peng. Deep supervised cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[123] 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. 1, 2, 3, 4

[124] Wenzhao Zheng, Chengkun Wang, Jiwen Lu, and Jie Zhou. Deep compositional metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9320–9329, June 2021. 5

[125] Yuehua Zhu, Muli Yang, Cheng Deng, and Wei Liu. Fewer is more: A deep graph metric learning perspective using fewer proxies. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 17792–17803. Curran Associates, Inc., 2020. 2, 5