AdaTriplet: Adaptive Gradient Triplet Loss with Automatic Margin Learning for Forensic Medical Image Matching

Khanh Nguyen*, Huy Hoang Nguyen*, and Aleksei Tiulpin

University of Oulu, Oulu, Finland
{khanh.nguyen,huy.nguyen,aleksei.tiulpin}@oulu.fi

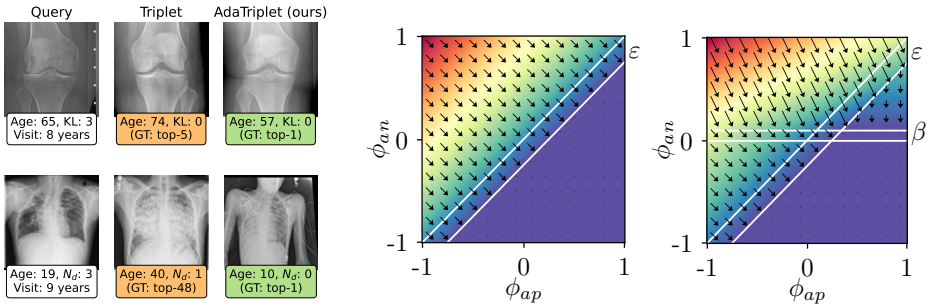
Abstract. This paper tackles the challenge of forensic medical image matching (FMIM) using deep neural networks (DNNs). FMIM is a particular case of content-based image retrieval (CBIR). The main challenge in FMIM compared to the general case of CBIR, is that the subject to whom a query image belongs may be affected by aging and progressive degenerative disorders, making it difficult to match data on a subject level. CBIR with DNNs is generally solved by minimizing a ranking loss, such as Triplet loss (TL), computed on image representations extracted by a DNN from the original data. TL, in particular, operates on triplets: anchor, positive (similar to anchor) and negative (dissimilar to anchor). Although TL has been shown to perform well in many CBIR tasks, it still has limitations, which we identify and analyze in this work. In this paper, we introduce (i) the AdaTriplet loss – an extension of TL whose gradients adapt to different difficulty levels of negative samples, and (ii) the AutoMargin method – a technique to adjust hyperparameters of margin-based losses such as TL and our proposed loss dynamically. Our results are evaluated on two large-scale benchmarks for FMIM based on the Osteoarthritis Initiative and Chest X-ray-14 datasets. The codes allowing replication of this study have been made publicly available at https://github.com/Oulu-IMEDS/AdaTriplet.

Keywords: Deep Learning · Content-based Image Retrieval · Forensic matching

1 Introduction

Content-based image retrieval (CBIR) describes the long-standing problem of retrieving semantically similar images from a database. CBIR is challenging due to the diversity of foreground and background color, context, and semantic changes in images [19]. Besides general computer vision [18,22], in the domain of medicine, content-based medical image retrieval (CBMIR)is growing [3,27], due to the increasing demand for effectively querying medical images from hospital picture archive and communication systems (PACS) [10].

* Equal contributions



(a) FMIM samples (b) Triplet loss (3) (c) AdaTriplet loss (5)

Fig. 1: Comparisons between the Triplet loss and our AdaTriplet loss. (a) Top-1 retrieved results. Green: if a ground truth (GT) is the top-1 in the ranked retrieval list, orange: otherwise. KL indicates the grade of knee osteoarthritis severity. N_d is the number of thorax diseases. (b-c) 2D loss surfaces and negative gradient fields of the two losses. Each point is a triplet. Loss values are represented by colors (increasing from purple to red). The arrows are negative gradient vectors.

In CBMIR, given a medical image (query), one aims to search in a database for images that are similar disease-wise or belonging to the same subject. The former problem is related to diagnostic applications, and the latter problem is of interest for forensic investigations. Hereinafter, we name this problem forensic medical image matching (FMIM). Unlike general CBIR, matching longitudinal medical imaging data of a person is challenging due to aging, progression of various diseases or surgical interventions (Figure 1a). Therefore, the FMIM domain poses new challenges for CBIR.

Deep learning (DL)-based methods have made breakthroughs in various fields, and in particular metric learning, which is the backbone of CBIR [3,12,22,27]. The aim of DL-based metric learning is to train a functional parametric mapping f_θ from the image space $\mathbb{R}^{C\times H\times W}$ to a lower-dimensional feature space $\mathbb{R}^D$. In this feature space, representations of semantically similar images are close, and ones of irrelevant images are distant. In our notation, C, H and W represent the number of channels, height, and width of an image, respectively.

The loss function is the central component of metric learning [13], and there exist two major types: (i) those that enforce the relationships between samples in each batch of data during stochastic optimization – *embedding losses* [9,12,19,25,27] and (ii) *classification losses* [6,22,28]. Two fundamental embedding losses that previous studies have built upon are Contrastive loss (CL) [4] and Triplet loss (TL) [9]. The idea of the CL, is to minimize the feature space distance between similar data points, and maximize it for the dissimilar ones. The TL, on the other hand, considers every triplet of samples – anchor, positive and negative, and aims to ensure that the distance between the anchor and positive samples is smaller than the distance between the anchor and negative ones.

In many practical applications, although the TL is more commonly used than the CL [1,24,26], it also has limitations. Firstly, the TL depends on a "margin"

hyperparameter, which is usually fixed and needs to be chosen empirically. Secondly, as we show in this work, the TL ignores the magnitude of the pair-wise distances, thus may overlook the case where anchors and negative samples are too close. In this paper, we tackle these limitations, and summarize our contributions as follows:

1. We theoretically analyze the TL, and propose *an adaptive gradient triplet loss*, called AdaTriplet, which has appropriate gradients for triplets with different hardness. That characteristic makes our loss distinct from the TL, as illustrated in Figure 1.
2. To address the issue of selecting margin hyperparameters, we propose a simple procedure – AutoMargin, which estimates margins adaptively during the training process, and eliminates the need for a separate grid-search.
3. Through a rigorous experimental evaluation on knee and chest X-ray image forensic matching problems, we show that AdaTriplet and AutoMargin allow for more accurate FMIM than a set of competitive baselines.

2 Methods

2.1 Problem Statement

Let $\mathbf{X} \times \mathbf{Y} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ be a dataset of medical images $\mathbf{x}_i$'s $\in \mathbb{R}^{C\times H\times W}$ and corresponding subjects' identifiers y_i's with $|\mathbf{Y}| \leq N$. We aim to learn a parametric mapping $f_\theta : \mathbb{R}^{C\times H\times W} \to \mathbb{R}^D$ such that $\forall(\mathbf{x}_i, y_i), (\mathbf{x}'_i, y_i), (\mathbf{x}_j, y_j) \in \mathbf{X}_{train} \times \mathbf{Y}_{train}, \mathbf{X}_{train} \subset \mathbf{X}, y_i \neq y_j$,

$$d(f_\theta(\mathbf{x}_i), f_\theta(\mathbf{x}'_i)) < d(f_\theta(\mathbf{x}_i), f_\theta(\mathbf{x}_j)). \tag{1}$$

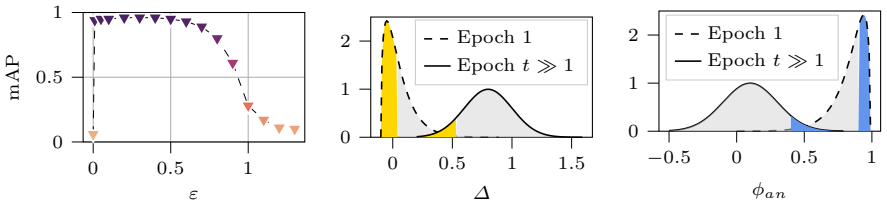
The learned mapping f_θ is expected to be generalizable to $\mathbf{X}_{test} = \mathbf{X} \setminus \mathbf{X}_{train}$ where $\mathbf{Y}_{test} \cap \mathbf{Y}_{train} = \varnothing$. Often, $\mathbf{x}_i$ is called an anchor point, $\mathbf{x}'_i$ – a positive point, and $\mathbf{x}_j$ – a negative point. Hereinafter, they are denoted as $\mathbf{x}_a, \mathbf{x}_p$, and $\mathbf{x}_n$, respectively. For simplicity, we also denote $\mathbf{f}_a = f_\theta(\mathbf{x}_a)$, $\mathbf{f}_p = f_\theta(\mathbf{x}_p)$, $\mathbf{f}_n = f_\theta(\mathbf{x}_n)$, $\phi_{ap} = \mathbf{f}_a^\intercal \mathbf{f}_p$, and $\phi_{an} = \mathbf{f}_a^\intercal \mathbf{f}_n$.

2.2 Triplet Loss

Let $\mathcal{T} = \{(\mathbf{x}_a, \mathbf{x}_p, \mathbf{x}_n) \mid y_a = y_p, y_a \neq y_n\}$ denote a set of all triplets of an anchor, a positive, and a negative data point. For each $(\mathbf{x}_a, \mathbf{x}_p, \mathbf{x}_n) \in \mathcal{T}$, the Triplet loss is formulated as [2,9]:

$$\mathcal{L}_{\text{Triplet}} = \left[\|\mathbf{f}_a - \mathbf{f}_p\|_2^2 - \|\mathbf{f}_a - \mathbf{f}_n\|_2^2 + \varepsilon\right]_+, \tag{2}$$

where $[\cdot]_+ = \max(\cdot, 0)$, and ε is a non-negative margin variable. Following common practice, we normalize all feature vectors, that is $\|\mathbf{f}_a\|_2 = \|\mathbf{f}_p\|_2 = \|\mathbf{f}_n\|_2 = 1$, as well since we can then derive that $\varepsilon \in [0, 4)$. Thereby, we can convert Eq. (2) to a slightly different objective, which is identical to optimize, but allows us to identify limitations of the TL.



(a) Effect of ε in Eq. (3) (b) Distribution of Δ (c) Distribution of ϕ_{an}

Fig. 2: (a) The sensitivity of the Triplet loss (3) with the change of ε. (b-c) The convergences of distributions of $\Delta = \phi_{ap} - \phi_{an}$ and ϕ_{an} under our loss. Yellow and blue areas, specified by Eqs. (7) and (8), indicate hard triplets and hard negative pairs, respectively.

Proposition 1. *Given* $\|\mathbf{f}_a\|_2 = \|\mathbf{f}_p\|_2 = \|\mathbf{f}_n\|_2 = 1$, *minimization of the Triplet loss* (2) *corresponds to minimizing*

$$\mathcal{L}^*_{\text{Triplet}} = [\phi_{an} - \phi_{ap} + \varepsilon]_+ , \ \varepsilon \in [0, 2). \tag{3}$$

Proof. See Suppl. Section 1.

Instead of depending on L2 distances between feature vectors as in (2), the TL in Eq. (3) becomes a function of the cosine similarities ϕ_{ap} and ϕ_{an} (i.e. $\cos(\mathbf{f}_a, \mathbf{f}_p)$ and $\cos(\mathbf{f}_a, \mathbf{f}_n)$, respectively). In Figure 1b, we graphically demonstrate the 2D loss surface of the TL (3) with $\varepsilon = 0.25$, treating ϕ_{ap} and ϕ_{an} as its arguments.

2.3 Adaptive Gradient Triplet Loss

The TL in Eq. (3) only aims to ensure that the distance between the feature vectors $\mathbf{f}_a$ and $\mathbf{f}_p$ is strictly less than the distance between the anchor and a negative $\mathbf{f}_n$. Such a formulation, however, allows for the existence of an unexpected scenario where both the distances are arbitrarily small. We present a simple intuition of the scenario in Suppl. Figure 1. Although increasing the margin ε should enlarge the distance of negative pairs, our empirical evidence in Figure 2a shows that using $\varepsilon > 0.5$ results in a significant drop in performance. Therefore, we propose to explicitly set a threshold on the *virtual angle* between $\mathbf{f}_a$ and $\mathbf{f}_n$, that is $\angle(\mathbf{f}_a, \mathbf{f}_n) \geq \alpha$, where $\alpha \in [0, \pi/2]$, which is equivalent to $\cos(\mathbf{f}_a, \mathbf{f}_n) - \cos(\alpha) \leq 0$. To enforce such a constraint, we minimize the following loss

$$\mathcal{L}_{\text{an}} = [\phi_{an} - \beta]_+ , \tag{4}$$

where $\beta = \cos(\alpha) \in [0, 1]$. Using this additional term, we introduce *an adaptive gradient triplet loss*, named AdaTriplet, that is a combination of $\mathcal{L}^*_{\text{Triplet}}$ and $\mathcal{L}_{\text{an}}$:

$$\mathcal{L}_{\text{AdaTriplet}} = [\phi_{an} - \phi_{ap} + \varepsilon]_+ + \lambda [\phi_{an} - \beta]_+ , \tag{5}$$

where $\lambda \in \mathbb{R}_+$ is a coefficient, $\varepsilon \in [0, 2)$ is a strict margin, and $\beta \in [0, 1]$ is a relaxing margin.

Proposition 2. *Consider* $\|\mathbf{f}_a\|_2 = \|\mathbf{f}_p\|_2 = \|\mathbf{f}_n\|_2 = 1$. *Compared to the Triplet loss, the gradients of AdaTriplet w.r.t.* ϕ_{ap} *and* ϕ_{an} *adapt the magnitude and the direction depending on the triplet hardness:*

$$\left(\frac{\partial \mathcal{L}_{\text{AdaTriplet}}(\tau)}{\partial \phi_{ap}}, \frac{\partial \mathcal{L}_{\text{AdaTriplet}}(\tau)}{\partial \phi_{an}}\right) = \begin{cases} (-1, 1+\lambda) & \text{if } \tau \in \mathcal{T}_+ \cap \mathcal{P}_+ \\ (0, \lambda) & \text{if } \tau \in (\mathcal{T} \backslash \mathcal{T}_+) \cap \mathcal{P}_+ \\ (-1, 1) & \text{if } \tau \in \mathcal{T}_+ \cap (\mathcal{T} \backslash \mathcal{P}_+) \\ (0, 0) & \text{otherwise} \end{cases}, \quad (6)$$

where $\mathcal{T}_+ = \{(\mathbf{x}_a, \mathbf{x}_p, \mathbf{x}_n) \mid \phi_{an} - \phi_{ap} + \varepsilon > 0\}$ *and* $\mathcal{P}_+ = \{(\mathbf{x}_a, \mathbf{x}_p, \mathbf{x}_n) \mid \phi_{an} > \beta\}$.

Proof. See Suppl. Section 2.

In Figure 1c, we illustrate the negative gradient field of AdaTriplet with $\varepsilon = 0.25, \beta = 0.1$, and $\lambda = 1$. As such, the 2D coordinate is partitioned into 4 sub-domains, corresponding to Eq. (6). The main distinction of the AdaTriplet loss compared to the TL is that our loss has different gradients depending on the difficulty of hard negative samples. In particular, it enables the optimization of easy triplets with $\phi_{an} > \beta$, which addresses the drawback of TL.

2.4 AutoMargin: Adaptive Hard Negative Mining

Hard negative samples are those, where feature space mapping $f_\theta(\cdot)$ fails to capture semantic similarity between samples. Recent studies have shown the benefit of mining hard or semi-hard negative samples for the optimization of TL-based metric learning methods [19,25]. These approaches rely on using *fixed* margin variables in training, which are selected experimentally. In this work, we hypothesize that learning the margin *on-line* is not only more computationally efficient, but also allows gaining better results [7,17,28].

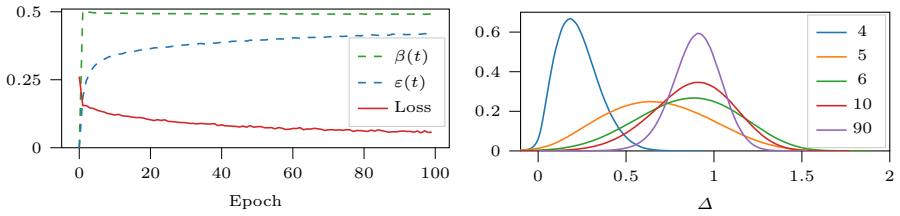
In AdaTriplet, instead of defining hard negatives as the ones for which $\phi_{an} > \phi_{ap}$, we have enforced the numerical constraint on the value of ϕ_{an} itself. Empirically, one can observe that this constraint becomes easier to satisfy as we train the model for longer.

Let $\Delta = \phi_{ap} - \phi_{an}$, we rewrite (5) as $\mathcal{L}_{\text{AdaTriplet}} = [\varepsilon - \Delta]_+ + \lambda [\phi_{an} - \beta]_+$. During the convergence of a model under our loss, the distributions of Δ and ϕ_{an} are supposed to transform as illustrated in Figures 2b and 2c, respectively. Here, we propose adjusting the margins ε and β according to the summary statistics of the Δ and ϕ_{an} distributions during the training:

$$\varepsilon(t) = \frac{\mu_\Delta(t)}{K_\Delta}, \quad (7) \qquad \beta(t) = 1 + \frac{\mu_{an}(t) - 1}{K_{an}}, \quad (8)$$

where $\mu_\Delta(t)$ and $\mu_{an}(t)$ are the means of $\{\Delta \mid (\mathbf{x}_a, \mathbf{x}_p, \mathbf{x}_n) \in \mathcal{T}\}$ and $\{\phi_{an} \mid (\mathbf{x}_a, \mathbf{x}_p, \mathbf{x}_n) \in \mathcal{T}\}$ respectively, and $K_\Delta, K_{an} \in \mathbb{Z}_+$ are hyperparameters.

The difference in $\varepsilon(t)$ and $\beta(t)$ can be observed from their definition: we aim to enforce the triplet constraint with the highest possible margin, and this progressively raises it. Simultaneously, we want to increase the virtual thresholding angle between anchors and negative samples, which leads to the decrease



(a) Margins and our training loss (b) Evolution of distribution of Δ

Fig. 3: Effects of AdaTriplet and AutoMargin. Colors in (b) represent epochs.

of $\beta(t)$. We provide a graphical illustration of adaptive margins in Figures 2b and 2c using yellow and blue colors, respectively.

3 Experiments

3.1 Datasets

Knee X-ray dataset. The Osteoarthritis Initiative (OAI) cohort, publicly available at https://nda.nih.gov/oai/, comprises 4,796 participants from 45 to 79 years old. The original interest of the cohort was to study knee osteoarthritis, which is characterized by the appearance of osteophytes, joint space narrowing, as well as textural changes of the femur and tibia. We used X-ray imaging data collected at baseline, 12, 24, 36, 48, 72, and 96-month follow-up visits. The detailed data description is presented in Suppl. Table 2. We utilized KNEEL [21] to localize and crop a pair of knees joints from each bilateral radiograph. Our further post-processing used augmentations that eventually produces input images with a shape of 256×256 (see Suppl. Table 1a for details).
Chest X-ray dataset. ChestXrays-14 (CXR) [23] consists of 112,120 frontal-view chest X-ray images collected from 30,805 participants from 0 to 95 years old. The radiographic data were acquired at a baseline and across time up to 156 months. The training and test data are further described in Suppl. Table 2. To be in line with the OAI dataset, we grouped testing data by year, and used the same set of augmentations, yielding 256×256 images.

3.2 Experimental Setup

We conducted our experiments on V100 Nvidia GPUs. We implemented our method and all baselines in PyTorch [15] and the Metric Learning library [14]. Following [13], the same data settings, optimizer hyperparameters, augmentations, and feature extraction module were used for all the methods. We utilized the Adam optimizer [11] with a learning rate of 0.0001 and a weight decay of 0.0001. We used the ResNet-18 network [8] with pretrained weights to extract embeddings with D of 128 from input images. We trained each method in 100

Table 1: Ablation studies (5-fold Cross-Validation; OAI dataset). CMC means CMC top-1. * indicates the results when the query and the database are 6 years apart. N_s is the number of scanned hyperparameter values.

(a) Impact of $\lambda\mathcal{L}_{\text{an}}$

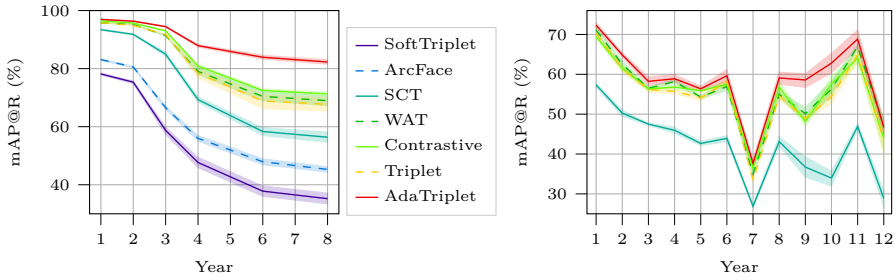
λ	mAP*	CMC*	mAP	CMC
0	95.6	93.4	96.6	93.6
0.5	96.1	94.4	96.9	94.6
1	**96.3**	**94.6**	**97.0**	**94.7**
2	94.5	92.1	95.6	92.3

(b) Triplet loss

Method	N_s	mAP	CMC
Q1	1	27.3	14.9
Q2	1	87.7	76.9
WAT [28]	4	96.5	93.5
Grid search	4	**96.6**	93.6
AutoMargin	2	**96.6**	**93.7**

(c) AdaTriplet loss

Method	N_s	mAP	CMC
Q1	1	94.3	89.4
Q2	1	88.9	79.5
Grid search	16	97.0	**94.8**
AutoMargin	4	**97.1**	94.7



(a) On OAI test set (b) On CXR test set

Fig. 4: Performance comparisons on the test sets of OAI and CXR (mean and standard error over 5 random seeds). Detailed quantitative results are in Suppl. Tables 4 and 5.

epochs with a batch size of 128. For data sampling in each batch, we randomly selected 4 medical images from each subject. We thoroughly describe lists of hyperparameters for all the methods in Suppl. Table 1b.

To evaluate forensic matching performance, we used mean average precision (mAP) [20], mAP@R [13], and cumulative matching characteristics (CMC) accuracy [5]. All experiments were run 5 times with different random seeds. All test set metrics represent the average and standard error over runs.

3.3 Results

Impact of $\mathcal{L}_{an}$. We performed an experiment in which we varied the coefficient λ in the AdaTriplet loss (5). The results on the OAI test set in Table 1a show that $\lambda = 1$ yielded the best performances according to both the mAP and CMC metrics. Notably, we observed that the differences are more apparent when querying images at least 6 years apart from images in the database. We thus set $\lambda = 1$ for our method in all other experiments.

Impact of AutoMargin. AutoMargin is applicable for both TL and AdaTriplet, and we investigated its impact in Tables 1b and 1c. For baselines, we used the Q1 and Q2 quartiles of distributions of Δ and ϕ_{an} to define the margins ε and β, respectively. In addition, we performed exhaustive grid searches for the

two losses' margins. Besides the naïve baselines, we compared our method to the weakly adaptive triplet loss (WAT) [28], which also allows for dynamic margin adjustment in the TL. Based on Suppl. Table 3, we set the constants (K_{Δ}, K_{an}) of AutoMargin to $(2, 2)$ and $(2, 4)$ for OAI and CXR, respectively.

AutoMargin helped both the losses to outperform the quartile-based approaches. Compared to the grid search, our method was at least 2-fold more efficient, and performed in par with the baseline. In the TL, AutoMargin was 2 time more efficient and achieved better results compared to WAT. Furthermore, on the independent test sets, the combination of AdaTriplet and AutoMargin gained substantially higher performances than WAT (Figure 4).

Effects of our methods in training. We demonstrate the behaviour of AdaTriplet and AutoMargin during training of one of the runs of the OAI experiments in Figure 3. Specifically, under our adaptive hard negative mining, the margin β drastically increased from 0 to 0.5 in a few epochs. While β was stable after the drastic increase in value, the margin ε gradually grew from 0 and converged around 0.4. As a result, our loss improved rapidly at the beginning, and continuously converged afterwards (see Figure 3a). During the process, the mean of Δ shifted away from 0 to 1 while its variance increased at first, and then gradually decreased (Figure 3b).

Comparison to baselines. Finally, We compared our AdaTriplet loss with AutoMargin to competitive metric learning baselines such as SoftTriplet [16], ArcFace [6], TL (Triplet) [9,19], CL (Contrastive) [4], WAT [28], and Selectively Contrastive Triplet (SCT) [25]. Whereas SoftTriplet and ArcFace are classification losses, the other baselines are embedding losses. In Figure 4, our empirical results show that the classification losses generalized poorly on the two test sets, especially on chest X-ray data. On both test sets, our loss outperformed all baselines across years. Notably, on the OAI data, the differences between our method and the baselines were more significant at later years. We present more detailed results in Suppl. Tables 4 and 5. Moreover, we demonstrate the retrieval results of our method alongside the baselines in Figure 1a and Suppl. Figure 2.

4 Discussion

In this work, we analyzed Triplet loss in optimizing hard negative samples. To address the issue, we proposed the AdaTriplet loss, whose gradients are adaptive depending on the difficulty of negative samples. In addition, we proposed the AutoMargin method to adjust margin hyperparameters during training. We applied our methodology to the FMIM problem, where the issue of hard negative samples is evident; many medical images may look alike, and it is challenging to capture relevant fine-grained information. Our experiments on two medical datasets showed that AdaTriplet and AutoMargin were robust to visual changes caused by aging and degenerative disorders. The main limitation of this work is that we did not test other neural network architectures, and used grayscale images. However, as recommended in [13], we aimed to make our protocol standard to analyze all the components of the method. Future work

should investigate a wider set of models and datasets. We hope our method will be used for other CBMIR tasks, and have made our code publicly available at https://github.com/Oulu-IMEDS/AdaTriplet.

Acknowledgments

The OAI is a public-private partnership comprised of five contracts (N01- AR-2-2258; N01-AR-2-2259; N01-AR-2- 2260; N01-AR-2-2261; N01-AR-2-2262) funded by the National Institutes of Health, a branch of the Department of Health and Human Services, and conducted by the OAI Study Investigators. Private funding partners include Merck Research Laboratories; Novartis Pharmaceuticals Corporation, GlaxoSmithKline; and Pfizer, Inc. Private sector funding for the OAI is managed by the Foundation for the National Institutes of Health.

We would like to thank the strategic funding of the University of Oulu, the Academy of Finland Profi6 336449 funding program, the Northern Ostrobothnia hospital district, Finland (VTR project K33754) and Sigrid Juselius foundation for funding this work. Furthermore, the authors wish to acknowledge CSC – IT Center for Science, Finland, for generous computational resources.

Finally, we thank Matthew B. Blaschko for useful discussions in relation to this paper. Terence McSweeney is acknowledged for proofreading this work and providing comments that improved the clarity of the manuscript.

References

1. Bai, X., Yang, M., Huang, T., Dou, Z., Yu, R., Xu, Y.: Deep-person: Learning discriminative deep features for person re-identification. Pattern Recognition **98**, 107036 (2020)
2. Chechik, G., Sharma, V., Shalit, U., Bengio, S.: Large scale online learning of image similarity through ranking. Journal of Machine Learning Research **11**(3) (2010)
3. Choe, J., Hwang, H.J., Seo, J.B., Lee, S.M., Yun, J., Kim, M.J., Jeong, J., Lee, Y., Jin, K., Park, R., et al.: Content-based image retrieval by using deep learning for interstitial lung disease diagnosis with chest ct. Radiology **302**(1), 187–197 (2022)
4. Chopra, S., Hadsell, R., LeCun, Y.: Learning a similarity metric discriminatively, with application to face verification. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). vol. 1, pp. 539–546. IEEE (2005)
5. DeCann, B., Ross, A.: Relating roc and cmc curves via the biometric menagerie. In: 2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS). pp. 1–8. IEEE (2013)
6. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4690–4699 (2019)
7. Harwood, B., Kumar BG, V., Carneiro, G., Reid, I., Drummond, T.: Smart mining for deep metric learning. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2821–2829 (2017)

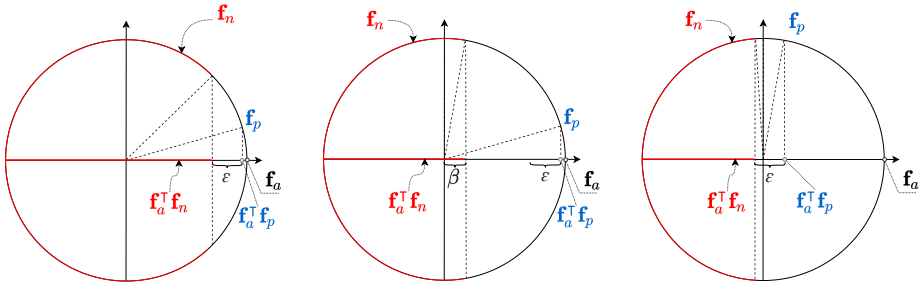
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
9. Hoffer, E., Ailon, N.: Deep metric learning using triplet network. In: International workshop on similarity-based pattern recognition. pp. 84–92. Springer (2015)
10. Hostetter, J., Khanna, N., Mandell, J.C.: Integration of a zero-footprint cloud-based picture archiving and communication system with customizable forms for radiology research and education. Academic radiology **25**(6), 811–818 (2018)
11. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
12. Liang, Y., Han, W., Qiu, L., Wu, C., Shao, Y., Wang, K., He, L.: Exploring forensic dental identification with deep learning. Advances in Neural Information Processing Systems **34** (2021)
13. Musgrave, K., Belongie, S., Lim, S.N.: A metric learning reality check. In: European Conference on Computer Vision. pp. 681–699. Springer (2020)
14. Musgrave, K., Belongie, S., Lim, S.N.: Pytorch metric learning (2020)
15. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
16. Qian, Q., Shang, L., Sun, B., Hu, J., Li, H., Jin, R.: Softtriple loss: Deep metric learning without triplet sampling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6450–6458 (2019)
17. Roth, K., Milbich, T., Ommer, B.: Pads: Policy-adapted sampling for visual similarity learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6568–6577 (2020)
18. Saritha, R.R., Paul, V., Kumar, P.G.: Content based image retrieval using deep learning process. Cluster Computing **22**(2), 4187–4200 (2019)
19. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 815–823 (2015)
20. Schütze, H., Manning, C.D., Raghavan, P.: Introduction to information retrieval, vol. 39. Cambridge University Press Cambridge (2008)
21. Tiulpin, A., Melekhov, I., Saarakkala, S.: Kneel: knee anatomical landmark localization using hourglass networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 0–0 (2019)
22. Tzelepi, M., Tefas, A.: Deep convolutional learning for content based image retrieval. Neurocomputing **275**, 2467–2478 (2018)
23. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2097–2106 (2017)
24. Wu, C.Y., Manmatha, R., Smola, A.J., Krahenbuhl, P.: Sampling matters in deep embedding learning. In: Proceedings of the IEEE international conference on computer vision. pp. 2840–2848 (2017)
25. Xuan, H., Stylianou, A., Liu, X., Pless, R.: Hard negative examples are hard, but useful. In: European Conference on Computer Vision. pp. 126–142. Springer (2020)
26. Yuan, Y., Chen, W., Yang, Y., Wang, Z.: In defense of the triplet loss again: Learning robust person re-identification with fast approximated triplet loss and label distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 354–355 (2020)

27. Zhang, K., Qi, S., Cai, J., Zhao, D., Yu, T., Yue, Y., Yao, Y., Qian, W.: Content-based image retrieval with a convolutional siamese neural network: Distinguishing lung cancer and tuberculosis in ct images. Computers in biology and medicine **140**, 105096 (2022)
28. Zhao, X., Qi, H., Luo, R., Davis, L.: A weakly supervised adaptive triplet loss for deep metric learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 0–0 (2019)

AdaTriplet: Adaptive Gradient Triplet Loss – Supplementary Material

Khanh Nguyen*, Huy Hoang Nguyen*, and Aleksei Tiulpin

University of Oulu, Oulu, Finland
{khanh.nguyen,huy.nguyen,aleksei.tiulpin}@oulu.fi



(a) Triplet loss ($\mathbf{f}_a$ and $\mathbf{f}_p$ are close)

(b) AdaTriplet loss ($\mathbf{f}_a$ and $\mathbf{f}_p$ are close)

(c) AdaTriplet loss ($\mathbf{f}_a$ and $\mathbf{f}_p$ are distant)

Fig. 1: Demonstration of triplets of 2D normalized feature vectors on unit circles. Assume that $\mathbf{f}_a$ is $(1, 0)$, then $\mathbf{f}_a^\intercal\mathbf{f}_p$ and $\mathbf{f}_a^\intercal\mathbf{f}_n$ are the projections of $\mathbf{f}_p$ and $\mathbf{f}_n$ on the horizontal axis, respectively. ε and β are margin variables. Red arcs indicate feasible values of $\mathbf{f}_n$ under a loss function's constraint, and red segments indicate corresponding values of $\mathbf{f}_a^\intercal\mathbf{f}_n$. (a) When the angle between $\mathbf{f}_p$ and $\mathbf{f}_a$ is small, $\mathbf{f}_n$ is allowed to be close to $\mathbf{f}_a$ under the constraint of the Triplet loss (3). (b) In the same scenario, our loss has a term to ensure $\mathbf{f}_n$ to be far from $\mathbf{f}_a$ at least $\arccos(\beta)$ radian. (c) When $\mathbf{f}_p$ is sufficiently far from $\mathbf{f}_a$, the margin ε overrides the effect of β.

1 Proof of Proposition 1

Proof. Consider that for vectors $\mathbf{a}$ and $\mathbf{b}$, s.t. $\|\mathbf{a}\|_2 = \|\mathbf{b}\|_2 = 1$. Then $\|\mathbf{a}-\mathbf{b}\|_2^2 = (\mathbf{a}-\mathbf{b})^\intercal(\mathbf{a}-\mathbf{b}) = 2 - 2\mathbf{a}^\intercal\mathbf{b}$. Therefore,

$$\mathcal{L}_{\text{Triplet}} = \left[2\phi_{an} - 2\phi_{ap} + \varepsilon\right]_+, \ \varepsilon \in [0, 4). \tag{9}$$

By simplifying the coefficient and adjust the range of ε accordingly, we derive Eq. (3).

* Equal contributions

2 Proof of Proposition 2

Proof. Let $\mathcal{T}_+ = \{(\mathbf{x}_a, \mathbf{x}_p, \mathbf{x}_n) \mid \phi_{an} - \phi_{ap} + \varepsilon > 0\}$ and $\mathcal{P}_+ = \{(\mathbf{x}_a, \mathbf{x}_p, \mathbf{x}_n) \mid \phi_{an} > \beta\}$ denote the set of all not-easy triplets and the set of all triplets with a hard negative pair, respectively. Then, the AdaTriplet loss intrinsically partitions the domain of the loss function in into 4 sub-domains:

$$\mathcal{L}_{\text{AdaTriplet}}(\tau) = \begin{cases} (1+\lambda)\phi_{an} - \phi_{ap} + \varepsilon - \lambda\beta & \text{if } \tau \in \mathcal{T}_+ \cap \mathcal{P}_+ \\ \lambda\phi_{an} - \lambda\beta & \text{if } \tau \in (\mathcal{T} \backslash \mathcal{T}_+) \cap \mathcal{P}_+ \\ \phi_{an} - \phi_{ap} + \varepsilon & \text{if } \tau \in \mathcal{T}_+ \cap (\mathcal{T} \backslash \mathcal{P}_+) \\ 0 & \text{otherwise,} \end{cases}, \tag{10}$$

where $\lambda \in \mathbb{R}_+$, and $\tau \in \mathcal{T}$ is a triplet. As a result, we can derive the partial derivatives of $\mathcal{L}_{\text{AdaTriplet}}$ with respect to ϕ_{ap} and ϕ_{an}

$$\left(\frac{\partial \mathcal{L}_{\text{AdaTriplet}}(\tau)}{\partial \phi_{ap}}, \frac{\partial \mathcal{L}_{\text{AdaTriplet}}(\tau)}{\partial \phi_{an}}\right) = \begin{cases} (-1, 1+\lambda) & \text{if } \tau \in \mathcal{T}_+ \cap \mathcal{P}_+ \\ (0, \lambda) & \text{if } \tau \in (\mathcal{T} \backslash \mathcal{T}_+) \cap \mathcal{P}_+ \\ (-1, 1) & \text{if } \tau \in \mathcal{T}_+ \cap (\mathcal{T} \backslash \mathcal{P}_+) \\ (0, 0) & \text{otherwise.} \end{cases} \tag{11}$$

In contrast, $\left(\frac{\partial \mathcal{L}^*_{\text{Triplet}}(\tau)}{\partial \phi_{ap}}, \frac{\partial \mathcal{L}^*_{\text{Triplet}}(\tau)}{\partial \phi_{an}}\right) = (-1, 1), \forall \tau \in \mathcal{T}_+$, which concludes the proof.

Table 1: (a) An ordered list of common transformations. (✓) indicates ones only used in the training phase. (b) Lists of hyperparameter values. Bold and underlined numbers indicate selected values for OAI and CXR, respectively.

(a)

Transformation	**Prob.**	**Parameter**
Resize	1	280×280
Gaussian noise (✓)	0.5	0.3
Rotation (✓)	1	[-10, 10]
Random cropping (✓)	1	256×256
Center cropping	1	256×256
Gamma correction (✓)	0.5	[0.5, 1.5]
Normalization	1	[0.5, 0.3] on OAI [0.5, 0.5] on CXR

(b)

Method	**Hyperparam.**	**List of values**
SCT	λ	{0, <u>0.5</u>, **1**, 2}
ArcFace	m	{5.7, **14.3**, <u>28.6</u>, 43}
WAT	β	{<u>**0.1**</u>, 0.25, 0.5, 0.75}
SoftTriplet	m	{<u>0.01</u>, **0.02**, 0.05, 0.1}
Contrastive	m_{neg}	{0.25, 0.5, <u>**0.75**</u>, 1}
	m_{pos}	{0, 0.25, <u>**0.5**</u>, 0.75}
Triplet + Grid search	ε	{0.1, **0.25**, <u>0.5</u>, 0.75}
+ AutoMargin	K_Δ	{<u>**2**</u>, 4}
AdaTriplet + Grid search	ε	{0.1, 0.25, <u>**0.5**</u>, 0.75}
	β	{0.1, 0.25, **0.5**, <u>0.75</u>}
+ AutoMargin	K_Δ	{<u>**2**</u>, 4}
	K_{an}	{**2**, <u>4</u>}

Table 2: Descriptions of the OAI and CXR datasets. Knee X-ray images with disease indicate those with KL grade greater than 1. Chest X-ray images with disease consist of those with at least one lung or heart disease. OAI test data are from the acquisition site C. The CXR data splits are given by the CXR's owner. Both the galleries contain only data points at their baselines. Queries are from the other follow-up visits.

(a) Overview descriptions

Dataset	**Phase**	**# Images**	**# Subjects**	**% Male**	**# Images with disease**
OAI	Training/validation	37410	3490	59.1	14240
	Test	15648	1306	53.8	5409
CXR	Training/validation	86524	28008	56.0	36324
	Test	25587	2797	58.1	15735

(b) Detailed descriptions of the test sets.

		Query (at year)											
Dataset	**Gallery**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
OAI	2610	2498	2430	2306	2252	0	1858	0	1694	0	0	0	0
CXR	13137	5662	1944	1216	1123	636	528	541	236	304	105	75	80

Table 3: Comparison between an exhaustive grid search for fixed margins and AutoMargin for adaptive margins in the AdaTriplet loss on OAI and CXR. SE means standard error.

(a) Exhaustive grid search

ε	β	**mAP**$_{OAI}$ (%)	**mAP**$_{CXR}$ (%)
0.1	0.1	95.90	79.33
	0.25	96.15	80.08
	0.5	96.27	79.50
	0.75	96.20	80.81
0.25	0.1	96.20	83.48
	0.25	96.66	84.15
	0.5	96.91	84.40
	0.75	96.75	85.27
0.5	0.1	95.70	84.07
	0.25	96.36	86.04
	0.5	**97.02**	85.84
	0.75	96.44	**86.65**
0.75	0.1	92.10	79.29
	0.25	95.59	85.42
	0.5	96.59	81.75
	0.75	95.33	82.89
Mean±SE		96.01±0.28	83.06±0.65

(b) AutoMargin

K_Δ	K_{an}	**mAP**$_{OAI}$ (%)	**mAP**$_{CXR}$ (%)
2	2	**97.08**	85.95
	4	96.70	**87.04**
4	2	96.58	83.71
	4	96.73	85.28
Mean±SE		96.77±0.09	85.50±0.70

Table 4: Performance comparisons on the OAI test set (mean and standard error over 5 random seeds). Bold values indicate the best performances, and underline values indicate ones that are substantially higher than the others. Rows corresponding to our method are highlighted.

Metric	Loss	1 year	2 years	3 years	4 years	6 years	8 years	All
mAP	SoftTriplet	92.1±0.2	91.2±0.3	85.8±0.6	79.4±1.1	73.3±1.2	71.3±1.4	83.2±0.7
	ArcFace	93.0±0.1	92.0±0.1	86.5±0.2	80.0±0.7	74.7±1.0	72.7±0.8	84.2±0.4
	SCT	97.1±0.1	96.6±0.1	94.4±0.2	87.5±0.7	83.4±0.9	81.6±0.9	90.9±0.4
	WAT	98.1±0.0	97.7±0.0	96.6±0.1	92.0±0.6	89.4±0.8	88.1±0.9	94.2±0.3
	Contrastive	98.3±0.0	97.9±0.0	97.1±0.0	93.4±0.2	91.1±0.2	90.2±0.2	95.1±0.1
	Triplet	98.1±0.1	97.8±0.0	96.7±0.1	92.4±0.5	89.6±0.8	88.6±0.8	94.4±0.3
	AdaTriplet	**98.5±0.0**	**98.3±0.0**	**97.9±0.0**	**96.2±0.1**	**95.0±0.2**	**94.2±0.3**	**96.9±0.1**
mAP@R	SoftTriplet	78.2±0.3	75.3±0.5	58.7±1.3	47.7±1.7	37.8±1.8	35.2±1.9	57.6±1.2
	ArcFace	83.1±0.2	80.6±0.3	66.5±0.6	56.0±1.1	47.9±1.2	45.3±1.1	65.1±0.6
	SCT	93.4±0.3	91.8±0.2	85.0±0.8	69.3±1.0	58.3±1.5	56.4±1.9	77.6±0.8
	WAT	95.7±0.1	95.2±0.1	91.4±0.2	79.0±1.1	70.4±1.7	69.0±1.9	84.9±0.7
	Contrastive	96.4±0.1	95.7±0.1	93.0±0.1	80.9±0.7	72.5±0.7	71.3±0.8	86.3±0.3
	Triplet	95.8±0.2	94.9±0.2	91.7±0.3	78.4±1.6	68.8±2.6	67.7±2.5	84.4±1.1
	AdaTriplet	**97.0±0.1**	**96.3±0.1**	**94.5±0.2**	**87.9±0.5**	**83.9±0.7**	**82.3±0.8**	**91.1±0.3**
CMC top 1	SoftTriplet	89.5±0.2	88.2±0.3	80.9±0.9	73.1±1.3	65.4±1.4	63.2±1.6	78.1±0.8
	ArcFace	90.6±0.1	89.2±0.1	81.8±0.3	74.0±0.9	67.3±1.2	64.7±1.0	79.2±0.5
	SCT	95.6±0.1	94.7±0.1	91.1±0.3	81.4±1.0	75.3±1.2	73.3±1.2	86.4±0.6
	WAT	97.3±0.1	96.7±0.1	94.8±0.1	87.9±0.8	83.8±1.1	82.4±1.2	91.3±0.5
	Contrastive	97.6±0.0	97±0.1.0	95.6±0.1	89.9±0.3	86.2±0.2	85.0±0.3	92.5±0.1
	Triplet	97.3±0.1	96.8±0.1	95.0±0.2	88.6±0.7	84.3±1.1	83.1±1.1	91.6±0.5
	AdaTriplet	**98.0±0.1**	**97.7±0.1**	**96.9±0.1**	**94.5±0.2**	**92.8±0.3**	**91.6±0.4**	**95.6±0.2**

Table 5: Performance comparisons on the CXR test set (mean and standard error over 5 random seeds). Results of our AdaTriplet loss are highlighted. The best performances are in bold, and underline values indicate ones that are substantially higher than the others.

Metric	Loss	1 year	2 years	3 years	4 years	5 years	6 years	7 years	8 years	9 years	10 years	11 years	12 years	All
mAP	SoftTriplet	27.3±0.1	21.6±0.3	19.3±0.3	17.4±0.4	17.5±0.3	18.3±0.3	10.6±0.3	15.3±0.8	18.4±2.1	12.6±0.6	13.6±0.6	6.5±0.6	22.3±0.1
	ArcFace	29.9±0.2	24.2±0.1	22.1±0.3	18.7±0.3	19.4±0.3	21.2±0.7	12.6±0.6	17.0±0.7	23.0±1.0	14.0±0.6	14.6±1.3	7.4±0.7	24.8±0.2
	SCT	70.2±0.6	61.5±0.6	59.0±0.3	57.3±0.5	54.0±0.7	55.2±0.6	34.9±0.5	57.7±1.2	56.1±1.7	50.4±1.1	62.5±1.0	44.2±3.0	62.6±0.5
	WAT	81.5±0.4	72.7±0.3	66.1±0.4	67.1±0.4	64.2±0.6	67.1±1.2	43.1±0.5	65.8±0.4	65.6±0.8	69.9±0.6	74.7±0.8	57.9±2.4	73.2±0.3
	Contrastive	79.7±0.4	71.4±0.6	65.5±0.2	66.3±0.3	64.6±0.3	66.1±1.1	43.7±0.5	65.7±1.2	65.2±1.5	69.7±1.4	72.7±0.7	52.3±3.0	72.0±0.4
	Triplet	80.9±0.6	71.3±0.4	65.3±0.6	65.9±0.3	64.4±0.3	67.7±0.5	42.2±0.3	66.2±1.6	66.2±0.4	66.8±1.9	73.2±0.3	55.4±4.3	72.5±0.4
	AdaTriplet	**82.5±0.5**	**74.5±0.2**	**67.0±0.7**	**68.0±0.5**	**65.6±0.6**	**68.7±1.1**	**44.0±0.7**	**68.8±1.0**	**71.1±1.6**	**72.3±1.7**	**74.8±1.1**	**59.6±1.3**	**74.5±0.4**
mAP@R	SoftTriplet	14.9±0.1	11.0±0.2	8.8±0.3	7.9±0.4	8.2±0.2	8.2±0.4	4.6±0.3	5.8±0.8	9.1±1.3	3.6±0.4	4.2±0.9	1.3±0.5	11.4±0.1
	ArcFace	16.7±0.1	12.2±0.3	10.4±0.3	9.0±0.5	9.0±0.6	9.5±0.4	5.5±0.5	7.0±0.3	9.5±1.6	5.1±0.8	5.2±0.5	1.4±0.3	12.9±0.2
	SCT	57.3±0.6	50.3±0.9	47.5±0.3	46.0±0.9	42.6±0.6	43.9±0.8	26.9±0.5	43.1±1.4	36.7±2.8	34.0±2.1	46.9±0.9	28.9±2.2	50.4±0.6
	WAT	71.1±0.7	62.5±1.0	56.5±0.4	58.2±0.5	54.3±0.5	56.9±1.0	35.0±0.6	55.0±0.8	50.1±1.9	56.3±2.1	67.0±0.5	**47.0±1.9**	63.0±0.6
	Contrastive	69.4±0.6	61.5±0.9	56.3±0.4	56.8±0.7	55.9±0.5	57.5±0.7	36.5±0.5	56.8±1.7	48.1±1.3	57.5±2.1	64.1±1.7	43.9±4.0	62.1±0.5
	Triplet	70.2±0.8	61.5±0.8	56.2±0.7	55.7±0.6	54.1±0.8	58.2±0.8	33.4±0.5	54.6±1.1	48.9±0.3	54.4±2.0	64.9±1.4	43.4±2.8	62.1±0.6
	AdaTriplet	**72.3±0.7**	**64.9±0.6**	**58.2±1.1**	**58.9±1.0**	**56.4±0.7**	**59.6±1.6**	**37.7±0.9**	**59.1±1.6**	**58.6±1.9**	**62.8±2.6**	**68.8±2.6**	46.7±1.8	**64.9±0.5**
CMC top 1	SoftTriplet	20.3±0.2	16.5±0.3	14.6±0.3	14.0±0.3	13.7±0.3	13.7±0.4	8.20±0.2	11.5±0.7	14.8±2.2	9.30±0.4	9.80±0.9	4.40±0.8	16.9±0.2
	ArcFace	22.9±0.2	18.9±0.2	17.2±0.3	15.3±0.3	15.3±0.3	16.3±0.5	10.0±0.6	13.5±0.5	18.5±1.0	11.1±0.8	11.0±1.3	5.4±0.8	19.2±0.1
	SCT	57.2±0.5	50.3±0.6	50.2±0.3	48.5±0.6	44.5±0.8	46.3±0.6	29.2±0.4	48.1±1.3	45.9±2.2	39.6±1.0	51.1±1.4	34.0±2.9	51.6±0.4
	WAT	71.4±0.5	64.1±0.5	58.9±0.4	59.9±0.4	56.1±0.6	58.0±0.9	37.2±0.5	58.5±0.4	56.2±1.2	60.4±0.6	68.2±0.8	47.4±2.7	64.3±0.4
	Contrastive	69.8±0.5	62.5±0.8	58.2±0.3	59.3±0.3	57.5±0.5	57.9±1.2	38.0±0.4	58.3±1.5	56.3±1.8	61.3±1.1	66.4±1.0	43.4±2.7	63.3±0.5
	Triplet	70.5±0.8	62.5±0.6	57.7±0.7	58.5±0.3	56.0±0.4	58.4±0.6	36.0±0.2	58.1±1.5	56.1±1.9	57.9±0.4	65.4±0.8	46.4±3.6	63.3±0.6
	AdaTriplet	**72.9±0.6**	**66.1±0.3**	**60.0±0.8**	**61.0±0.6**	**58.1±0.6**	**60.3±0.9**	**38.3±0.8**	**61.3±0.8**	**62.6±2.0**	**63.7±1.8**	**68.5±1.7**	**47.7±0.9**	**66.0±0.4**

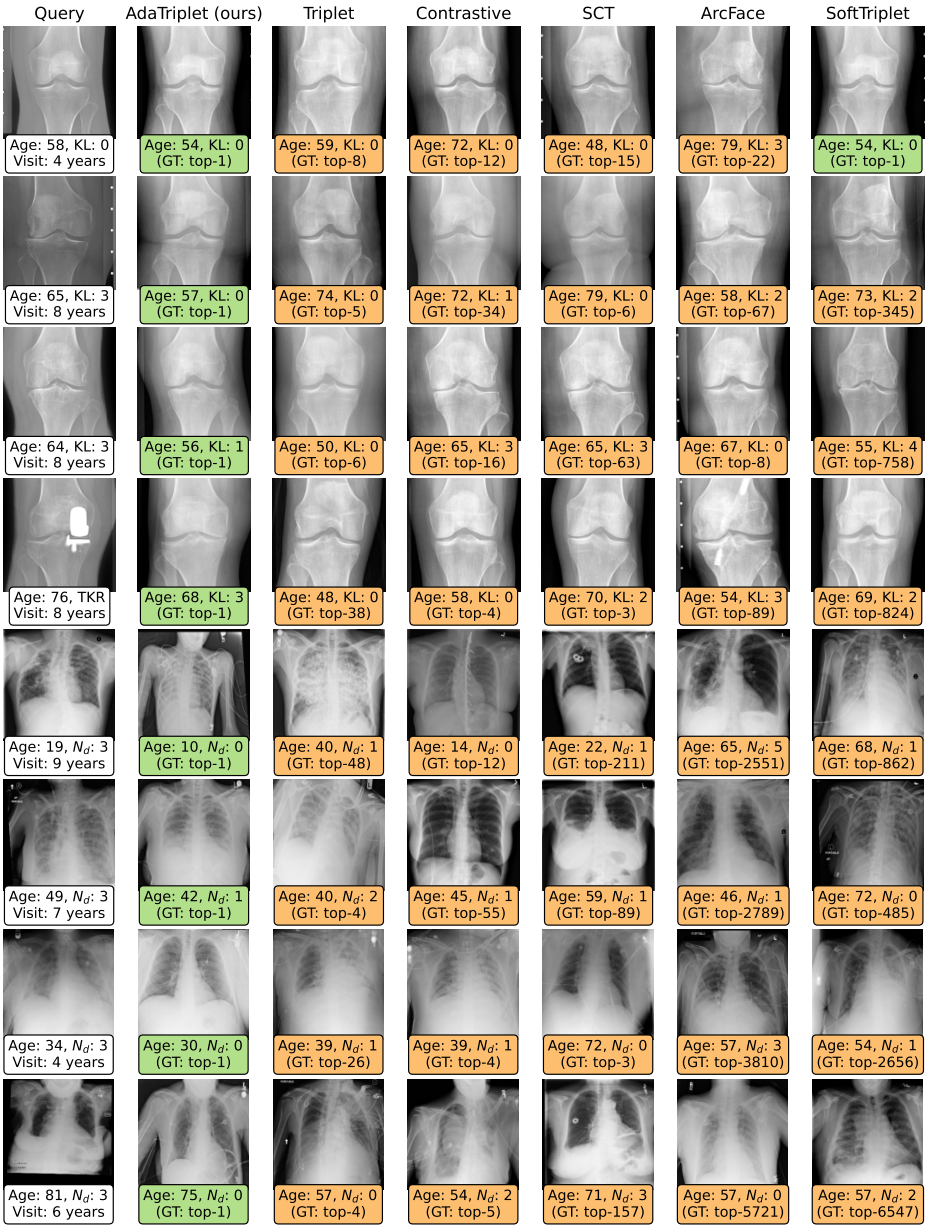


Fig. 2: Matching samples of our method and the other baselines. Columns 2-7 are the top-1 matched images of the corresponding methods. Top-k indicates the position of ground truth (GT). Green: top-1 prediction is the correct person (GT is top-1), orange: otherwise. KL means the Kellgence-Lawrence grade, assessing the stage of knee osteoarthritic severity. TKR indicates knees undergone total knee replacement surgery. N_d is the number of thorax diseases.