SDGMNet: Statistic-Based Dynamic Gradient Modulation for Local Descriptor Learning

Yuxin Deng, Jiayi Ma*

Electronic Information School, Wuhan University, Wuhan 430072, China
dyx_acuo@whu.edu.cn, jyma2010@gmail.com

Abstract

Rescaling the backpropagated gradient of contrastive loss has made significant progress in descriptor learning. However, current gradient modulation strategies have no regard for the varying distribution of global gradients, so they would suffer from changes in training phases or datasets. In this paper, we propose a dynamic gradient modulation, named SDGMNet, for contrastive local descriptor learning. The core of our method is formulating modulation functions with dynamically estimated statistical characteristics. Firstly, we introduce angle for distance measure after deep analysis on backpropagation of pair-wise loss. On this basis, auto-focus modulation is employed to moderate the impact of statistically uncommon individual pairs in stochastic gradient descent optimization; probabilistic margin cuts off the gradients of proportional triplets that have achieved enough optimization; power adjustment balances the total weights of negative pairs and positive pairs. Extensive experiments demonstrate that our novel descriptor surpasses previous state-of-the-art methods in several tasks including patch verification, retrieval, pose estimation, and 3D reconstruction.

Introduction

Feature extraction is a fundamental problem in many computer vision tasks, such as image classification, matching, and retrieval. In the image matching pipeline, features are firstly extracted and then matched for downstream applications (Fan et al. 2022). Undoubtedly, the quality of the feature determines the upper limit of the pipeline (Fan et al. 2019). Moreover, the independent studies on feature extraction (Gleize, Wang, and Feiszli 2023; Wang et al. 2023), especially those involving feature descriptions only (Tian et al. 2020; Wang, Zhang, and Huang 2022, 2023) are more compatible with mature pipelines and applicable to more tasks. Thus, we think feature description deserves further study.

Benefiting from the great potentials of Deep Neural Networks (DNN), deep feature description dispenses with heuristic designs to acquire transform or illumination invariance as early efforts (Lowe 2004) did. Overall, descriptor learning is exactly a branch of contrastive learning (Cui et al. 2023). Specifically, this task aims to encode images or local patches into descriptors, and then predict whether pairs of images belong to the same category or not according to distances between descriptors. To train the encoder, we need to minimize the distance of correspondence/positive pairs and maximize non-correspondences/negative ones in the loss function. To this end, various pair-wise losses are used, such as triplet loss (Mishchuk et al. 2017; Xue, Budvytis, and Cipolla 2023), negative cross entropy loss (Tian, Fan, and Wu 2017; Gleize, Wang, and Feiszli 2023), and ranking loss (He, Lu, and Sclaroff 2018).

So, what is the principle for advanced loss design? Hard example mining (HEM) is one basic principle. Specifically, a positive pair of descriptors would be a hard example, if the distance between the two descriptors is too large. To optimize the hard positive example, its back-propagated gradients from the loss should be weighted as shown in Fig. 1. In contrast, the magnitude of the gradients *i.e.*, the weights for a hard negative pair that is closer should be larger. Moreover, the hardness of Siamese pairs that share the same anchor can be measured by relative distance and also deserves attention. Those hard Siamese pairs should be emphasized with weights increasing with the relative distance. Many efforts have been made to modulate the weights of back-propagated gradients in the field of general contrastive learning (Huang et al. 2020; Boutros et al. 2022; Zhou et al. 2023). Their successes demonstrate the significance of gradient modulation. However, most modulations are static. The values of the modulation functions depend on the distance of a few sampled pairs, but do not involve the training phase or the global information. Such modulations might suffer from the changes in training phases and datasets. Thus, making the gradients adapt to global statistics, which vary over time and datasets, is reasonable for learning better descriptors.

In this paper, we propose SDGMNet, a statistic-based dynamic gradient modulation strategy for contrastive local descriptor learning. Firstly, we analyze the back-propagated gradient of pair-wise loss, and explore that angular distance provides flattening magnitude of gradients before modulation. SDGMNet is easily implemented in pseudo loss composite of weighted included angles. Secondly, we propose auto-focus modulation to modulate gradients for individual pairs. Auto-focus modulation utilizes the statistics of distances between individual pairs. Rather than following strict HEM, it mines reliable pairs whose distances lay around

*Corresponding author

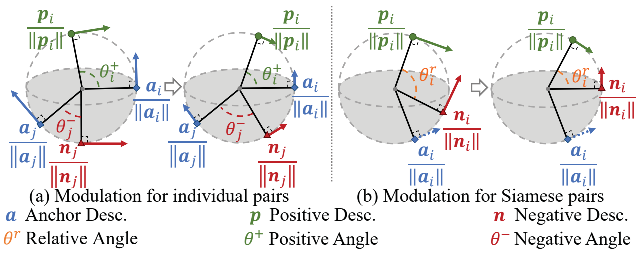


Figure 1: Illustration of the gradient modulation for HEM. Arrows denote the gradients of normalized descriptors during training. Relative angle θ_i^r is equal to $\theta_i^+ - \theta_i^-$. (a) The magnitude of the gradient of an individual corresponding pair $\{\boldsymbol{a}, \boldsymbol{p}\}$ should increase with θ^+. In contrast, the one of a non-corresponding pair $\{\boldsymbol{a}, \boldsymbol{n}\}$ should decrease with θ^-. (b) Siamese pair, *i.e.*, a triplet $\{\boldsymbol{a}, \boldsymbol{p}, \boldsymbol{n}\}$ in SDGMNet desires a smaller weight, when relative angle θ^r is diminishing.

the location of the distribution to orient the optimization. Thirdly, probabilistic margin employs statistics of the relative distance of Siamese pairs, *i.e.*, triplets. It is applied to cut off the gradients of proportional Siamese pairs that are believed to reach the optimums. Meanwhile, the novel margin emphasizes harder examples with increasing weights. Finally, we adjust the ratio of positive and negative total weights with weight normalization and attenuation coefficient. All statistics are estimated dynamically with rough Bayesian sequential update (Bishop 2006). Extensive experiments in five tasks, including patch verification, matching, retrieval, pose estimation, and 3D reconstruction, confirm the superiority of the descriptors learned by SDGMNet.

Our contributions can be summarized as follows:

1) We explore the special characteristic of angular distance in backpropagation, which can provide a theoretical unbiased modulation for elaborated modifications.
2) We propose statistics-based auto-focus modulation to moderate the adverse impacts of the extremely hard individual pairs so that the training can converge more stably.
3) We propose probabilistic margin, which combine CDF-based soft and hard margins, to further optimize those hard Siamese pairs in a more explainable way.
4) We propose power adjustment to rebalance total weights of negative and positive pairs for better generalization.

Related Works

Gradient Modulation for Contrastive Loss

Modulating gradients has been a theme of designing contrastive loss for some time. Modulation strategies can be categorized into two classes: Modulation for individual pairs and ones for Siamese pairs as drawn in Fig. 1. Let $\mathcal{L}$ denote a general loss, $d^+(\boldsymbol{a}, \boldsymbol{p})$ denote a general distance between a correspondence $\{\boldsymbol{a}, \boldsymbol{p}\}$, and $d^-(\boldsymbol{a}, \boldsymbol{n})$ denote the one between a non-correspondence $\{\boldsymbol{a}, \boldsymbol{n}\}$. $d^+ - d^-$ is referred to relative distance, denoted by d^r.

Modulation for Individual Pairs. Following HEM, $\partial\mathcal{L}/\partial d^+$ should be modulated with an increasing function *w.r.t.* d^+, while $-\partial\mathcal{L}/\partial d^-$ needs a decreasing one. In recent years, Circle Loss (Sun et al. 2020) satisfies respective demands of d^+ and d^- with circle margin. SFace (Zhong et al. 2021) employs sigmoid functions to mine hard pairs. For local descriptor learning, Scale-aware Loss (Keller et al. 2018) modulate $\partial\mathcal{L}/\partial d_i^+$ and $-\partial\mathcal{L}/\partial d_i^-$ with functions symmetrical about $(d_i^+ + d_i^-)/2$ for a triplet. Exp-TL (Wang et al. 2019a) conducts more strict HEM with exponential loss. HyNet (Tian et al. 2020) also observes a hidden modulation in deep backpropagation. It replaces common similarity with hybrid similarity, whose gradient can balance the needs of two kinds of pairs.

Modulation for Siamese Pairs. The relative hardness of Siamese pairs should be also considered in HEM. For triplets, the harder ones with larger d^r should be emphasized with larger $\partial\mathcal{L}/\partial d^r$. Balntas *et al.* (Balntas et al. 2016) introduce a static hard margin to prevent easy triplets from descriptor learning. Quadratic triplet loss (Tian et al. 2019) and Scale-aware Loss (Keller et al. 2018) modulate the gradients with continuous elementary functions that monotonically increase with d^r. CDFDesc (Zhang and Rusinkiewicz 2019) further enrolls cumulative distribution function (CDF) for dynamic modulation. In other tasks, angular margin (Deng et al. 2019; Boutros et al. 2022; Zhou et al. 2023) cuts off easy Siamese pairs and improves face recognition performance. Moreover, Multi-Simi Loss (Wang et al. 2019b) separates Siamese pairs into positive and negative groups, and gradients of samples in each group are modulated independently. In contrast, Circle Loss (Sun et al. 2020) considers two kinds of Siamese pairs together.

As discussed above, few works regard a systematic and dynamic solution adaptive to training steps or datasets. Therefore, in this work we focus on designing such a strategy for descriptor learning.

Feature-based Image Matching

Image matching is one of the major downstream applications of descriptor learning. Feature-based image matching firstly extracts features on images, and then matches two feature sets via special matching methods, *e.g.*, nearest neighbor matching, to acquire correspondences. Feature extraction methods can be classified into two categories: detecting-then-describing and detecting-and-describing. Detecting-then-describing methods (Lowe 2004; Mishchuk et al. 2017; Wang, Zhang, and Huang 2023) require off-the-shelf detectors (Barroso-Laguna et al. 2019; Lee, Kim, and Cho 2022) to extract keypoints, and then encode the patches centered at detected keypoints. Detecting-and-describing methods (Tyszkiewicz, Fua, and Trulls 2020; Zhao et al. 2022; Wang et al. 2023) employ a network to regress a dense descriptor map and a detected score map, and sample sparse descriptors according to the score map. Moreover, deep feature matching catches increasing attention in current years. SuperGlue (Sarlin et al. 2020) matches the features with both descriptor and keypoint information, which boosts the recall of inliers and the successful rate of matching performance. LOFTR (Sun et al. 2021) eliminates the detector and directly matches descriptors of all pixels, which leads the recent trend. Obviously, descriptors are indispensable for any image matching method, so we think descriptor learning still merits independent study. Moreover, those descriptors can be applied to not only image matching, but also image retrieval, patch similarity verification, and so on.

Methodology

Descriptor learning aims to encode a patch $\boldsymbol{I}$ centered at a detected keypoint into a descriptor $\boldsymbol{x}(\boldsymbol{\Omega}, \boldsymbol{I})$ with an encoder parameterized by $\boldsymbol{\Omega}$. We omit $(\boldsymbol{\Omega}, \boldsymbol{I})$ in subsequent analysis for clarity. $\boldsymbol{a}$, $\boldsymbol{p}$ and $\boldsymbol{n}$ are instances of $\boldsymbol{x}$. To train the encoder, we should construct a loss function $\mathcal{L}$, in which the positive distance d^+ between a corresponding $\{\boldsymbol{a}, \boldsymbol{p}\}$ should be minimized, while the negative distance d^- between $\{\boldsymbol{a}, \boldsymbol{n}\}$ should be maximized.

As discussed above, a better loss function should give each d a proper weight for better optimization. For example, following HEM, the weight for $\partial d^+/\partial\boldsymbol{\Omega}$ should be large, if d^+ is large. Those weights would reflect in $\partial\mathcal{L}/\partial d$ during backpropagation. And SDGMNet majors in modulating $\partial\mathcal{L}/\partial d$ for better descriptor learning.

Since triplet loss is the most popular loss function for descriptor learning, we take triplet loss as our baseline and the example for theoretical analysis. Let N denote batch size, $\boldsymbol{D}$ denote the pair-wise distance batch $\{d_1^-, d_2^-, \ldots, d_N^-; d_1^+, d_2^+, \ldots, d_N^+\}$. Given a risk function $f(\,\bullet\,)$ and a distance batch $\boldsymbol{D}$, the general triplet loss $\mathcal{L}$ can be represented as

$$\mathcal{L}(\boldsymbol{D}) = f(d_1^-, d_2^-, \ldots, d_N^-; d_1^+, d_2^+, \ldots, d_N^+). \quad (1)$$

The back-propagated gradient of the loss *w.r.t.* the encoder parameters $\boldsymbol{\Omega}$ can be computed by chain rule as

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\boldsymbol{\Omega}} =& \sum_{i=1}^N \frac{\partial\mathcal{L}}{\partial d_i^+}\left(\frac{\partial d_i^+}{\partial\boldsymbol{a}_i}\frac{\partial\boldsymbol{a}_i}{\partial\boldsymbol{\Omega}} + \frac{\partial d_i^+}{\partial\boldsymbol{p}_i}\frac{\partial\boldsymbol{p}_i}{\partial\boldsymbol{\Omega}}\right) + \\ & \sum_{i=1}^N \frac{\partial\mathcal{L}}{\partial d_i^-}\left(\frac{\partial d_i^-}{\partial\boldsymbol{a}_i}\frac{\partial\boldsymbol{a}_i}{\partial\boldsymbol{\Omega}} + \frac{\partial d_i^-}{\partial\boldsymbol{n}_i}\frac{\partial\boldsymbol{n}_i}{\partial\boldsymbol{\Omega}}\right),\end{aligned} \quad (2)$$

where $(\boldsymbol{D})$ is omitted. In Eq. (2), $\partial\mathcal{L}/\partial d$ is a scalar, which reveals how much the corresponding pair contributes to the update of parameters. Gradient modulation focuses on rescaling $\partial\mathcal{L}/\partial d$ with a function about d.

Angular Distance

Consider the term $\partial d/\partial\boldsymbol{x}$. After L_2 normalization, descriptors are embedded onto unit hypersphere. Included angular θ, a distance measure, is defined as

$$\theta(\boldsymbol{x}, \boldsymbol{y}) = \text{acos}\frac{\boldsymbol{x}^T\boldsymbol{y}}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|}, \quad (3)$$

where $\|\bullet\|$ denotes L_2 normalization, $\boldsymbol{y}$ is the other descriptor in a pair. Cosine similarity s and L_2 distance l are more common metrics for measuring a distance:

$$s(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x}^T\boldsymbol{y}}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|}, \qquad l(\boldsymbol{x}, \boldsymbol{y}) = \left\|\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|} - \frac{\boldsymbol{y}}{\|\boldsymbol{y}\|}\right\|. \quad (4)$$

They are equivalent instances of d in forward propagation but distinguishing in backpropagation.

In backpropagation, $\partial\theta/\partial\boldsymbol{x}$, $-\partial s/\partial\boldsymbol{x}$, $\partial l/\partial\boldsymbol{x}$ share the same optimal direction which is orthogonal to $\boldsymbol{x}$ as illustrated in Fig. 1. However, they own special magnitudes as

$$\left\|\frac{\partial\theta}{\partial\boldsymbol{x}}\right\| = \frac{1}{\|\boldsymbol{x}\|}, \quad (5)$$

$$\left\|\frac{\partial s}{\partial\boldsymbol{x}}\right\| = \frac{1}{\|\boldsymbol{x}\|}\sqrt{1 - s^2}, \quad (6)$$

$$\left\|\frac{\partial l}{\partial\boldsymbol{x}}\right\| = \frac{1}{\|\boldsymbol{x}\|}\frac{\sqrt{4 - l^2}}{2}. \quad (7)$$

As shown above, an implicit modulation takes effect after a metric is chosen. The magnitude of $\partial\theta/\partial\boldsymbol{x}$ depends on $\|\boldsymbol{x}\|$ only, which would not disturb the modulation function about θ we design later. Thus, θ is a suitable choice for our intention. And, we would free $1/\|\boldsymbol{x}\|$ for two reasons. Firstly, related works (Ranjan, Castillo, and Chellappa 2017; Salimans and Kingma 2016) observe such natural scales can accelerate the training and better reflect the data variance. Moreover, $\boldsymbol{a}$, $\boldsymbol{p}$ and $\boldsymbol{n}$ are equivalent over training steps.

In short, $\partial\theta/\partial\boldsymbol{x}$ owns the optimal direction and a plain magnitude for learning. Thus, we employ θ for distance measure in SDGMNet. As a result, we can dedicate to modulating gradients of pairs, *i.e.*, formulating $\partial\mathcal{L}/\partial d$ (*i.e.* $\partial\mathcal{L}/\partial\theta$). Eq. (2) can be reformulated with θ into

$$\frac{\partial\mathcal{L}}{\partial\boldsymbol{\Omega}} = \sum_{i=1}^N w_i^+\frac{\partial\theta_i^+}{\partial\boldsymbol{\Omega}} - \sum_{i=1}^N w_i^-\frac{\partial\theta_i^-}{\partial\boldsymbol{\Omega}}, \quad (8)$$

where $w_i^+ \triangleq \partial\mathcal{L}/\partial\theta_i^+$ and $w_i^- \triangleq -\partial\mathcal{L}/\partial\theta_i^-$ are larger than 0. They represent the magnitudes of $\partial\mathcal{L}/\partial\theta$ and the weights of $\partial\theta/\partial\,\boldsymbol{\Omega}$. We decompose w_i into $w_s \times w_c$ in SDGMNet.

Auto-focus Modulation

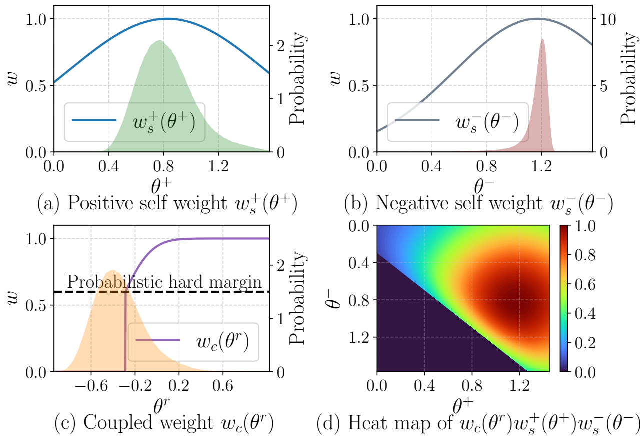
Ideally, θ^+/θ^- reaches its optimum at $0/\pi$. Following HEM, the gradient of θ that is further away from its optimum should be weighted more heavily. In other words, the gradients of positive pairs should be modulated with w_s^+ that is monotonously increasing *w.r.t.* θ^+, and the negative with decreasing w_s^-. However, θ^+ and θ^- might be unreliable. Although hard positive pairs with large θ^+ are validated by ground truth, extreme distortions they carry would damage the global optimization. For the hardest negative pairs of patches, while the real distance between them cannot be evaluated, we should not simply push their descriptors away. Thus, extreme individual pairs should be treated more cautiously. Successes of HyNet (Tian et al. 2020) and SFace (Zhong et al. 2021) also imply that excessive HEM on individual pairs should not be advocated.

To neutralize HEM and extreme individual pairs suppression, we formulate dynamic self weight w_s^+ and w_s^- for individual pairs in SDGMNet as

$$w_s^+(\theta^+) = \exp\left(-\frac{(\theta^+ - \mathrm{E}_t[\theta^+])^2}{2(\pi/6 + \mathrm{Std}_t[\theta^+])^2}\right), \quad (9)$$

$$w_s^-(\theta^-) = \exp\left(-\frac{(\theta^- - \mathrm{E}_t[\theta^-])^2}{2(\pi/6 + \mathrm{Std}_t[\theta^-])^2}\right), \quad (10)$$

where $\mathrm{E}_t[\bullet]$ represents the expectation, $\mathrm{Std}_t[\bullet]$ denotes the standard deviance, and the subscript t means the statistics are dynamically estimated over time. The modulation originates from Gaussian blur. It is referred to auto-focus modulation, because it automatically focuses on the samples near the expectation of distribution, while moderating the



(a) Positive self weight $w_s^+(\theta^+)$
(b) Negative self weight $w_s^-(\theta^-)$
(c) Coupled weight $w_c(\theta^r)$
(d) Heat map of $w_c(\theta^r)w_s^+(\theta^+)w_s^-(\theta^-)$

Figure 2: Visualization of modulation functions and related data distributions at the last training epoch on *Liberty*. Curves in (a), (b) and (c) illustrate three kinds of weights introduced in the text. Shadows denote the probability distribution of variables. (d) is a heat map of $w_s^+(\theta^+)w_s^-(\theta^-)w_c(\theta^r)$ with θ^+ and θ^- as x-axis and y-axis. The dark red in (d) indicates a strong impact on optimization. Our formulation does not turn the spotlight on hardest triplets that should lay at the top right. Easy triplets at the bottom left are eliminated by the hard margin.

impacts of harder and easier examples as illustrated in Fig. 2. We limit the lower bound of the blur radius to $\pi/2$, *i.e.*, add $\pi/6$ to the standard deviance, because the angle of positive pairs and negative pairs mainly spread on $[0, \pi/2]$ (Zhang et al. 2017). If no such constraint, the long-tailed hardest examples will be cleaned out with extremely small weight.

Probabilistic Margin

The optimization does not converge until $\partial\mathcal{L}/\partial\boldsymbol{\Omega}$ approaches zero. Thus, margins are necessary to force w to be zeros near the optimums. For example, θ_i^+ holds an ideal optimum at 0, so the $w_i^+|_{\theta_i<0+m_i}$ should be 0, where the exclusive margin m_i can be an infinitesimal. However, searching for m_i for each sampled pair is infeasible. Furthermore, an unified margin for all triplets proves more effective in the practice, which means a single margin m designed as a function of θ^r is favored. The function modulates Siamese pairs with the same weight and we name it coupled weight w_c.

How large m should be set for the global optimum is still unclear. Instead of employing a fixed empirical m, we elaborate a hard margin $m \in [0, 1]$ that believes $100 \times m\%$ examples have reached the optimum, so-called probabilistic hard margin (PHM). Given a probabilistic hard margin m, we combine probabilistic soft margin (Zhang and Rusinkiewicz 2019) and hard margin to formulate w_c in SDGMNet as

$$w_c(\theta^r) = \begin{cases} \mathrm{CDF}_t(\theta^r), & \theta^r > \mathrm{iCDF}_t(m), \\ 0, & \theta^r \leq \mathrm{iCDF}_t(m), \end{cases} \tag{11}$$

where CDF($\boldsymbol{\cdot}$) denotes cumulative distribution function, iCDF($\boldsymbol{\cdot}$) denotes inverse CDF. Triplets that carry θ^r smaller than iCDF(m) are top $100 \times m\%$ easy examples empirically. These easy examples are believed to approach the optimum and will be isolated from further optimization. The others are preserved and weighted by a monotonously increasing CDF for HEM. Due to probabilistic hardness, the modulation is dynamic and adaptive to training data and stage. To facilitate the implementation, we approximate the data distribution with a normal distribution. The curve of $w_c(\theta^r)$ is drawn in Fig. 2 (c), where we set $m = 0.6$.

Power Adjustment

Power is defined as the total weight of a class of pairs:

$$P^+ = \textstyle\sum_{i=1}^N w_i^+, \qquad P^- = \sum_{i=1}^N w_i^-. \tag{12}$$

Power describes how strongly a class of positive or negative pairs guides the training with gradients. Before modulation, *i.e.*, $w = 1$, the positive power P^+ and the negative power P^- hold balanced. However, a bias has been introduced once the scale factors, *i.e.*, $1/\sqrt{2\pi}\sigma$, are dropped in auto-focus modulation. Rather than determining proper scale factors, we define power which involves both individual and Siamese modulation to reconsider the bias. Intuitively, the positive power guides the model to identify the images with the same label. In contrast, the negative power forces the model to discriminate negative examples. An inductive bias on the negative power P^- might be preferred, because the model does not need to identify all labels well which would not appear in the test. Moreover, discriminating ability can promote identifying for both human beings and machine learning. To adjust the ratio of power, we propose weight normalization that divides the weights by the expectation of the power. Then, attenuation is adopted on the positive side. Finally, SDGMNet is finished as:

$$\frac{\partial\mathcal{L}}{\partial\boldsymbol{\Omega}} = \sum_{i=1}^N \frac{\alpha w_i^+}{E_t[P^+]}\frac{\partial\theta_i^+}{\partial\boldsymbol{\Omega}} - \sum_{i=1}^N \frac{w_i^-}{E_t[P^-]}\frac{\partial\theta_i^-}{\partial\boldsymbol{\Omega}}, \tag{13}$$

where α is the attenuation coefficient. Once normalization is activated, the balance of powers can be quantified and adjusted by the attenuation coefficient. Such a ratio can adapt to random data, arbitrary modulations, running training phases and finally benefit the training.

Implementation Details

Triplet Sampling. We follow HardNet's triplet sampling strategy (Movshovitz-Attias et al. 2017) to construct loss for descriptor learning. Briefly, HardNet follows L2Net (Tian, Fan, and Wu 2017) to sample N corresponding pairs. For a corresponding pair, HardNet mines for the nearest negative neighbor in batch as the negative sample of triplet.

Network Architecture. HyNet (Tian et al. 2020) encode a 32×32 patch into a 128-d descriptor by an encoder equipped with learnable Filter Response Normalization (FRN) and Threshold Logic Unit (TLU) (Singh and Krishnan 2020). We adopt this network architecture and illustrate it in Fig. 3.

Statistics Estimation. There are some statistics in SDGMNet varying over training time. We employ rough Bayesian update (Bishop 2006) to estimate these variables:

$$\boldsymbol{\beta}_t = 0.999\boldsymbol{\beta}_{t-1} + 0.001\boldsymbol{\mu}_t, \tag{14}$$

Figure 3: Network architecture for descriptor learning, where Conv $3 \times 3, 32, /2$ denotes a convolution layer with a kernel size of 3, output channel of 32, and stride of 2; BN denotes Batch Normalization.

Algorithm 1: SDGMNet for local descriptor learning

Input: m and α, initial $\boldsymbol{\beta}_0$, model, dataset, optimizer.
$t = 1$;
while training **do**
  Sample a data batch from datasets;
  Obtain HardNet triplets in batch;
  Update corresponding statistics by Eq. (14);
  Compute weights by Eqs. (9), (10), and (11);
  if warming **then**
    Set $w_s^+(\theta^+)$, $w_s^-(\theta^-)$ and $w_c(\theta^r)$ to 1
  end if
  Compute powers P^+ and P^- by Eq. (12);
  Update the expectation of powers by Eq. (14);
  Construct pseudo loss by Eq. (15);
  Update the model with the optimizer;
  $t = t + 1$;
end while
Output: Well-trained model.

where $\boldsymbol{\beta}_t$ is the vector of approximated global statistics and $\boldsymbol{\mu}_t$ is the estimation in batch at the tth iteration.

Pseudo Loss. The modulated gradient in SDGMNet contains CDF that is a non-elementary function so that we cannot find a simple direct loss that has gradients as Eq. (13) to guide the training. Motivated by general pair weighting framework (Wang et al. 2019b), we define pseudo loss as:

$$\mathcal{L}_{\mathcal{P}} = \frac{\alpha}{E_t[P^+]} \sum_{i=1}^{N} w_i^+ \theta_i^+ - \frac{1}{E_t[P^-]} \sum_{i=1}^{N} w_i^- \theta_i^-, \quad (15)$$

where α, w, $E_t[P^+]$ and $E_t[P^-]$ are all constant with regard to the model parameter $\boldsymbol{\Omega}$. The gradient of pseudo loss is the same as Eq. (13) so it can be used to train SDGMNet.

Training. We train SGMNet on UBC PhotoTour dataset (Winder and Brown 2007) with Algorithm 1. where we set $m = 0.6$ and $\alpha = 0.9$ for the best performance. The network is trained for 200 epochs (200K iterations) with batch size of 1024 and SGD optimizer. Moreover, the training is warmed up with $w = 1$ in the first 10% of iterations. During warming, only $E[P^+]$ and $E[P^-]$ take effect, and all statistics are estimated in every iteration. As a result, only the initial values of $E[P^+]$ and $E[P^-]$ contribute to the full SDGMNet training. We initialize them with 1000.

Embedded into Detecting-and-describing Methods. We embed SDGMNet into existing joint methods, including DISK (Tyszkiewicz, Fua, and Trulls 2020) and ALIKE (Zhao et al. 2022), to investigate its broader impact. The total loss composite of the original one and our SDGM loss is used to finetune the pre-trained models in 10 epochs with the released official codes. The best checkpoint with the highest validation accuracy is chosen for the test.

Train	ND	YOS	LIB	YOS	LIB	ND	Mean
Test	LIB		ND		YOS		
SIFT	29.84		22.53		27.29		26.55
L2Net	2.36	4.70	0.72	1.29	2.57	1.17	2.23
HardNet	1.49	2.51	0.53	0.78	1.96	1.84	1.51
CDFDesc	1.21	2.01	0.39	0.68	1.51	1.29	1.38
SOSNet	1.08	2.12	0.34	0.67	1.03	0.95	1.03
PUDesc	2.91	4.58	0.91	1.21	3.57	2.43	2.60
HyNet	0.89	**1.37**	**0.34**	0.61	0.88	0.96	0.84
SDGMNet	**0.88**	1.41	**0.34**	**0.46**	**0.82**	**0.69**	**0.78**

Table 1: Patch verification performance on UBC PhotoTour. The best FPR@95(%) are highlighted in bold. Dash lines separate different network architectures. LIB: *Liberty*, YOS: *Yosemite*, ND: *Notredame*.

Experiments

We test SDGMNet on five benchmarks: UBC PhotoTour (Winder and Brown 2007), HPatches (Balntas et al. 2017), Image Matching Challenge (Jin et al. 2021), ScanNet (Dai et al. 2017) and ETH 3D reconstruction (Schonberger et al. 2017). The results are compared with SIFT (Lowe 2004), L2Net (Tian, Fan, and Wu 2017), HardNet (Mishchuk et al. 2017), CDFDesc (Zhang and Rusinkiewicz 2019), SOSNet (Tian et al. 2019), HyNet (Tian et al. 2020), PUDesc (Wang, Zhang, and Huang 2022), DISK (Tyszkiewicz, Fua, and Trulls 2020), ALIKE (Zhao et al. 2022) and AWDesc (Wang et al. 2023). All methods output 128-dimensional descriptors that can be evaluated with L_2 distance. Note that, DISK, ALIKE and AWDesc are detecting-and-describing methods, while the others belong to 'then' methods. ALIKE and DISK finetuned with SDGM are suffixed with '+SDGM'. Our codes are available at https://github.com/ACuOoOoO/SDGMNet.

Patch Verification on UBC PhotoTour

UBC PhotoTour (Winder and Brown 2007) is the most widely used dataset for local descriptor learning. It consists of three subsets *Liberty*, *Yosemite* and *Notredame*. Deep descriptors are trained on one subset and tested on the other two. In the standard protocol, the test aims to verify 100K pairs of patches matched or not. We report the false positive rate at 95% recall (FPR@95) of verification results in Table 1. Let A-B represent the result trained on A and then tested on B. SDGMNet outperforms the current state-of-the-art method HyNet on YOS-ND and ND-YOS with certain margins of 0.15 and 0.27, when only on YOS-LIB a small gap exists. The better performance of SDGMNet proves the significance of making the gradient modulation dynamic.

Image Matching and Retrieval on Hpatches

HPatches (Balntas et al. 2017) is a more comprehensive benchmark that evaluates descriptors on three tasks: patch

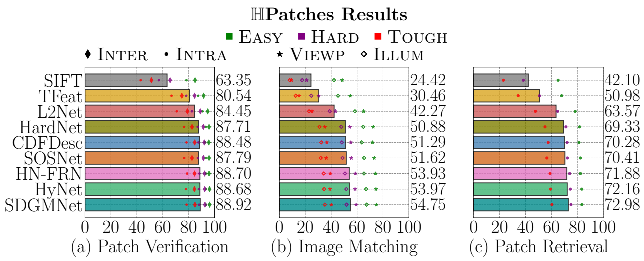


Figure 4: Test on HPatches split 'a'. We report mean average precision (mAP) (%) as evaluation metric. Results of subtasks are marked with different colors and patterns. The bars show the mean scores of subtasks.

verification, image matching and patch retrieval. According to the degree of distortion, subtasks are categorized into *Easy*, *Hard* and *Tough*. Furthermore, patch pairs from the same or different image sequences are separated into two test subsets for verification, denoted by *Intra* and *Inter*, respectively. And the matching task is designed to evaluate the viewpoint (*VIEWP*) and illumination (*ILLUM*) invariance of descriptors. For a fair comparison, all models are trained on *Liberty* of UBC PhotoTour. The results are shown in Fig. 4, where HN-FRN denotes the HyNet's network trained with HardNet loss. As shown, while there is no gap between HN-FRN and HyNet on image matching task, an improvement of 0.82 on HN-FRN is achieved by SDGMNet. Moreover, SDGMNet exceeds HyNet with a certain margin of 0.82 on the patch retrieval task, which suggests that the dynamic modulation can help descriptor learning and improve the applicability of descriptors on various tasks.

Outdoor Pose Estimation

Image Matching Challenge (IMC) (Jin et al. 2021) focuses on the performances of local features on outdoor stereo and multiview (MV) matching. In the standard pipeline of IMC, local descriptors trained on *Liberty* are extracted with DOG detector. Then Fast Library for Approximate Nearest Neighbors (FLANN) is used to match the features for downstream tasks with the optimal ratio test. We use DEGENSAC (Chum, Werner, and Matas 2005) for geometric verification with the recommended settings. Please refer to (Jin et al. 2021) for more details. The performances of camera pose estimation on IMC are shown in Table 2 with mean Average Accuracy to 10° angular error (mAA@10) as the evaluation metric. SDGMNet achieves state-of-the-art performance on all subtasks and finally obtains a gain of about 1% on mean mAA, which demonstrates its practical applicability for outdoor image matching.

Moreover, we compare the learned descriptor, *i.e.*, the methods above the middle line, to the joint detecting-and-describing methods including DISK, ALIKE and AWDesc. As we can see, two-stage methods composite of classic DOG detector and learned descriptors maintain competitive to the joint methods. Additionally, DISK and ALIKE can be boosted by the embedding of our SDGMNet loss, which demonstrates the significance of independent study of descriptor learning and our novel loss function.

#Keypoints	2k		8k		Mean
Task	Stereo	MV	Stereo	MV	
HardNet	54.32	61.64	67.07	74.24	64.32
HyNet	54.71	62.13	67.29	74.78	64.73
SDGMNet	**55.62**	**63.51**	**68.13**	**75.53**	**65.70**
DISK	61.25	81.54	62.97	82.33	72.02
+SDGM	**62.37**	**81.78**	**63.85**	**82.50**	**72.65**
ALIKE	54.31	73.40	56.55	76.25	65.12
+SDGM	56.20	74.21	58.72	76.21	66.33
AWDesc	45.11	65.06	46.66	74.08	57.77

Table 2: Pose estimation on the validation set of *PhotoTourism* in IMC. The best mAA (%) is marked in bold. The middle line separate methods belong to DOG+learned descriptor or joint detection and description methods.

AUC	5°	10°	20°
HardNet	11.68	24.49	39.72
SDGMNet	**11.72**	**25.27**	**41.51**
DISK	6.64	15.50	27.23
+SDGM	8.14	18.92	32.95
ALIKE	6.12	14.88	27.62
+SDGM	7.59	17.60	30.26
AWDesc	**10.77**	**23.59**	**38.88**

Table 3: Relative pose estimation in indoor dataset ScanNet. AUC at different thresholds with 2048 features are reported.

Indoor Pose Estimation

ScanNet (Dai et al. 2017) provides large-scale indoor sequences with ground-truth camera poses and depth images. Those indoor scenes are harder than those scenes in IMC due to the lack of textures. We select 1500 pairs of images for the test. Up to 2048 features are extracted and then matched with FLANN. Finally, RANSAC is employed for geometric verification. Local descriptors are trained on *Liberty*, and joint methods are trained on MegaDepth (Li and Snavely 2018). The area under the cumulative pose error curve (AUC) at different thresholds is reported in Table 3.

Compared to HardNet, our SGMNet brings extract improvements in pose estimation accuracy, especially at 20°. Compared to joint detection and description methods, detecting-then-describing methods obtain better performance on indoor scenes, while all the models are trained on outdoor scenes. It reveals that joint learning methods might overfit the training scene and confirms the applicability of 'then' methods. Moreover, SDGM loss improves the performance of the joint methods. Especially, it boosts DISK at 20° with a large margin of 5.7%, which further supports the significance of our proposals.

ETH 3D Reconstruction

ETH benchmark (Schonberger et al. 2017) shows more interest in how the matching performance affects the more practical 3D reconstruction tasks, *i.e.*, structure-from-motion (SFM) and Multi-View Stereo (MVS) (Schonberger and Frahm 2016; Schönberger et al. 2016). We test the local

Scene	Feature	#Reg. Images (↑)	#Sparse Points (↑)	#Dense Points (↑)
M. M.	HardNet	723	258K	1.08M
	HyNet	807	353K	1.25M
	SDGMNet	**850**	**429K**	**1.35M**
G.	HardNet	1163	795K	3.15M
	HyNet	1171	836K	3.13M
	SDGMNet	**1184**	**841K**	**3.29M**

Table 4: 3D reconstruction on Madrid Metropolis (M. M.) and Gendarmenmarkt (G.) of ETH benchmark. Three crucial metrics from the benchmark are reported.

descriptors trained on Liberty on two subsets of ETH benchmark: Madrid Metropolis (M. M.) and Gendarmenmarkt (G.). There are 1334/1463 images in the M. M./ G. The number of registered images, reconstructed sparse points, and dense points are reported in Table 4. All of these metrics indicate the completeness of reconstruction. As we can see, SDGMNet generates the most complete model containing the most sparse and dense points, which shows its superiority in large 3D reconstruction tasks.

Discussion

Impact of Hyperparameters

SDGMNet contains two hyperparameters, namely probabilistic hard margin m and attenuation coefficient α. To evaluate the impacts of the hyperparameters, we train SDGMNet on *Liberty* with one of them changing. Then, we report the average FPR@95 of the top 10 checkpoints for higher confidence. The curves of FPR@95 versus m and α are drawn in Fig. 5 (a). Probabilistic hard margin smaller than 0.6 degenerates the performance slightly, which demonstrates that probabilistic hard margin can isolate easy triplets more completely than naive CDF-based soft margin. This modification slightly improves the descriptors, but an excessively large hard margin would release only a few examples into optimization, *e.g.*, fewer than 30% with $m = 0.7$. Too few examples make the model overfit on the current batch and weaken the performance. Compared with m, varying α leads to larger fluctuation. A large α indicates a bias to identify hard positive correspondence. An excessive preference on the positive side more significantly degenerates the generalization of the model, because those extreme distortions the hard positive examples carry in the training sets would not appear in the test, and forcing the model to identify those positive examples is harmful for discriminating negative examples, which are naturally distributed (Zhang et al. 2017).

Ablation Study

SDGMNet contains four components including angular distance, auto-focus modulation (AF), probabilistic margin (PM) and power adjustment (PA). We think AF and implicit modulations are kinds of self weight. Let $\&\theta$, $\&s$, and $\&l_2$ denote self weights computed by Eqs. (5), (6),and (7), respectively. &AF represents our formulation. We assess frameworks that combine four kinds of self weight with PM-based coupled weight. Moreover, to test the efficiency of PA,

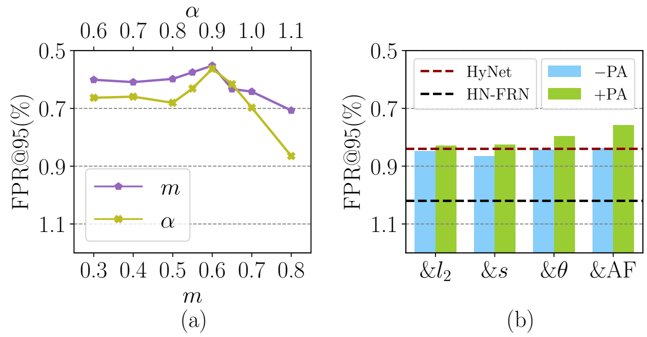


Figure 5: Ablation study on different hyperparameters and components. (a) Performances of SDGMNet on *Liberty* with different probabilistic margin m or attenuation coefficient α. (b) Effectiveness of proposed components on full UBC PhotoTour. The red and black dashed lines denote the mean FPR@95 of HyNet and HN-FRN, respectively.

we embed PA to those raw frameworks. The performances on full UBC PhotoTour are shown in Fig. 5 (b).

Without PA, all frameworks outperform the HardNet (Movshovitz-Attias et al. 2017) embedded with FRN (HN-FRN). Their scores have been floating near the previous record of HyNet (Tian et al. 2020). It is worth mentioning that $\&l_2-$PA is equivalent to the HN-FRN upgraded with PM, which brings a gain of about 0.18. However, angular distance and AF reveal only a little distinction without PA. AF is not so effective probably because a bias is introduced in AF formulation. The bias would mightily mislead the training and degenerate the performance. So an inductive bias is introduced by PA to fix the problem. After PA is equipped, all raw frameworks advance. In such circumstance, $\&\theta$+PA and &AF+PA show their superiority. These outcomes suggest the advantages of the proposed angular distance, AF and PA.

Conclusion

In this paper, we propose a statistic-based dynamic gradient modulation for local descriptor learning, called SDGMNet. SDGMNet devotes to dynamically rescaling the gradients of pair-wise loss. Firstly, SDGMNet conducts deep analysis on backpropagation and chooses included angle which is unbiased in theory for distance measure. Secondly, auto-focus modulation is applied to modulate the gradients of individual pairs. It neutralizes the HEM and noisy example suppression according to the statistical characteristics of individual pairs. Thirdly, SDGMNet combines hard and soft statistic-based probabilistic margins to modulate the gradients of Siamese pairs, *i.e.*, triplets. Finally, total weights, *i.e.*, powers of two kinds of pairs are adjusted by gradient normalization and attenuation coefficient. Local descriptors learned in SDGM strategy show superiority on various tasks and datasets. Every modification in SDGMNet proves efficient through extensive experiments.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62276192).

References

Balntas, V.; Lenc, K.; Vedaldi, A.; and Mikolajczyk, K. 2017. HPatches: A benchmark and evaluation of handcrafted and learned local descriptors. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 5173–5182.

Balntas, V.; Riba, E.; Ponsa, D.; and Mikolajczyk, K. 2016. Learning local feature descriptors with triplets and shallow convolutional neural networks. In *Proc. British Mach. Vis. Conf.*, 1–11.

Barroso-Laguna, A.; Riba, E.; Ponsa, D.; and Mikolajczyk, K. 2019. Key. net: Keypoint detection by handcrafted and learned cnn filters. In *Proc. IEEE Int. Conf. Comput. Vis.*, 5836–5844.

Bishop, C. M. 2006. *Pattern Recognition and Machine Learning*. Springer.

Boutros, F.; Damer, N.; Kirchbuchner, F.; and Kuijper, A. 2022. Elasticface: Elastic margin loss for deep face recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 1578–1587.

Chum, O.; Werner, T.; and Matas, J. 2005. Two-view geometry estimation unaffected by a dominant plane. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 772–779.

Cui, J.; Zhong, Z.; Tian, Z.; Liu, S.; Yu, B.; and Jia, J. 2023. Generalized parametric contrastive learning. *IEEE Trans. Pattern Anal. Mach. Intell.*

Dai, A.; Chang, A. X.; Savva, M.; Halber, M.; Funkhouser, T.; and Nießner, M. 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 5828–5839.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 4690–4699.

Fan, B.; Kong, Q.; Wang, X.; Wang, Z.; Xiang, S.; Pan, C.; and Fua, P. 2019. A performance evaluation of local features for image-based 3D reconstruction. *IEEE Trans. Image Process.*, 28(10): 4774–4789.

Fan, B.; Zhou, J.; Feng, W.; Pu, H.; Yang, Y.; Kong, Q.; Wu, F.; and Liu, H. 2022. Learning semantic-aware local features for long term visual localization. *IEEE Trans. Image Process.*, 31: 4842–4855.

Gleize, P.; Wang, W.; and Feiszli, M. 2023. SiLK–Simple Learned Keypoints. *Proc. IEEE Int. Conf. Comput. Vis.*

He, K.; Lu, Y.; and Sclaroff, S. 2018. Local descriptors optimized for average precision. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 596–605.

Huang, Y.; Wang, Y.; Tai, Y.; Liu, X.; Shen, P.; Li, S.; Li, J.; and Huang, F. 2020. Curricularface: adaptive curriculum learning loss for deep face recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 5901–5910.

Jin, Y.; Mishkin, D.; Mishchuk, A.; Matas, J.; Fua, P.; Yi, K. M.; and Trulls, E. 2021. Image matching across wide baselines: From paper to practice. *Int. J. Comput. Vis.*, 129(2): 517–547.

Keller, M.; Chen, Z.; Maffra, F.; Schmuck, P.; and Chli, M. 2018. Learning deep descriptors with scale-aware triplet networks. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2762–2770.

Lee, J.; Kim, B.; and Cho, M. 2022. Self-Supervised Equivariant Learning for Oriented Keypoint Detection. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 4847–4857.

Li, Z.; and Snavely, N. 2018. Megadepth: Learning single-view depth prediction from internet photos. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2041–2050.

Lowe, D. G. 2004. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.*, 60(2): 91–110.

Mishchuk, A.; Mishkin, D.; Radenovic, F.; and Matas, J. 2017. Working hard to know your neighbor's margins: Local descriptor learning loss. In *Adv. Neural Inf. Process. Syst.*, 4829–4840.

Movshovitz-Attias, Y.; Toshev, A.; Leung, T. K.; Ioffe, S.; and Singh, S. 2017. No fuss distance metric learning using proxies. In *Proc. IEEE Int. Conf. Comput. Vis.*, 360–368.

Ranjan, R.; Castillo, C. D.; and Chellappa, R. 2017. L2-constrained softmax loss for discriminative face verification. *arXiv preprint arXiv:1703.09507*.

Salimans, T.; and Kingma, D. P. 2016. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In *Adv. Neural Inf. Process. Syst.*, volume 29.

Sarlin, P.-E.; DeTone, D.; Malisiewicz, T.; and Rabinovich, A. 2020. Superglue: Learning feature matching with graph neural networks. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 4938–4947.

Schonberger, J. L.; and Frahm, J.-M. 2016. Structure-from-motion revisited. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 4104–4113.

Schonberger, J. L.; Hardmeier, H.; Sattler, T.; and Pollefeys, M. 2017. Comparative evaluation of hand-crafted and learned local features. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 1482–1491.

Schönberger, J. L.; Zheng, E.; Frahm, J.-M.; and Pollefeys, M. 2016. Pixelwise view selection for unstructured multi-view stereo. In *Proc. Europ. Conf. Comput. Vis.*, 501–518.

Singh, S.; and Krishnan, S. 2020. Filter response normalization layer: Eliminating batch dependence in the training of deep neural networks. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 11237–11246.

Sun, J.; Shen, Z.; Wang, Y.; Bao, H.; and Zhou, X. 2021. LoFTR: Detector-free local feature matching with transformers. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 8922–8931.

Sun, Y.; Cheng, C.; Zhang, Y.; Zhang, C.; Zheng, L.; Wang, Z.; and Wei, Y. 2020. Circle loss: A unified perspective of pair similarity optimization. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 6398–6407.

Tian, Y.; Barroso Laguna, A.; Ng, T.; Balntas, V.; and Mikolajczyk, K. 2020. HyNet: Learning Local Descriptor with Hybrid Similarity Measure and Triplet Loss. In *Adv. Neural Inf. Process. Syst.*, 7401–7412.

Tian, Y.; Fan, B.; and Wu, F. 2017. L2-net: Deep learning of discriminative patch descriptor in euclidean space. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 661–669.

Tian, Y.; Yu, X.; Fan, B.; Wu, F.; Heijnen, H.; and Balntas, V. 2019. Sosnet: Second order similarity regularization for local descriptor learning. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 11016–11025.

Tyszkiewicz, M.; Fua, P.; and Trulls, E. 2020. DISK: Learning local features with policy gradient. *Adv. Neural Inf. Process. Syst.*, 33: 14254–14265.

Wang, C.; Xu, R.; Lv, K.; Xu, S.; Meng, W.; Zhang, Y.; Fan, B.; and Zhang, X. 2023. Attention Weighted Local Descriptors. *IEEE Trans. Pattern Anal. Mach. Intell.*

Wang, S.; Li, Y.; Liang, X.; Quan, D.; Yang, B.; Wei, S.; and Jiao, L. 2019a. Better and Faster: Exponential Loss for Image Patch Matching. In *Proc. IEEE Int. Conf. Comput. Vis.*, 4812–4821.

Wang, W.; Zhang, L.; and Huang, H. 2022. Progressive Unsupervised Learning of Local Descriptors. In *Proc. ACM Int. Conf. Multimedia*, 2371–2379.

Wang, W.; Zhang, L.; and Huang, H. 2023. Revisiting Unsupervised Local Descriptor Learning. In *Proc. AAAI Conf. Artif. Intell.*, volume 37, 2680–2688.

Wang, X.; Han, X.; Huang, W.; Dong, D.; and Scott, M. R. 2019b. Multi-similarity loss with general pair weighting for deep metric learning. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 5022–5030.

Winder, S. A.; and Brown, M. 2007. Learning local image descriptors. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 1–8.

Xue, F.; Budvytis, I.; and Cipolla, R. 2023. SFD2: Semantic-guided Feature Detection and Description. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 5206–5216.

Zhang, L.; and Rusinkiewicz, S. 2019. Learning local descriptors with a CDF-based dynamic soft margin. In *Proc. IEEE Int. Conf. Comput. Vis.*, 2969–2978.

Zhang, X.; Yu, F. X.; Kumar, S.; and Chang, S.-F. 2017. Learning spread-out local feature descriptors. In *Proc. IEEE Int. Conf. Comput. Vis.*, 4595–4603.

Zhao, X.; Wu, X.; Miao, J.; Chen, W.; Chen, P. C.; and Li, Z. 2022. Alike: Accurate and lightweight keypoint detection and descriptor extraction. *IEEE Trans. Multimedia*.

Zhong, Y.; Deng, W.; Hu, J.; Zhao, D.; Li, X.; and Wen, D. 2021. SFace: Sigmoid-Constrained Hypersphere Loss for Robust Face Recognition. *IEEE Trans. Image Process.*, 30: 2587–2598.

Zhou, X.; Zhong, Y.; Cheng, Z.; Liang, F.; and Ma, L. 2023. Adaptive Sparse Pairwise Loss for Object Re-Identification. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 19691–19701.