Dynamic Metric Learning: Towards a Scalable Metric Space to Accommodate Multiple Semantic Scales

Yifan Sun[1,3], Yuke Zhu[1], Yuhan Zhang[2], Pengkun Zheng[1], Xi Qiu[1], Chi Zhang[1], Yichen Wei[1]
[1]Megvii Technology [2]Beihang University [3]Baidu Research
sunyf15@tsinghua.org.cn Fzhangchi,weiyichenG@megvii.com

Abstract

This paper introduces a new fundamental characteristic, i.e.*, the dynamic range, from real-world metric tools to deep visual recognition. In metrology, the dynamic range is a basic quality of a metric tool, indicating its flexibility to accommodate various scales. Larger dynamic range offers higher flexibility. In visual recognition, the multiple scale problem also exist. Different visual concepts may have different semantic scales. For example, "Animal" and "Plants" have a large semantic scale while "Elk" has a much smaller one. Under a small semantic scale, two different elks may look quite* different *to each other . However, under a large semantic scale (*e.g.*, animals and plants), these two elks should be measured as being* similar.

Introducing the dynamic range to deep metric learning, we get a novel computer vision task, i.e.*, the Dynamic Metric Learning. It aims to learn a scalable metric space to accommodate visual concepts across multiple semantic scales. Based on three types of images,* i.e.*, vehicle, animal and online products, we construct three datasets for Dynamic Metric Learning. We benchmark these datasets with popular deep metric learning methods and find Dynamic Metric Learning to be very challenging. The major difficulty lies in a conflict between different scales: the discriminative ability under a small scale usually compromises the discriminative ability under a large one, and vice versa. As a minor contribution, we propose Cross-Scale Learning (CSL) to alleviate such conflict. We show that CSL consistently improves the baseline on all the three datasets. The datasets and the code will be publicly available at https://github.com/SupetZYK/DynamicMetricLearning.*

1. Introduction

This papers consider the deep metric learning for visual recognition and supplements it with an important concept in metrology, *i.e.*, the dynamic range. In metrology, the dynamic range is defined as the ratio between the largest and the smallest scale that a metric tool can provide. It is a basic quality of a metric, indicating the flexibility to accommodate various scales. We argue that such flexibility is also important for deep metric learning, because different visual concepts indeed correspond to different semantic scales. However, after a rethink on current deep metric learning tasks, we find that they all give NO consideration to the dynamic range. Therefore, we introduce the dynamic range to endow a single deep metric with flexibility among multiple semantic granularities. Potentially, it may reveal a new perspective in understanding the generalization ability of deep visual recognition.

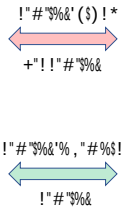

Figure 1. Visual concepts have various semantic scales, which impact on the similarity measuring result. Under the small scale of "Elk", these two elks look quite different. Under the large scale of "Animal", they should be measured as being similar.

We explain the importance of "dynamic range" with a comparison between the deep metric learning and the real-world metric tools. In real world, a metric tool typically has a dynamic range. For example, a ruler has a lot of markings to indicate different lengths ranged from "1 mm" to "10 cm" or even to several meters. Rulers with only one single scale in "1 mm" or "10 cm" would have no use in daily life. Arguably, the dynamic range is essential to a metric tool, enabling it to measure objects of different sizes. In visual recognition, the visual concepts also have various semantic scales. For example, "Animals" and "Plants" have a large semantic scale, while "Elk" has a much smaller one. When we try to measure the similarity between two images, which is the aim of metric learning, the underlying semantic scales impact on the result. In Fig. 1, the two elks look quite different to each other. However, under the large scale

of “Animals”, they should be judged as being similar.

There is no consideration for the dynamic range in current deep metric learning tasks, *e.g.*, face recognition [30, 28, 5, 27, 34, 19, 12, 3], person re-identification [29, 37, 25, 23, 24], and vehicle re-identification (re-ID) [15, 7, 38, 39]. They all focus on learning a metric for a single specified semantic scale (*e.g.*, the identity of face, pedestrian and vehicles, respectively). The single-scaled metric lacks flexibility and may become inaccurate if the scale of interest changes. We validate this point with a toy scenario based on vehicle retrieval. In Fig. 2, two users use a same query image with different intentions. In the first row, the intention is to retrieve the cars with the same *identity*, while the second row is to retrieve the cars with the same *body type* (*i.e.*, “SUV”). A discriminative metric for vehicle re-ID (which learns to identify each vehicle) satisfies the first intention. With a similarity threshold $\mathbf{T}$, it accurately separates the true matches and the false matches to the query image. However, it lacks discriminative ability for recognizing the same body type, which corresponds to a larger semantic scale than the identity. Maintaining $\mathbf{T}$ as the threshold, it fails to recall all the true matches. If we lower the similarity thresh hold to $\mathbf{T}$ to promote the recall rate, the accuracy dramatically decreases (refer to Section 6.1 for experimental evidence). We thus infer that a single-scaled metric does not fit novel semantic scales due to the lack of flexibility.

Introducing the dynamic range to deep metric learning, we get a new task, *i.e.*, the *Dynamic Metric Learning* (DyML). DyML aims to learn a scalable metric space to accommodate multiple semantic scales. In another word, a metric for DyML should be discriminative under several semantic granularities across a wide range. To promote the research on DyML, we construct three datasets based on vehicle, animal and product, respectively. All these datasets have three different semantic scales, *i.e.*, fine, middle and coarse. We benchmark these datasets with a variety of popular deep metric learning methods, *e.g.*, Cosface [31], Circle Loss [22], triplet loss [20], N-pair loss [21]. Extensive experiments show that DyML is very challenging. Even when the deep model learns from all the semantic scales in a multi-task manner, it does not naturally obtain a good dynamic range. The major difficulty lies in a conflict between different scales: the discriminative ability under a small scale usually compromises the discriminative ability under a large one, and vice versa. To alleviate such conflict, we device a simple method named *Cross-Scale Learning* (CSL). CSL uses the within-class similarity of the smallest scale as the unique reference to contrast the between-class similarity of all the scales, simultaneously. Such learning manner is similar to the fact that all the markings on a ruler share “0” as the start. Experimental results confirm that CSL brings consistent improvement over the baselines.

To sum up, this paper makes the following four contributions:

- We propose Dynamic Metric Learning by supplementing deep metric learning with dynamic range. In contrary to canonical metric learning for visual recognition, DyML desires discriminative ability across multiple semantic scales.
- We construct three datasets for DyML, *i.e.*, DyML-Vehicle, DyML-Animal and DyML-Product. All these datasets contain images under multiple semantic granularities for both training and testing.
- We benchmark these DyML datasets with popular metric learning methods through extensive experiments. Experimental investigations show that DyML is very challenging due to a conflict between different semantic scales.
- As a minor contribution, we propose Cross-Scale Learning for DyML. CSL gains better dynamic range and thus consistently improves the baseline.

2. Related work

2.1. Deep Metric Learning.

Deep metric learning (DML) plays a crucial role in a variety of computer vision applications, *e.g.*, face recognition [30, 22, 5, 27, 12, 3], person re-identification [29, 37, 25, 23, 24], vehicle re-identification [15, 7, 38, 39] and product recognition [17, 1, 6]. Generally, these tasks aim to retrieve all the most similar images to the query image.

During recent years, there has been remarkable progresses [30, 28, 5, 27, 34, 19, 12, 3, 22] in deep metric learning. These methods are usually divided into two types, *i.e.*, pair-based methods and classification-based methods. Pair-based methods (*e.g.*, Triplet loss [20], N-pair loss [21], Multi-Simi loss [33]) optimize the similarities between sample pairs in the deeply-embedded feature space. In contrast, classification-based methods learn the embedding by training a classification model on the training set, *e.g.*, Cosface[31], ArcFace[5], NormSoftmax[35] and proxy NCA[16]. Moreover, a very recent work, *i.e.*, Circle Loss[22], considers these two learning manners from a unified perspective. It provides a general loss function compatible to both pair-based and classification-based learning.

Compared with previous metric learning researches, the Dynamic Metric Learning lays emphasis on the capacity to simultaneously accommodate multiple semantic scales. This new characteristic significantly challenges previous metric learning methods (as to be detailed in Section 6.2).

2.2. Hierarchical Classification.

We clarify the difference between DyML and a “look-alike” research area, *i.e.*, the hierarchical classification [26, 9, 2, 10, 13]. Dynamic Metric Learning organizes multiple semantic scales in a hierarchical manner (as to be detailed in Section 3.2), which may seem similar to the hierarchical

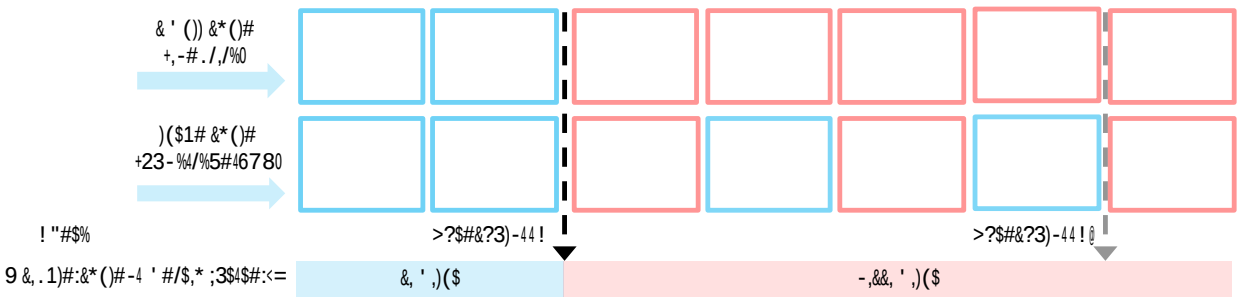

Figure 2. A single-scaled metric does not fit novel semantic scales. An accurate metric for vehicle re-ID (a small semantic scale) becomes inaccurate for recognizing the same body type "SUV" (*i.e.*, a relatively larger semantic scale) in the second row. Using the threshold $\mathbf{T}$ fails to recall all the true matches, while using a lower threshold $\mathbf{T}$ incurs false positive matches. The positive and negative matches are bounded with red and blue boxes, respectively. The images are from DyML-Vehicle.

classification. However, DyML significantly differs from hierarchical classification in two major aspects.

First, DyML belongs to metric learning domain, in which the training data and testing data has no class intersections. Correspondingly, the learned metric has to be generalized to unseen classes. In contrast, the hierarchical classification belongs to image classification task. The training data and the testing data share same classes, so there is no unseen classes during testing.

Second, DyML uses the hierarchical information for learning a single hidden layer (*i.e.*, the deep embedding layer). All the semantic scales (coarse, medium and fine levels) are equally important for DyML. In contrast, the hierarchical classification methods mainly cares about the accuracy on the fine level, and all the other semantic scales are used only for auxiliary supervision (in preceding layers before the final classification layer).

Moreover, the hierarchical data structure is NOT prerequisite for DyML. We organize the multiple semantic scales in hierarchy mainly for efficiency consideration.

3. Dynamic Metric Learning

3.1. Task Formulation

Let us assume there are $\mathbf{C}$ categories of images $\{\mathbf{l}_1, \mathbf{l}_2, \cdots, \mathbf{l}_C\}$. Each $\mathbf{l}_i$ is consisted of several images. Given a random category $\mathbf{l}_i$, if there exists another category $\mathbf{l}_{j \neq i}$ looks very similar to $\mathbf{l}_i$, we consider these categories (as well as the corresponding images) jointly form a small semantic scale. In contrast, given a random category $\mathbf{l}_i$, if its nearest neighbor $\mathbf{l}_{j \neq i}$ looks quite different from $\mathbf{l}_i$, we consider these categories jointly form a large semantic scale. To be intuitive, we take the animals as an example. According to biological taxonomy, animals may be divided by "phylum", "class", "order", "family", "genus", *etc*. The categories in the "phylum" form a large semantic scale, while the categories in "genus" form a relatively small one.

Given multiple semantic scales $\mathbf{S}^1, \mathbf{S}^2, \cdots, \mathbf{S}^M$, DyML ensembles all of them to expand a wide semantic range R, which is formulated by:

$$\mathrm{R} = \{\mathbf{S}^1, \mathbf{S}^2, \cdots, \mathbf{S}^M\}, \tag{1}$$

in which a random semantic scale $\mathbf{S}^i$ contains $\mathbf{N}^i$ labeled images, *i.e.*, $\{(\mathbf{x}^i_j, \mathbf{l}^i_j)\}$ $(\mathbf{j} = 1, 2, \cdots, \mathbf{N}^i)$.

DyML aims to learn a single metric space $\mathbf{M}$ with discriminative ability across the whole semantic range R. Concretely, through $\mathbf{M}$, samples of a same class are close to each other, and the samples of different classes are far away, regardless of the underlying semantic scale $\mathbf{S}^i$. To evaluate the discriminative ability under each scale, DyML adopts the image retrieval paradigm. Given a query image $\mathbf{x}^i_q$ from $\mathbf{S}^i$, DyML employs $\mathbf{M}$ to calculate its similarity scores between all the other images $\mathbf{x}^i_{j \neq q}$. According to descending order of the similairty scores, DyML get a ranking list, denoted as $\{\mathbf{r}^i_1, \mathbf{r}^i_2, \cdots, \mathbf{r}^i_{N_i}\}$, where $\mathbf{r}^i_{j \neq q}$ is the sorted index of image $\mathbf{x}^i_{j \neq q}$. An ideal ranking list is to place all the positive images (*i.e.*, images from a same class) in front of the negative images. In another word, $\mathbf{r}^i_j$ should be small if $\mathbf{x}^i_j$ and $\mathbf{x}^i_q$ are within a same class (*i.e.*, $\mathbf{l}^i_j = \mathbf{l}^i_q$). Formally, the objective function of DyML is formulated as,

$$\min \sum_{j=1}^{N_j} \mathbf{r}^i_{j \neq q} \, \mathbf{l}(\mathbf{l}^i_j, \mathbf{l}^i_p), \qquad \square \mathbf{S}^i \square \, \mathrm{R} \tag{2}$$

in which $\mathbf{l}(\mathbf{l}^i_j, \mathbf{l}^i_p) = 1$ if $\mathbf{l}^i_j = \mathbf{l}^i_q$ and $\mathbf{l}(\mathbf{l}^i_j, \mathbf{l}^i_p) = 0$ if $\mathbf{l}^i_j \neq \mathbf{l}^i_q$.

3.2. Hierarchical Modification for Efficiency

We note that multiple semantic scales may share images with each other to reduce the cost for image collection. In another word, a single image $\mathbf{x}$ may simultaneously belong to multiple $\mathbf{S}$ (with the superscript omitted). It is because an image in itself may have several visual concepts. For example, an image of elk (in Fig. 1) may correspond to "Elk" (small scale), "Deer" (medium scale) or "Mammal"

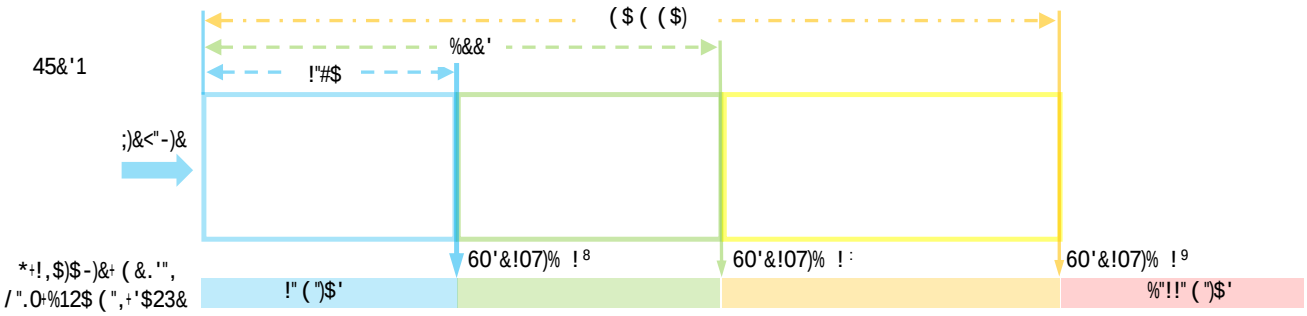

Figure 3. A multi-scaled metric with a dynamic range. It is capable to accommodate different semantic scales. Given a "sika" as the query image, it retrieves images of "sika", "deer" and "mammal" with descending similarity thresholds, *i.e.*, $T^1 > T^2 > \cdots > T^M$. The images are from DyML-Animal.

(large scale). In practice, when preparing the dataset (as to be detailed in Section 4), we annotate each image x_j with a set of labels L_j, which is formulated as:

$$R = \{(x_1, L_1), (x_2, L_2), \cdots, (x_N, L_N)\}, \quad (3)$$

in which N is the total number of images, $L_j = \{l_j^1, l_j^2, \cdots, l_j^M\}$ is the complete label set for x_j. Specifically, l_j^i denotes the label of x_j under semantic scale S_i. $S^1, \cdots, S^M$ have gradually ascending scales, *i.e.*, $S^1 \sqsubset S^2 \sqsubset \cdots \sqsubset S^M$. Multiple classes under S^i may belong to a same large-scale class in $S^j (j > i)$. Consequentially, the semantic scales of DyML follow a hierarchical ordering.

3.3. A Multi-scaled Metric for DyML

Under the hierarchical modification, *i.e.*, $S^1 \sqsubset S^2 \sqsubset \cdots \sqsubset S^M$, we analyze the property of a multi-scaled metric M. To be accurate under all the scales, M should satisfy the following two criteria:

• **Single-scale criterion.** Under a same scale, M maintains within-class compactness, as well as between-class discrepancy, *i.e.*, the within-class similarity scores are larger than the between-class similarity scores. It is a common criterion in all metric learning tasks.

• **Cross-scale criterion.** Second, under any two different scales $S^i \sqsubset S^j$, the within-class similarity scores under S^i should be larger than that under S^j. It is a unique criterion in Dynamic Metric Learning.

With both the single-scale criterion and the cross-scale criterion satisfied, M uses descending thresholds (*i.e.*, $T^1 > T^2 > \cdots > T^M$) to recall true matches under ascending semantic scales (*i.e.*, $S^1 \sqsubset S^2 \sqsubset \cdots \sqsubset S^M$), as illustrated in Fig. 3. Given an image of "sika deer" as the query, the gallery images of "sika deer", "deer (but not sika)", "mammal (but not deer)", "bird", "reptile" are recognized as having gradually-decreasing similarities. It thus looks like a ruler with multiple markings to accommodate objects with various sizes.

4. DyML Datasets

4.1. Description

Overview. This paper provides three datasets for Dynamic Metric Learning research, *i.e.*, DyML-Vehicle, DyML-Animal and DyML-Product. We collect all the source images from publicly available datasets and supplement them with some manual annotations to enrich the semantic scales. Overall, these datasets have the following common properties:

• *Three hierarchical levels of labelling*. Each image may have at most three labels corresponding to coarse, middle and fine levels, respectively. A coarse class contains several middle classes. Similarly, a middle class contains several fine classes.

• *Abundant semantic scales*. Although the datasets are hierarchically organized into three levels, the actual semantic scales are even more abundant. It is because each level may contain several semantic scales. For example, in DyML-Animal, the visual concepts contained in the middle level are consisted of "order", "family" and "genus".

• *No class intersections between training and testing set*. In accordance to the popular metric learning settings, the testing classes are novel to training classes (except for the coarse level). The total classes under coarse level are very limited. To obtain enough training / testing classes, we allow intersections under the coarse level, and insist open-set setting. In another word, under the coarse level, some testing classes exist in the training set, while some other testing classes are novel.

The quantitative descriptions of all these three datasets are summarized in Table 1.

DyML-Vehicle merges two vehicle re-ID datasets PKU VehicleID [11], VERI-Wild [14]. Since these two datasets have only annotations on the identity (fine) level, we manually annotate each image with "model" label (*e.g.*, Toyota Camry, Honda Accord, Audi A4) and "body type" label (*e.g.*, car, suv, microbus, pickup). Moreover, we label all

Datasets		DyML-Vehicle		DyML-Animal		DyML-Product	
		Train	Test	Train	Test	Train	Test
Coarse	Classes	5	6	5	5	36	6
	Images	343.1 K	5.9 K	407.8 K	12.5 K	747.1 K	1.5 K
Middle	Classes	89	127	28	17	169	37
	Images	343.1 K	34.3 K	407.8 K	23.1 K	747.1 K	1.5 K
Fine	Classes	36,301	8,183	495	162	1,609	315
	Images	343.1 K	63.5 K	407.8 K	11.3 K	747.1 K	1.5 K

Table 1. Three datasets, *i.e.*, DyML-Vehicle, DyML-Animal, DyML-Product for Dynamic Metric Learning. We collect the raw images from publicly available datasets and supplement them with abundant multi-scale annotations. Each dataset has three hierarchical labels ranging from coarse to fine. Some level contains several semantic scales. Under the middle level and fine level, there is no intersection between training and testing classes. The coarse level allows certain class intersections and yet insists on the open-set setting.

the taxi images as a novel testing class under coarse level.

DyML-Animal is based on animal images selected from ImageNet-5K [4]. It has 5 semantic scales (*i.e.*, classes, order, family, genus, species) according to biological taxonomy. Specifically, there are 611 "species" for the fine level, 47 categories corresponding to "order", "family" or "genus" for the middle level, and 5 "classes" for the coarse level. We note some animals have contradiction between visual perception and biological taxonomy, *e.g.*, whale in "mammal" actually looks more similar to fish. Annotating the whale images as belonging to mammal would cause confusion to visual recognition. So we take a detailed check on potential contradictions and intentionally leave out those animals.

DyML-Product is derived from iMaterialist-2019 [1], a hierarchical online product dataset. The original iMaterialist-2019 offers up to 4 levels of hierarchical annotations. We remove the coarsest level and maintain 3 levels for DyML-Product.

4.2. Evaluation Protocol

The overall CMC and mAP. DyML sets up the evaluation protocol based on two popular protocols adopted by image retrieval, *i.e.*, the Cumulated Matching Characteristics (CMC) [32] and the mean Average Precision (mAP) [36] The criterion of CMC indicates the probability that a true match exists in the top-K sorted list. In contrast, the criterion of mAP considers both precision and recall of the retrieval result. When there are multiple ground-truth for a query (which is the common case), mAP lays emphasis on the capacity of recognizing all the positive matches, especially those difficult ones.

To get an **overall evaluation** on the discriminative ability under all the semantic scales, DyML first evaluates the performance under each level and then averages the results under three levels (*i.e.*, fine, middle and coarse). Notablly, the level information is not accessible to the evaluated metric. Manually using the level information of the query to fit the underlying scale is not allowed. It thus prohibits learning several single-scaled metrics and manually choosing an appropriate one to fit each query. The reason is that, in reality, 1) user will not know which metric exactly fits the underlying scale and 2) enumerating all the metrics online is impractical.

The average set intersection (ASI). ASI is a popular protocol for evaluating the similarity of two ranking list. Given two ranking list $\mathbf{A} = \{a_1, a_2, \cdots, a_N\}$ and $\mathbf{B} = \{b_1, b_2, \cdots, b_N\}$, the set intersection at depth $\mathbf{k}$ is defined as:

$$\mathbf{SI(k)} = \frac{|\{a_1, a_2, \cdots, a_k\} \cap \{b_1, b_2, \cdots, b_k\}|}{k}, \quad (4)$$

in which $|\bullet|$ denotes the operation of counting the number of a set.

ASI averages SI at random depths by:

$$\mathbf{ASI} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{SI(i)} \quad (5)$$

In DyML, we use the ground truth ranking list and the predicted ranking list to calculate ASI. ASI naturally takes all the semantic scales into account.

5. Methods

5.1. Multi-scale learning Baseline

Basically, we use a deep model (backboned on ResNet-34 [8]) to map the raw input images into a feature space. Given the deep features, we first enforce a independent supervision through a specified loss function (*e.g.*, the softmax loss, Cosface [31], Circle Loss [22], Triplet loss [20], N-pair loss [21], Multi-Simi loss [33]). Then we sum up the losses on all the semantic scales in the multi-task learning manner. The multi-scale learning baseline has the following characteristics:

• First, it is superior to the single-scaled metric learning w.r.t. to the overall accuracy. Since it combines the supervisions under multiple semantic scales, the improvement on the overall accuracy is natural. The details are to be accessed in Section 6.1.

• Second, it is confronted with a mutual conflict among different scales. To illustrate this point, let us assume two

[1] https://github.com/MalongTech/imaterialist-product-2019

samples $\mathbf{x_1}$ and $\mathbf{x_2}$ with $\mathbf{l_1^i \equiv l_2^i}$ and $\mathbf{l_1^{i+1} = l_2^{i+1}}$. In another word, under the small scale $\mathbf{S^i}$, they belong to two different classes, while under the larger scale $\mathbf{S^{i+1}}$, they belong to a same class. Under $\mathbf{S^i}$, the baseline is to push $\mathbf{x_1}$ and $\mathbf{x_2}$ far away. In contrast, under $\mathbf{S^{i+1}}$, the baseline is to pull them close. These two contrary optimization objectives compromise each other. In Section 6.4, we experimentally validate the above-described mutual conflict.

5.2. Our Method: Cross-Scale Learning

In response to the mutual conflict, we propose Cross-Scale Learning (CSL). Under $\mathbf{S^i}$, we denote the within-class similarity as $\mathbf{s_p^i}$, and the between-class similarity as $\mathbf{s_n^i}$. CSL uses the within-class similarity under the smallest scale (*i.e.*, $\mathbf{s_p^1}$) as the unique reference to contrast the between-class similarities under all the scales (*i.e.*, $\mathbf{s_n^i, i = 1, 2, \cdots, M}$), yielding the so-called Cross-Scale Learning. Formally, CSL desires:

$$\mathbf{s_p^1 \geq s_n^i + m^i, \qquad i = 1, 2, \cdots, M} \tag{6}$$

in which $\mathbf{m^i}$ is the similarity margin under $\mathbf{S^i}$. Intuitively, we set $\mathbf{m^1 < m^2 < \cdots < m^M}$ in accordance to the increasing scope of $\mathbf{S^1, S^2, \cdots, S^M}$.

Eq. 6 enables a joint optimization across all the semantic scales, with $\mathbf{S^1}$ as the reference scale. The advantages of CSL are two-fold. First, it does not enforce explicit within-class compactness under $\mathbf{S^i (i > 1)}$, and thus avoids the mutual conflict between different scales. Second, using $\mathbf{S^1}$ as the shared reference scale (for optimizing the between-class similarities under all the scales) makes the learning more stable. In reality, a ruler has all its markings annotated with distances from the "0" point.

The loss function for Cross-Scale Learning is correspondingly defined as:

$$\mathbf{L_{CSL} = \sum_{i=1}^{M} \log\Big(1 + \sum_{k=1}^{C^i} \exp\alpha(s_{n,k}^i - s_p^1 + m^i)\Big),} \tag{7}$$

in which $\boldsymbol{\alpha}$ is a scaling factor, $\mathbf{C^i}$ is the total number of training classes under $\mathbf{S^i}$, $\mathbf{s_{n,k}^i}$ is the $\mathbf{k}$-th between-class similarity under $\mathbf{S^i}$.

Besides the cross-scale optimization, CSL further eliminates the mutual suppression with a *proxy-sharing strategy*. Basically, CSL adopts the classification-based training manner: The within-class similarity $\mathbf{s_p}$ is calculated as the cosine similarity between feature $\mathbf{x}$ and the weight vector of its target class. Meanwhile, a between-class similarity $\mathbf{s_n}$ is calculated as the cosine similarity between feature $\mathbf{x}$ and a corresponding weight vector of a non-target class. In CLS, only the classes under $\mathbf{S^1}$ has an independent weight vector as the class prototype. A high-level class under the higher semantic scale $\mathbf{S^i (i > 1)}$ uses the set of weight vectors consisted of its sub-classes in $\mathbf{S^1}$ as the prototype. It thus allows a high-level class to have relatively large within-class diversity [18]. To measure the between-class similarity between $\mathbf{x}$ and a high-level class, CSL has to compare $\mathbf{x}$ against a set of weight vectors. To this end, we follow the common practice of hard mining strategy, *i.e.*, choosing the hardest (closest) negative weight vector against $\mathbf{x}$ to represent the whole vector set.

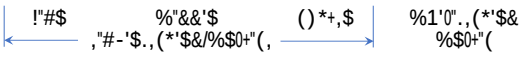

Figure 4. Comparison between single-scaled metrics and a multi-scaled metric baseline on DyML-Animal. The multi-scaled metric surpasses all the single-scaled metrics on the overall accuracy.

6. Experiments

The experiments are arranged as follows. Section 6.1 first experimentally validates that a single-scaled metric does not fit novel semantic scales. Section 6.2 benchmarks the DyML datasets with popular metric learning methods. Section 6.3 evaluates the proposed method of Cross-Scale Learning. Section 6.4 experimentally validates that CSL alleviates the mutual conflict among different scales.

6.1. Analysis on Single-scaled Metric

We investigate the generalization capacity of single-scaled metrics on DyML-Vehicle. Specifically, we learn a deep metric using only the coarse-level, middle-level and the fine-level labels, respectively. We compare them against the multi-scale learning baseline in Fig. 4. For fair comparison, we use a same loss function, *i.e.*, Cosface [31] under all the settings. We draw three observations as follows:

First, each single-scaled metric shows relatively high accuracy under its dedicated scales. For example, the "fine" metric (*i.e.*, the metric learned with fine-level labels) achieves 54.7% Rank-1 accuracy under the fine-level testing. Second, a single-scaled metric does not naturally generalize well to another scale. For example, under the coarse-level testing, the "fine" metric only achieves 26.2% Rank-1, which is lower than the "coarse" metric by **−55.4%**. It validates that a single-scaled metric does not fit a novel scale. Third, with consideration of the overall performance, we find that multi-scale training performs the best. Under the

	DyML-Vehicle					DyML-Animal					DyML-Product				
	ASI	mAP	R@1	R@10	R@20	ASI	mAP	R@1	R@10	R@20	ASI	mAP	R@1	R@10	R@20
Triplet Loss	18.3	10.0	13.8	52.6	65.1	19.3	11.0	18.2	55.5	66.3	9.2	9.3	11.2	43.6	53.3
MS Loss	19.7	10.4	17.4	56.0	67.9	19.9	11.6	16.7	53.5	64.8	9.8	10.0	12.7	45.7	56.4
N-Pair Loss	19.4	10.5	16.4	55.7	68.1	**45.7**	30.3	39.6	69.6	78.8	15.7	15.3	20.3	55.5	65.6
Softmax Loss	22.7	12.0	22.9	61.6	72.9	35.9	25.8	49.6	81.7	88.8	26.8	26.1	50.2	81.6	87.7
Cosface Loss	23.0	12.0	22.9	62.1	73.4	39.6	28.4	45.1	75.7	83.3	25.5	25.0	49.3	81.3	87.7
Circle Loss	22.8	12.1	23.5	62.0	73.3	44.5	30.6	41.5	72.2	80.3	15.8	15.0	26.7	61.5	70.3
CSL	**23.0**	**12.1**	**25.2**	**64.2**	**75.0**	45.2	**31.0**	**52.3**	**81.7**	**88.3**	**29.0**	**28.7**	**54.3**	**83.1**	**89.4**

Table 2. Evaluation of six popular deep learning methods and the proposed Cross-Scale Learning (CSL) on DyML-Vehicle, DyML-Animal and DyML-Product. For CMC and mAP, we report the overall results averaged from three scales. The ASI is an overall evaluation protocol in its nature. Best results are in **bold**.

overall evaluation, it surpasses the "fine" metric, the "middle" metric and the "coarse" metric by +7.2%, + 10.4%, + 19.9%, respectively. It implies that multi-scale training benefits from the multi-scale information and thus improves the overall performance.

Since the objective of DyML is the discriminative ability under all the semantic scales, we use multi-scale supervision for all the evaluated baselines to promote the overall performance.

6.2. Methods Evaluation

We benchmark all the DyML-datasets with 6 popular deep metric learning methods, including 3 pair-based methods (*i.e.*, the Triplet loss [20], N-pair loss [21], Multi-Simi loss [33]) and 3 classification-based methods (*i.e.*, the Softmax Loss, Cosface [31] and Circle Loss [22]). For each method, we use multi-scale supervision for training. For evaluation with CMC and mAP, we average the performance under all the scales and only report the overall performance. The results are reported in Table 2, from which we draw three observations.

First, DyML is very challenging. The overall performance is low under all the three baselines. For example, the "Cosface loss" only achieves 23.0%, 39.6%, 25.5% ASI and 12.0%, 30.7% and 25.0% mAP on DyML-Vehicle, DyML-Animal and DyML-Product, respectively. Second, the classification-based methods generally surpasses the pair-based methods, indicating that the classification training manner usually achieves higher discriminative ability. It is consistent with the observation in many other metric learning tasks [30, 27, 34]. We infer that in spite of the fundamental difference of dynamic range, DyML shares a lot of common properties with the canonical deep metric learning. Third, comparing three classification-based methods (*i.e.*, softmax, Cosface and Circle Loss) against each other, we find that they all achieve very close performance. Though Cosface and Circle Loss marginally surpasses the softmax loss in canonical deep metric learning tasks [30, 27, 22], they do NOT exhibit obvious superiority for DyML. One potential reason is that Cosface and Circle Loss has more hyper-parameters (*i.e.*, the scale and the margin) for each semantic scale. DyML has multiple semantic scales and thus make the optimization of these hyper-parameters more difficult.

Method	Scale	DyML-Animal		DyML-Product	
		mAP	R@1	mAP	R@1
Cosface	Fine	8.7	18.3	11.1	20.3
	Middle	28.4	46.6	16.9	47.6
	Coarse	48.2	70.5	47.1	80.0
	Overall	28.4	45.1	27.8	49.3
CSL	Fine	10.3	25.3	15.6	26.2
	Middle	30.1	53.9	20.1	53.2
	Coarse	52.7	77.7	50.4	83.7
	Overall	31.0	52.3	28.7	54.3

Table 3. Comparison between Cosface and the proposed CSL in three specified scales (besides the overall performance). We report mAP and Rank-1 accuracy. CSL exhibits consistent improvement under all the scales.

6.3. Effectiveness of Cross-Scale Learning

We compare the proposed method, *i.e.*, Cross-Scale Learning with all the six existing methods in Table 2. It clearly shows that CSL is superior to the competing methods w.r.t. the overall performance. For example, CSL surpasses "Cosface" by +2.3%, +7.2% and +5.0% Rank-1 accuracy on DyML-Vehicle, DyML-Animal and DyML-Product, respectively. To be more concrete, Table 3 compares CSL against Cosface under each (fine, middle and coarse) semantic scale, respectively. We observe that CSL achieves improvement not only on the overall (averaged) accuracy, but also on every single scale level. It indicates that CSL does not have a bias towards certain specified semantic scales. Instead, it generally improves the discriminative ability of the learned deep metric under (almost) all the scales, indicating better generalization across multiple scales.

6.4. Reasons for the Superiority of CSL

We investigate the mutual conflict in multi-scale learning baseline, as well as the reason for the superiority of CSL.

Mutual conflict between different scales. During the multi-scale training, we record both the positive similar-

Figure 5. The distribution of similarity scores indicates conflicts between different scales.

ity scores and the negative similarity scores under three scales in Fig. 5. Overall, the similarity scores of "fine positive", "middle postive", "coarse positive" and "negative" pairs are naturally sorted in a descending order. It is because, some negative pairs under the fine scale are actually positive pairs under the middle / coarse scale. When the baseline enforces between-class discrepancy under the fine scale, it tries to decreases their similarity scores, which consequentially decreases the positive similarity scores under the middle / coarse scale. In a word, in the multi-scale learning baseline, the between-class discrepancy in the fine scale compromises the within-class compactness in the middle / coarse scales, and vice versa. We thus conclude that the mutual suppression between different scales hinders the multi-task learning baselines.

CSL alleviates the mutual conflict. We compare the training process of softmax baseline, Cosface baseline and CSL in Fig. 6. Specifically, we record the classification accuracy under all the three semantic scales and make the following three observations:

First, both the softmax and the cosface achieves relatively low classification accuracy under the fine scale. After convergence, they achieve 80.3% and 78.4% classification accuracy under the fine scale, respectively. It further evidences the mutual conflict phenomenon. Second, comparing CSL against softmax and cosface, we find that CSL facilitates faster and better convergence, especially under the fine scale. After convergence, CSL achieves 84.1% classification accuracy under the fine scale. It validates that CSL alleviates the mutual conflict. Third, comparing the training process of a same method, we find that in DyML, the convergence under the fine scale is hard to achieve, while the convergence under the middle and raw scales are relatively easier. It is because the fine-grained visual concepts are inherently harder to recognize.

Compatibility to both classification-based and pair-based training manner. We note that CSL is compatible to both the classification-based and pair-based training manner. In Table 2, CSL adopts the classification training manner. We further compare three training manners, *i.e.*, the classification-based training, pair-based training and the joint training for CSL. The results are shown in Table 4. We observe that CSL trained through classification surpasses its pair-based counterpart, and the joint training further brings incremental improvement. Considering that joint training doubles the hyper-parameters and the improvement is slight, we recommend classification training for CSL.

Figure 6. Comparison between the training process of softmax, cosface and CSL on DyML-Animal. Compared with the multi-scale learning baselines, CSL obtains faster convergence and higher classification accuracy on the training set, because it alleviates the mutual conflict between different semantic scales.

	ASI	mAP	R@1
CSL (Pair)	20.2	10.9	18.2
CSL (Cls)	23.0	12.1	25.2
CSL (Cls+Pair)	23.7	12.6	26.1

Table 4. Comparison between different training manners for CSL. We present the overall accuracy on DyML-Vehicle.

7. Conclusion

In this paper, we introduce the concept of "dynamic range" from real-world metric tools to deep metric for visual recognition. It endows a single metric with scalability to accommodate multiple semantic scales. Based on dynamic range, we propose a new task named Dynamic Metric Learning, construct three datasets (DyML-Vehicle, DyML-Animal and DyML-Product), benchmark these datasets with popular metric learning methods, and design a novel method.

Acknowledgement

This research was supported by China's "scientific and technological innovation 2030 - major projects" (No. 2020AAA0104400).

References

[1] Yalong Bai, Yuxiang Chen, Wei Yu, Linfang Wang, and Wei Zhang. Products-10k: A large-scale product recognition dataset. *arXiv preprint arXiv:2008.10545*, 2020. 2

[2] Thomas Berg, Jiongxin Liu, Seung Woo Lee, Michelle L Alexander, David W Jacobs, and Peter N Belhumeur. Birdsnap: Large-scale fine-grained visual categorization of birds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2011–2018, 2014. 2

[3] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 67–74. IEEE, 2018. 2

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

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

[6] Eran Goldman, Roei Herzig, Aviv Eisenschtat, Jacob Goldberger, and Tal Hassner. Precise detection in densely packed scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5227–5236, 2019. 2

[7] Bing He, Jia Li, Yifan Zhao, and Yonghong Tian. Part-regularized near-duplicate vehicle re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3997–4005, 2019. 2

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[9] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC)*, volume 2, 2011. 2

[10] 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. 2

[11] Hongye Liu, Yonghong Tian, Yaowei Wang, Lu Pang, and Tiejun Huang. Deep relative distance learning: Tell the difference between similar vehicles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2167–2175, 2016. 4

[12] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017. 2

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

[14] Y. Lou, Y. Bai, J. Liu, S. Wang, and L. Duan. Veri-wild: A large dataset and a new method for vehicle re-identification in the wild. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3230–3238, 2019. 4

[15] Yihang Lou, Yan Bai, Jun Liu, Shiqi Wang, and Ling-Yu Duan. Embedding adversarial learning for vehicle re-identification. *IEEE Transactions on Image Processing*, 28(8):3794–3807, 2019. 2

[16] 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

[17] Jingtian Peng, Chang Xiao, Xun Wei, and Yifan Li. Rp2k: A large-scale retail product dataset forfine-grained image classification. *arXiv preprint arXiv:2006.12634*, 2020. 2

[18] Qi Qian, Lei Shang, Baigui Sun, Juhua Hu, Hao Li, and Rong Jin. Softtriple loss: Deep metric learning without triplet sampling. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 6

[19] Rajeev Ranjan, Carlos D Castillo, and Rama Chellappa. L2-constrained softmax loss for discriminative face verification. *arXiv preprint arXiv:1703.09507*, 2017. 2

[20] 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. 2, 5, 7

[21] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *NIPS*, 2016. 2, 5, 7

[22] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. *arXiv preprint arXiv:2002.10857*, 2020. 2, 5, 7

[23] Yifan Sun, Qin Xu, Yali Li, Chi Zhang, Yikang Li, Shengjin Wang, and Jian Sun. Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 393–402, 2019. 2

[24] Y. Sun, L. Zheng, W. Deng, and S. Wang. Svdnet for pedestrian retrieval. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3820–3828, 2017. 2

[25] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *The European Conference on Computer Vision (ECCV)*, September 2018. 2

[26] Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. *arXiv preprint arXiv:2002.08347*, 2020. 2

[27] Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu. Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25(7):926–930, 2018. 2, 7

[28] Feng Wang, Xiang Xiang, Jian Cheng, and Alan Loddon Yuille. Normface: L2 hypersphere embedding for face verification. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1041–1049. ACM, 2017. 2

[29] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. *2018 ACM Multimedia Conference on Multimedia Conference - MM '18*, 2018. 2

[30] 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. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5265–5274, 2018. 2, 7

[31] 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. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5265–5274, 2018. 2, 5, 6, 7

[32] Xiaogang Wang, Gianfranco Doretto, Thomas Sebastian, Jens Rittscher, and Peter H. Tu. Shape and appearance context modeling. In *IEEE 11th International Conference on Computer Vision, ICCV 2007, Rio de Janeiro, Brazil, October 14-20, 2007*, 2007. 5

[33] 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 Conference on Computer Vision and Pattern Recognition*, pages 5022–5030, 2019. 2, 5, 7

[34] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pages 499–515. Springer, 2016. 2, 7

[35] Andrew Zhai and Hao-Yu Wu. Classification is a strong baseline for deep metric learning. *arXiv preprint arXiv:1811.12649*, 2018. 2

[36] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015. 5

[37] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[38] Yi Zhou and Ling Shao. Vehicle re-identification by adversarial bi-directional lstm network. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 653–662. IEEE, 2018. 2

[39] Jianqing Zhu, Huanqiang Zeng, Jingchang Huang, Shengcai Liao, Zhen Lei, Canhui Cai, and Lixin Zheng. Vehicle re-identification using quadruple directional deep learning features. *IEEE Transactions on Intelligent Transportation Systems*, 2019. 2