IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 3 March 2023, accepted 16 March 2023, date of publication 27 March 2023, date of current version 30 March 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3262271

RESEARCH ARTICLE

Batch Hard Contrastive Loss and Its Application to Cross-View Gait Recognition

MOHAMAD AMMAR ALSHERFAWI ALJAZAERLY[1], YASUSHI MAKIHARA[1], DAIGO MURAMATSU[2], (Member, IEEE), AND YASUSHI YAGI[1], (Senior Member, IEEE)

[1]Department of Intelligent Media, Institute of Scientific and Industrial Research, Osaka University, Osaka 567-0047, Japan
[2]Faculty of Science and Technology, Seikei University, Musashino, Tokyo 180-8633, Japan

Corresponding author: Mohamad Ammar Alsherfawi Aljazaerly (alsherfawi@am.sanken.osaka-u.ac.jp)

This work was supported in part by the Japanese Society for the Promotion of Science (JSPS) KAKENHI under Grant JP19H05692 and Grant JP20H00607, and in part by the MEXT "Innovation Platform for Society 5.0" Program under Grant JPMXP0518071489.

ABSTRACT Biometric person authentication comprises two tasks: the identification task (i.e., one-to-many matching) and the verification task (i.e., one-to-one matching). In this paper, we propose a loss function called batch hard contrastive loss (BHCn) for the deep learning-based verification task. For this purpose, we consider batch mining techniques developed in the identification task and translate them to the verification task. More specifically, inspired by batch mining triplet losses to learn a relative distance for the identification task, we propose BHCn to learn an absolute distance that better represents verification in general. Our method preserves the identity-agnostic nature of the contrastive loss by selecting the hardest pair of samples for each pair of identities in a batch instead of selecting the hardest pair for each sample. We validate the effectiveness of the proposed method in cross-view gait recognition using three networks: a lightweight input, structure, and output network we call GEI + CNN (Gait Energy Image Convolutional Neural Network) as well as the widely used GaitSet and GaitGL, which have sophisticated inputs, structures, and outputs. We trained these networks with the publicly available silhouette-based datasets, the OU-ISIR Gait Database Multi-View Large Population (OU-MVLP) dataset and the Institute of Automation Chinese Academy of Sciences Gait Database Multiview (CASIA-B) dataset. Experimental results show that the proposed BHCn outperforms other loss functions, such as a triplet loss with batch mining as well as the conventional contrastive loss.

INDEX TERMS Biometrics, deep learning, forensics, gait recognition.

I. INTRODUCTION

Biometrics are significant assets for many applications such as surveillance, access control, and forensics. Biometrics have a substantial impact on society, which is why they draw the research community's attention. Many biometrics are available, such as the face, finger veins, fingerprints, voice, iris, and gait. Each biometric has its advantages and disadvantages that should be considered for an application. Using gait as a biometric is highly suitable for applications such as surveillance because gait biometrics are observable even at a distance, unobtrusively, and without requiring the subject's cooperation. At the same time, gait is harder to hide because of its behavioral nature rather than physical. Anyone can hide their face, for example, by simply using a mask. Almost all recent research on this topic uses deep neural networks (DNNs). Researchers have improved many aspects of DNN, such as network structure, pre/post-processing, sampling, and loss functions. Among them, one of the most essential components of the DNN is the loss function, and suitable loss functions depend on the target scenario.

We note there are two main scenarios that use biometrics: identification and verification. In an identification scenario, given an instance of biometrics, the goal is to determine the best match in a gallery of biometric instances, i.e., a one-to-many comparison. In the case of gait, a police officer might want to re-identify a suspect from one (closed-circuit television) CCTV video in another CCTV video that contains

The associate editor coordinating the review of this manuscript and approving it for publication was Zahid Akhtar.

many other non-suspect gaits. For identification, the triplet loss [1] is widely adopted [2], [3], [4], [5], [6], [7], [8], [37]. The triplet loss operates on three embeddings (i.e., sample points in the discriminative feature space), which are the outputs of the DNN, and where exactly two out of the three input samples have the same identity (i.e., an anchor sample, a positive sample, and a negative sample). This loss penalizes the relative distance between the distance of a positive pair (the anchor and positive embeddings) and that of a negative pair (the anchor and negative embeddings) to make the positive pair relatively closer than the negative pair. Even if the distance between a positive pair is big in absolute terms (i.e., bigger than some other positive pair), the loss can be small as long as the distance between the negative pair is bigger, which corresponds well with the identification scenario but not well with the following scenario.

In verification scenarios, given two instances of biometrics, the goal is to determine whether these two instances are of the same identity or not, i.e., a one-to-one comparison. In the case of a verification scenario using gait, a police officer might want to compare the gait of the perpetrator to the gait of a suspect. For verification, the contrastive loss (Cn) [9] is more suitable than the triplet loss [2], [10], [11] and its variations have been proposed in [7], [12]. The contrastive loss is a pair loss, and therefore it handles the positive/negative pairs differently. To reduce this loss, the embeddings of the positive pair have to be as close as possible to each other (i.e., the distance should be ideally 0), whereas the embeddings of the negative pair have to be farther apart than a margin distance (i.e., greater than margin m). It is sufficient for a positive pair to be relatively closer than a negative pair in a triplet. In contrastive loss, the positive pair has to be absolutely close (i.e., closer than margin m at the least), and the negative pair has to be absolutely far (i.e., farther than margin m). We call this kind of distance learned with such loss an absolute distance, which corresponds well with the one-to-one matching.

In addition to the loss function itself, effectively sampling the data for training using a triplet or a pair loss dictates the performance of any deep learning model. We identified three main sampling methods, random sampling, sample mining, and batch sampling, as described below.

A. RANDOM SAMPLING

This sampling method is the simplest way to create a group of samples (e.g., pairs or triplets) by randomly taking samples from the training dataset [9], [13], [14], [15], [16]. However, after a few epochs of training the network performance increases. As a result, the embeddings of these random groups become more likely to be well separated in the discriminative feature space. Therefore the loss of such groups is zero, which means they do not participate in updating the weights (i.e., the loss is inactive because the gradients of zero loss are also zeros). The number of these inactive groups gradually increases, and in the worst case, all groups could be inactive; therefore, training would require additional iterations to determine active groups. Consequently, the training process may become slow because of these inactive groups with zero losses.

B. SAMPLE MINING

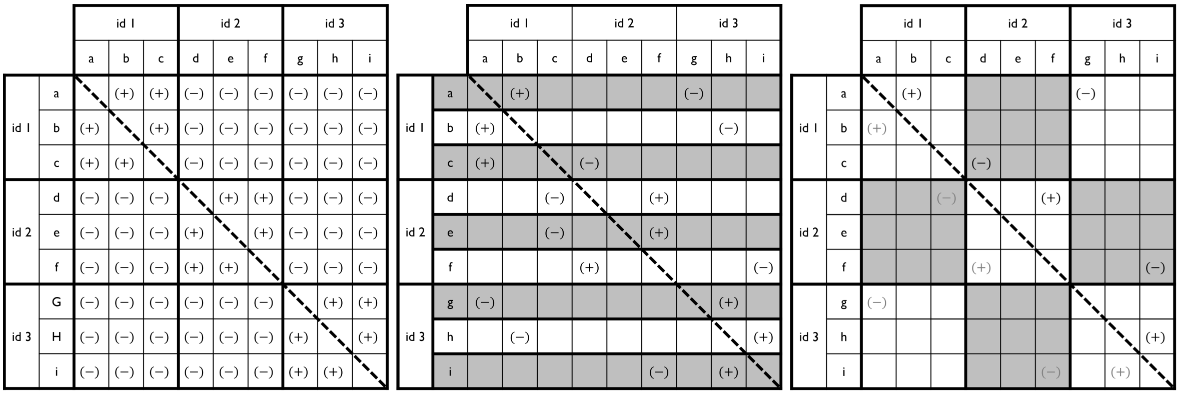
In this method, an algorithm searches the dataset to determine active groups, i.e., mine the dataset, to obtain a more efficient training process. When each training iteration has an adequate number of active groups, the network learns a more discriminative space efficiently. Studies such as [17], [18], [19], [20] start with random initial samples and mine the dataset for the corresponding effective groups, which are where the positive pair embeddings are far apart and the negative pair embeddings are close to each other. The mining process requires inferences about the samples to be made to obtain the embeddings and then searching for the most effective groups while discarding the other groups, which is computationally expensive. This creates a computational resources assignment problem between mining and training.

C. BATCH SAMPLING

This sampling method does not base the mining process on random initial samples. It instead starts with a mini-batch and considers all available groups created in the batch. A mini-batch has K random samples from each of P random identities. After embedding selects samples based on the pair-wise distance matrix of all $P \times K$ embeddings of the mini-batch (see Fig. 1 for an illustration of this method). Current methods reduce the waste of mining and achieve better performance [3], [4], [5], [6], [7], [8], [11], [21], [22], [23], [24]. Most of these methods can be adapted to work with the triplet loss, where there is an anchor sample, a positive sample, and a negative sample. As shown in Fig. 1a, they follow batch-all sampling for triplets, i.e., each sample in the mini-batch is an anchor.Samples of the same identity as the anchor are candidates for the positive sample, and samples of a different identity to the anchor are candidates for the negative sample. The simplest approach is to determine the loss for all possible triplets. Alternatively, some methods [5], [8], [11] only use hard triplets, that is, the $(+)$ and $(-)$ in Fig. 1b, which illustrates such hard examples in a mini-batch.

The studies mentioned above developed their sampling strategy according to the triplet loss, which is suitable for the identification scenario. By contrast, a batch-sampling method for verification is still missing from the literature, to the best of our knowledge.

We therefore propose a batch-sampling method for verification scenarios. Our batching method is heavily inspired by the progress on the triplet loss in the identification literature. However, there is a mismatch between the relative distance learned by the triplet loss and verification, which requires absolute distances. Therefore we construct our method on top of the contrastive loss. Moreover, instead of using the batch all sampling method (Fig. 1a), Our method



		id 1			id 2			id 3		
		a	b	c	d	e	f	g	h	i
id 1	a		(+)	(+)	(−)	(−)	(−)	(−)	(−)	(−)
	b	(+)		(+)	(−)	(−)	(−)	(−)	(−)	(−)
	c	(+)	(+)		(−)	(−)	(−)	(−)	(−)	(−)
id 2	d	(−)	(−)	(−)		(+)	(+)	(−)	(−)	(−)
	e	(−)	(−)	(−)	(+)		(+)	(−)	(−)	(−)
	f	(−)	(−)	(−)	(+)	(+)		(−)	(−)	(−)
id 3	G	(−)	(−)	(−)	(−)	(−)	(−)		(+)	(+)
	H	(−)	(−)	(−)	(−)	(−)	(−)	(+)		(+)
	i	(−)	(−)	(−)	(−)	(−)	(−)	(+)	(+)	

(a) Batch all sampling

		id 1			id 2			id 3		
		a	b	c	d	e	f	g	h	i
id 1	a		(+)					(−)		
	b	(+)							(−)	
	c	(+)			(−)					
id 2	d			(−)			(+)			
	e			(−)			(+)			
	f				(+)					(−)
id 3	g	(−)							(+)	
	h		(−)							(+)
	i						(−)		(+)	

(b) Sample-based batch sampling

		id 1			id 2			id 3		
		a	b	c	d	e	f	g	h	i
id 1	a		(+)					(−)		
	b	(+)								
	c				(−)					
id 2	d			(−)			(+)			
	e									
	f				(+)					(−)
id 3	g	(−)								
	h									(+)
	i						(−)		(+)	

(c) Identity-based batch sampling

FIGURE 1. Different batch-sampling methods visualized using a pair-wise dissimilarity matrix. The batch consists of a number of identities, and each identity has the same number of samples. Each sub-figure shows the selected positive (+) and negative (−) pairs. a Batch all sampling, where every pair is selected. b Sample-based sampling, where for each sample, the hardest positive pair and hardest negative pair are selected. c Identity-based batch sampling, where the hardest pair is selected for each pair of identities (a) Batch all sampling. (b) Sample-based batch sampling. (c) Identity-based batch sampling.

uses hard sampling. However, the sample-based batch sampling method introduces imbalance in the sampled identities, which is reflected in the asymmetric samples in the pair-wise distance matrix (Fig. 1a). This imbalance does not correspond well with the one-to-one comparison nature of verification. Our method samples the hardest pair for each pair of identities (as illustrated in Fig. 1c). We name this batching method, when coupled with the contrastive loss, batch hard contrastive loss (BHCn).

In addition to BHCn, we define and investigate batch all contrastive loss (BACn), which is a natural extension of the state-of-the-art batch all triple loss (BATr), by replacing the triplet loss with contrastive loss. Furthermore, an extensive evaluation is performed on two popular gait recognition datasets, the OU-ISIR Gait Database Multi-View Large Population (OU-MVLP) dataset [25] and the Institute of Automation Chinese Academy of Sciences Gait Database Multiview (CASIA-B) dataset [26]. Different network architectures are used, a light convolutional neural network (CNN), GaitSet [3], and GaitGL [4].

The outline of this paper is as follows: We summarize related work in Section 2, and describe the training network and loss for gait recognition in Section 3. In Section 4, we present the performance evaluation for gait verification, and discuss the proposed loss in Section 5. Finally, we present our conclusions and discuss future work in Section 6.

II. RELATED WORK

In this section, we introduce each family of sampling methods for mini-batch construction and the rationale behind our proposed loss.

A. BATCH ALL SAMPLING

This family makes the most of the information in each batch, not by increasing the batch size but by constructing all possible pairs from a given batch [5], [21], [22]. A batch is generally composed of multiple identities, and each identity contains multiple samples (the number of samples per identity is sometimes the same among the identities, as in [3], [4], [5], [37]). Each sample in the batch is regarded as an anchor and then all possible positive pairs (i.e., the anchor and a sample of the same identity) and negative pairs (i.e., the anchor and a sample of a different identity) are composed. All the possible triplets for the anchor are further constructed by combining the positive and negative pairs that have a common anchor.

Some techniques apply different grouping strategies, for example, [23] applies batch all sampling to make negative pairs only, whereas [8] applies it to make positive pairs only. Moreover, Sohn [24] generalized the triplet loss to accept a positive pair and multiple negative embeddings at one time. The mini-batch contains $2n$ samples that form n pairs of positive samples from n different identities, i.e., each positive pair has a unique identity in the mini-batch. They then apply the loss on each positive pair and the $n-1$ negative samples from the other pairs. Chen et al. [6] proposed quadruplet loss as a constraint to the triplet loss so that the minimum inter-class distance is more prominent than the maximum intra-class distance. They consider all possible triplets and quadruplets in a batch.

B. BATCH WEIGHTED SAMPLING

Instead of giving all groups the same importance, which could cause the model to become stuck in a local minimum, batch weighted sampling gives different groups different weights to contribute to the weight training gradient. Wu et al. [7] focus on the distances between samples in a batch. This approach treats each sample as an anchor, and for each anchor, it selects all positive embeddings but uniformly samples negative embeddings according to their distance to the anchor.

C. BATCH HARD SAMPLING

This family takes the most difficult samples in the batch for efficient training. Hermans et al. [5] propose a standard framework on a batch of P identities and K samples for each identity, i.e., PK samples in total. Once a sample is set to an anchor, the other samples are categorized into $(K - 1)$ positive samples (i.e., the same identity) and $(P - 1)K$ negative samples (i.e., different identities). They then select the hardest positive sample (i.e., whose distance to the anchor is the largest among the $(K - 1)$ positive samples) and the hardest negative sample too (i.e., whose distance to the anchor is the smallest among the $(P-1)K$ negative samples, as shown in Fig. 1b. Gao et al. [23] propose a fusion method of batch all and batch hard sampling strategies, i.e., employing batch hard sampling for positive samples while employing the batch all sampling for negative pairs.

Yuan et al. [11] propose a cascade-structure model in which only a percentage of each hardest negative and hardest positive pairs are forwarded to the next sub-model in the cascade. Therefore, the last model trains on the hardest positive and the hardest negative pairs. Song et al. [8] start their sampling method by considering all positive pairings in a batch and then selecting the hardest negative sample for each sample in the positive pairs. Yuan et al. [27] train using a quadruplet loss with exactly three samples of the same identity. They take the hardest positive pair and the hardest negative pair for each identity.

D. OTHER METHODS

There are some other studies that improve discrimination capability by designing more suitable loss functions. Lezama et al. [28] use the matrix trace norm to push the same identity in a low-rank subspace and different identities so that they are linearly independent, i.e., orthogonal. The columns of intra-identity matrices are created from the batch's identity embeddings, whereas the inter-identity matrix is created from all embeddings in the batch. However, this loss is set up to support a softmax loss, not a standalone loss. Another approach [29], [30], [31], [32], [33], [34] reformulates the metric learning problem as a classification problem. They perform sampling in the same manner as classification, i.e., by feeding the data iteratively to the model for training. Wang et al. [29] proposed two reformulations of the contrastive loss and triplet loss to create a classification problem by grouping the samples with classes vectors instead of other samples. Other studies [30], [31], [32], [33], [34] proposed adding angular margin-based modification to cross-entropy softmax classification loss to better embed high-level features.

1) RATIONALE BEHIND BHCn LOSS

Our BHCn loss, which is based on Euclidean distance, is suitable for the gait verification task for the following reasons. Many other biometrics (e.g., faces and fingerprints) often suffer from illumination variation (e.g., faces and contactless fingerprint) or measured intensity variation (e.g., low-contrast latent fingerprint images), and hence an intensity-normalized dissimilarity measure such as cosine distance is more appropriate. By contrast, gait silhouettes are binary and do not suffer from illumination variation. This binary nature of the gait silhouettes is why the Euclidean distance is effective for gait recognition. Moreover, target scenarios of biometrics fall into two main types of task, identification and verification, and suitable loss functions depend on the target scenario [2]. In identification scenarios, the relative distances of positive and negative pairs are important, and hence a triplet loss with an anchor, a positive sample, and a negative sample or its variant is often employed. However, in verification scenarios, the absolute distances of positive and negative pairs are important, hence a contrastive loss is a sensible choice. Our proposed BHCn aims to use contrastive loss for the most meaningful pairs while preserving the identity balance in the batch, and this is the rationale behind our method.

III. GAIT RECOGNITION WITH BHCn

A. NETWORK STRUCTURE

We design a network that takes a gait image as an input and outputs an embedding (i.e., a discriminative feature vector).

There are many gait representations that can be used as input. The two main representations are 1) a temporally compressed image over a gait cycle (i.e., a kind of gait template image) and 2) a sequence or set of gait images. The most common representation for the first category is the gait energy image (GEI) [35], which is generated by averaging the aligned silhouettes over one gait cycle. This representation is compact and lightweight yet effective and hence has been employed for a long time in the gait recognition community. We design a simple CNN that takes a GEI and is named GEI + CNN. The details of its network architecture are shown in Fig. 2.

The temporally compressed image, however, loses the fine-grained temporal information, and hence the gait representation from the second category has become more popular recently. We therefore adopt GaitSet [3] as the second category. GaitSet directly feeds an unordered set of silhouettes to the network so as not to lose each frame's information. Moreover, GaitSet employs horizontal pyramid pooling, in which a gait silhouette is vertically divided into parts (i.e., horizontal stripes) at multiple scales so as to capture both local and global information. As such, GaitSet outputs multiple vectors corresponding to each horizontal stripe, and the loss function is computed for each output separately, unlike the existing GEI-based network (e.g., GEINet [36]), which outputs a single vector for the whole body.

The unordered image set does not contain the characteristics of gait that change over time, and hence more recently, sequence representation has become more popular [4], [37]. We adopted GaitGL [4] as a second category to cover both set and sequence gait images. GaitGL represents time as the third

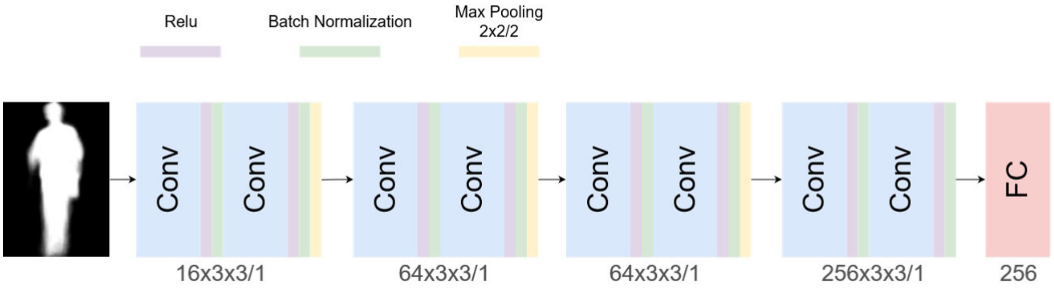


FIGURE 2. Our GEI + CNN structure. A simple network with four convolution blocks and a final fully connected layer. The input is a GEI with 128 × 88 pixels, and the output is a 256-dimensional vector.

dimension and employs 3D convolution to extract spatio-temporal features. Moreover, GaitGL applies these convolutions both on the whole body (i.e., globally) and on horizontal strips of the body (i.e., locally) to capture both local and global spatio-temporal information. The final features are per horizontal strip, and as such, like GaitSet, GaitGL outputs a multi-vector gait representation.

B. LOSS FUNCTIONS

We first introduce widely used relevant loss functions such as the triplet loss and its extension and the conventional contrastive loss as preliminaries to make this paper self-contained. The proposed method, named BHCn, is then introduced as an extension to the contrastive loss [9].

1) TRIPLET LOSS AND ITS VARIANTS

The triplet loss is defined by a triplet of an anchor, a positive sample whose identity is the same as the anchor, and a negative sample whose identity is different from the anchor. More specifically, the triplet loss penalizes if the relative distance of the negative pair to the positive pair does not exceed a pre-defined margin m. Given embeddings of the anchor, the positive sample, and the negative samples as y_a, y_p, and y_n, respectively, the triplet loss is defined as

$$\mathcal{L}_{Tr}(y_a, y_p, y_n) = [m + D(y_a, y_p) - D(y_a, y_n)]_+, \quad (1)$$

where $D(\cdot, \cdot)$ is a distance function (typically, Euclidean distance), and $[\cdot]_+$ is a non-negative clipping function defined as $[\cdot]_+ = \max(0, \cdot)$.

Hermans et al. [5] extends the triplet loss to the batch all triplet loss to leverage the available information in a batch. Assume that a batch is composed of P identities and each identity has K samples, i.e., PK samples in total. The embedding of the i-th identity of the j-th sample is denoted as y^i_j and a set of the embeddings as $Y = \{y^i_j\}$, and the batch all triplet loss is then defined as

$$\mathcal{L}_{BATr}(Y) = \frac{1}{|\mathcal{L}^+|} \sum_{i=1}^{P} \sum_{a=1}^{K} \sum_{\substack{p=1 \\ p\neq a}}^{K} \sum_{\substack{j=1 \\ j\neq i}}^{P} \sum_{n=1}^{K} [m + D(y^i_a, y^i_p) - D(y^i_a, y^j_n)]_+ \quad (2)$$

where $|\mathcal{L}^+|$ is the number of nonzero triplet losses over all triplet combinations and $1/|\mathcal{L}^+|$ is the active triplets averaging weight, i.e., the active losses are averaged and the inactive losses are discarded.

They also proposed the batch hard triplet loss, which selects the positive and negative samples on which the anchor performs the worst in a batch as

$$\mathcal{L}_{BHTr}(Y) = \frac{1}{|\mathcal{L}^+|} \sum_{i=1}^{P} \sum_{a=1}^{K} [m + \max_{p=1,\ldots,K} D(y^i_a, y^i_p) - \min_{\substack{j=1,\ldots,P \\ n=1,\ldots,K \\ j\neq i}} D(y^i_a, y^j_n)]_+ \quad (3)$$

where $1/|\mathcal{L}^+|$ is the active anchors averaging weight.

2) CONTRASTIVE LOSS AND ITS VARIANTS

The contrastive loss is defined on a pair of embeddings. Unlike the triplet loss, which considers the relative distance, the contrastive loss considers the absolute distance of positive/negative pairs. This attribute makes the contrastive loss better suited for verification scenarios. Given a pair of embeddings y_i and y_j, the contrastive loss is defined as

$$\mathcal{L}_{Cn}(y_i, y_j) = \begin{cases} D(y_i, y_j), & (positive\ \ pair) \\ [m - D(y_i, y_j)]_+, & (negative\ \ pair), \end{cases} \quad (4)$$

where m is a margin for the negative pair.

We then consider extending the contrastive loss so as to leverage all the available information in a batch. A batch hard variant of the contrastive loss is proposed, which is analogous to the batch hard triplet loss. A straightforward application of sample-based selection (1b) in the batch hard triplet loss may lead to a problem. The hardest negative for all the samples may be biased toward a specific identity (i.e., the so-called *lamb* in a Doddington biometric zoo). The training result may over-fit to the lamb as a consequence.

The proposed method is therefore to select the most challenging sample in an identity pair-wise way instead of the sample-wise way. Similar to the batch all/hard triplet loss case, assume that a batch contains P identities and each identity contains K samples, i.e., a total of PK samples.

More specifically, given a pair of identities, $i, j \in \{1, \ldots, P\}$,The method constructs K^2 pairs of samples within the identity pair and then selects the hardest sample pair from the K^2 pairs. In summary, the BHCn loss is defined as

$$\mathcal{L}_{BHCn}(Y) = \frac{1}{|\mathcal{L}^+|} \sum_{i=1}^{P} \sum_{j=1}^{P} \max_{\substack{k=1,\ldots,K \\ l=1,\ldots,K}} (\mathcal{L}_{Cn}(y_k^i, y_l^j))^2 \quad (5)$$

where $1/|\mathcal{L}^+|$ is the active identity pairs averaging weight.

a: GROUPED BHCn

When the network structure outputs multiple embedding vectors (e.g., GaitSet and horizontal pyramid mapping (HPM) [38]), the loss function is usually defined for each embedding (i.e., each group) separately and then summed later.

Given the number of multiple embeddings as G, where the g-th embedding indicated by a subscript g, the grouped version of the BHCn is defined as

$$\mathcal{L}_{BHCn}(Y) = \frac{1}{|\mathcal{L}^+|} \sum_{i=1}^{P} \sum_{j=1}^{P} \max_{\substack{k=1,\ldots,K \\ l=1,\ldots,K}} \times \left(\frac{1}{G} \sum_{g=1}^{G} \mathcal{L}_{Cn}(y_{g,k}^i, y_{g,l}^j) \right)^2 \quad (6)$$

where $1/|\mathcal{L}^+|$ is the active identity pairs averaging weight. To test the effectiveness of the proposed BHCn loss, we also define two versions of it. First, the BACn loss is defined in the same way as BATr by replacing the triplet loss in the BATr loss with contrastive loss, i.e., single-step averaging over the samples of all pairs of identities.

$$\mathcal{L}_{BACn}(Y) = \frac{1}{G} \sum_{g=1}^{G} \frac{1}{|\mathcal{L}^+|} \sum_{i=1}^{P} \sum_{a=1}^{K} \sum_{j=1}^{P} \sum_{b=1}^{K} \mathcal{L}_{Cn}(y_{g,a}^i, y_{g,b}^j) \quad (7)$$

where $1/|\mathcal{L}^+|$ is the active pairs averaging weight.

Second, the BACn2 is defined loss by replacing the hard selection in BHCn with averaging over all pairs, i.e., two-step averaging, which consists of averaging over samples for each pair of identities and averaging over pairs of identities.

$$\mathcal{L}_{BACn2}(Y) = \frac{1}{|\mathcal{L}^+|} \sum_{i=1}^{P} \sum_{j=1}^{P} \frac{1}{G} \sum_{g=1}^{G} \frac{1}{|\mathcal{L}_{ijg}^+|} \sum_{a=1}^{K} \sum_{b=1}^{K} \times \mathcal{L}_{Cn}(y_{g,a}^i, y_{g,b}^j) \quad (8)$$

where for a pair of identities i and j, $1/|\mathcal{L}_{ijg}^+|$ is the active sample pairs averaging weight in group g and $1/|\mathcal{L}^+|$ is the active identity pairs averaging weight.

IV. EXPERIMENTS

A. DATASETS

1) OU-MVLP DATASET

The OU-MVLP [25] is as of now the largest publicly available silhouette-based gait dataset. It includes 10,307 subjects, captured from 14 view angles 0° (frontal view), 15°, 30°, 45°, 60°, 75°, 90° (side view), 180°, 195°, 210°, 215°, 240°, 255°, and 270°. Each subject has two walking sequences labeled "00" and "01," which are assigned to a gallery and probe, respectively. As a result, the total number of sequences per subject is $14 \times 2 = 28$.

For evaluation, we use the same training/test split defined in [2], where there are 5,153 training subjects and the rest are the 5,154 test subjects.

2) CASIA-B DATASET

The CASIA-B [26] is another popular dataset that includes 124 subjects captured under three walking statuses, i.e., six normal walking (NM) sequences, two walking sequences with a bag (BG), and two walking sequences wearing a coat (CL), from 11 view angles: 0°, 18°, 36°, 54°, 72°, 90°, 108°, 126°, 144°, 162°, and 180°. Because our aim is pure cross-view gait recognition evaluation, we limit our evaluation to NM. As a result, the total number of sequences per subject is $6 \times 11 = 66$.

For evaluation, we use three training/test splits, which are well known in the literature. The settings are small-sample training (ST), where subjects labeled $1, \ldots, 24$ are for training and the rest are for testing, medium-sample training (MT), where subjects labeled $1, \ldots, 64$ are for training and the rest are for testing, and large-sample training (LT), where subjects labeled $1, \ldots, 74$ are for training and the rest are for testing.

B. PRE-PROCESSING

Both the OU-MVLP and CASIA-B datasets provide binary images of the extracted silhouettes at the original image resolution; hence, the apparent heights differ among subjects.

Because the networks take a set of fixed-height binary silhouettes as an input, preprocessing register and size-normalize the extracted silhouette based on the method of [2], [3]. One exception is the use of bilinear interpolation instead of cubic interpolation, because cubic interpolation returns negative pixel values at the silhouette edges, which become decimated when the image is converted to 8-bit storage. Then GEIs are generated from the size-normalized silhouettes by averaging over one gait cycle. The image resolution of the size-normalized silhouettes (i.e., for the GaitSet and GaitGL networks) is 64×44 pixels. The GEI (i.e., for the GEI + CNN), it is 128×88 pixels to be compatible with the network architectures.

C. LOSSES

To evaluate our proposed method (BHCn), we compare it with different losses. The first loss is the state-of-the-art loss for gait identification, BATr. The second loss is the traditional contrastive loss (Cn). The last losses are the defined BACn losses, BACn and BACn2. With the exception of the experiments using GaitGL, we also compare the combined BATr and cross entropy (BATr_CE) losses, as detailed in [4].

Although we wanted to compare the proposed loss with the batch hard triplet loss, the network did not converge under this loss.

TABLE 1. Batch size settings described in the format of (#identities) × (#samples per identity).

Network \ Dataset	CASIA-B	OU-MVLP
GEI + CNN	8×16	128×8
GaitSet	8×16	64×8
GaitGL	8×16	32×8

TABLE 2. #Training iterations depending on network architecture and dataset.

Data Net.	CASIA-B			OU-MVLP
	ST	MT	LT	
GEI + CNN	25 × 2048	30 × 2048	40 × 2048	40 × 2048
GaitSet	25 × 2048	30 × 2048	40 × 2048	123 × 2048
GaitLGL	25 × 2048	30 × 2048	40 × 2048	105 × 2000

D. SETUP

The batch size is set depending on the dataset and the network architectures, as presented in Table 1. The margin m for the triplet-type loss was 0.2, whereas that for the contrastive-type loss was 256. The network structure parameters of GaitSet and GaitGL were set depending on the dataset in the same manner as in their respective original papers [3], [4]. Adam optimizer is used for training and selected learning rates between 10^{-2} and 10^{-6}. The number of training iterations is set differently depending on the network architectures and datasets by considering the number of network parameters and number of available training samples in the dataset, as described in Table 2.

As for the evaluation metrics, we consider the equal error rate (EER) for the verification scenario, i.e., the trade-off point between a false acceptance rate and false rejection rate [39]. We also report the standard deviation of the false rejection rate (FRR) at the EER threshold, based on [40].

To mitigate the relatively large variation in accuracies due to the training/test split from a limited number of training/test samples in CASIA, we repeated each experiment 10 times and report the average and FRR standard deviation at the EER over the 10 runs.

E. RESULTS

1) OU-MVLP

We report the EERs averaged over gallery views for the five benchmarks BHCn, BATr, BACn, BACn2, and the conventional Cn, as presented in Tables 3 to 5. The results reveal that our BHCn outperforms the other losses on all architectures and almost all views. GEI + CNN, GaitSet, and GaitGL trained with BHCn have average EERs of 0.66%, 0.50%, and 0.53%, respectively.

2) CASIA-B

We compare the accuracies for the same four benchmarks on OU-MVLP, as listed in Tables 6 to 8. The BHCn outperforms the other losses on all settings for GEI + CNN, while it achieves second place for GaitSet and GaitGL. We consider the results are derived from a trade-off among the number of network parameters (i.e., the model complexity), the number of sampled pairs in each loss, and the training set size. GEI + CNN is a simpler network than GaitSet and GaitGL (i.e., the number of parameters is smaller), and hence the proposed BHCn, which uses a smaller number of sampled pairs than BACn, still works well. By contrast, GaitSet and GaitGL have more parameters, and hence batch all variants, which uses the full combination of pairs in a batch, works better, in particular for the CASIA dataset, where the training set size is smaller than that of OU-MVLP.

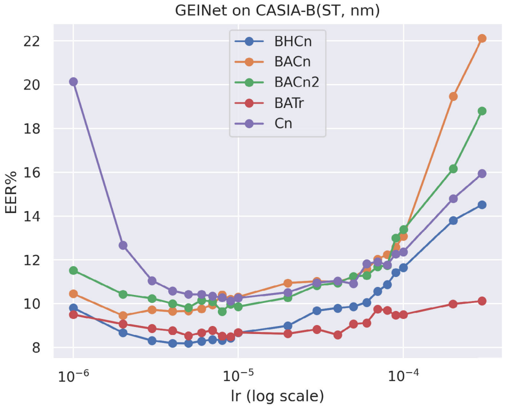


FIGURE 3. Sensitivity analysis of the learning rate on EERs using GEI + CNN on the CASIA-B ST setting.

F. ABLATION STUDIES OF THE SAMPLING METHODS

To verify that identity-based batch sampling is a better option for verification than other well-established batch-sampling methods, we conducted ablation studies on it. We show the EERs of BHCn with sample-based batch sampling and with identity-based batch sampling (the proposed method) in Table 9. We confirmed that identity-based batch sampling performs better than sample-based batch sampling.

G. SENSITIVITY ANALYSIS OF THE BATCH SIZE

The batch hard strategy used in our loss dramatically reduces the number of active samples (i.e., the samples that directly contribute to the loss function and update the network parameters). For P identities in a batch, our batch hard method selects $P(P+1)/2$ active pairs, regardless of the number K of samples per identity. As shown in Table 10, for the same number of samples per identity K, EERs steadily increase as the number of identities is doubled (i.e., $PK = 512$) and quadrupled (i.e., $PK = 1,024$). Moreover, we notice that $K = 8$ is a reasonable choice given the total samples PK on average because we achieve the best or the second-best EERs under the fixed total samples PK.

H. SENSITIVITY ANALYSIS OF THE LEARNING RATE

Because the learning rate impacts performance, we analyzed the sensitivity of the learning rate for GEI + CNN with the CASIA-B, ST setting. As shown in Fig. 3, most methods have

TABLE 3. EERs [%] averaged over gallery views on OU-MVLP using GEI + CNN. lr: learning rate, SD: standard deviation. The bold font indicates the best accuracy. Underlined font indicates the second best accuracy. These conventions are common for the following tables.

Loss (lr)	Probe view														mean ± SD
	0°	15°	30°	45°	60°	75°	90°	180°	195°	210°	225°	240°	255°	270°	
Cn (10^{-3})	2.50	1.83	1.33	1.34	1.55	1.41	1.44	2.21	1.87	1.46	1.44	1.61	1.45	1.60	1.65 ± 0.174
BATr (10^{-4})	1.63	0.98	0.60	0.59	0.77	0.72	0.80	1.34	0.98	0.61	0.62	0.78	0.74	0.86	0.86 ± 0.121
BACn (10^{-3})	1.69	1.15	0.78	0.75	0.89	0.78	0.85	1.35	1.22	0.90	0.89	1.00	0.91	0.97	1.01 ± 0.136
BACn2 (10^{-3})	1.34	0.85	0.56	0.55	0.64	0.55	0.63	1.06	0.93	0.63	0.73	0.75	0.67	0.73	0.76 ± 0.117
BHCn (10^{-3})	**1.17**	**0.72**	**0.48**	**0.47**	**0.59**	**0.52**	**0.57**	**0.93**	**0.77**	**0.55**	**0.57**	**0.64**	**0.57**	**0.63**	**0.66 ± 0.109**

TABLE 4. EERs averaged over gallery views on OU-MVLP using GaitSet.

Loss (lr)	Probe view														mean ± SD
	0°	15°	30°	45°	60°	75°	90°	180°	195°	210°	225°	240°	255°	270°	
Cn (10^{-3})	2.19	1.41	1.00	1.07	1.29	1.07	1.12	2.16	1.57	1.13	1.14	1.25	1.16	1.32	1.35 ± 0.156
BATr (10^{-4})	0.92	0.57	**0.42**	**0.43**	0.54	0.43	0.49	0.80	0.69	0.44	0.51	0.52	0.46	0.52	0.55 ± 0.100
BACn (10^{-3})	0.96	0.56	0.47	0.51	0.58	0.46	0.51	0.91	0.69	0.55	0.60	0.61	0.55	0.59	0.61 ± 0.106
BACn2 (10^{-3})	1.68	0.99	0.69	0.71	0.86	0.70	0.75	1.65	1.10	0.75	0.78	0.86	0.77	0.94	0.94 ± 0.129
BHCn (10^{-3})	**0.80**	**0.54**	0.43	**0.43**	**0.48**	**0.38**	**0.42**	**0.76**	**0.59**	**0.41**	**0.45**	**0.46**	**0.42**	**0.45**	**0.50 ± 0.096**

TABLE 5. EERs averaged over gallery views on OU-MVLP using GaitGL.

Loss (lr)	Probe view														mean ± SD
	0°	15°	30°	45°	60°	75°	90°	180°	195°	210°	225°	240°	255°	270°	
Cn (2×10^{-4})	1.65	1.09	0.88	0.97	0.95	0.83	0.89	1.21	1.34	1.14	1.16	1.20	1.12	1.19	1.12 ± 0.145
BATr (10^{-4})	0.82	**0.55**	0.45	0.47	0.47	0.39	0.41	0.61	**0.64**	0.53	0.60	0.60	0.55	0.56	0.55 ± 0.101
BATr_CE (10^{-4})	0.83	0.62	0.44	**0.46**	0.46	**0.37**	0.41	0.64	0.67	**0.48**	**0.52**	**0.55**	**0.52**	0.55	0.54 ± 0.099
BACn (2×10^{-4})	1.09	0.72	0.61	0.68	0.60	0.54	0.58	0.77	0.91	0.79	0.82	0.84	0.76	0.78	0.75 ± 0.119
BACn2 (2×10^{-4})	1.28	0.84	0.69	0.73	0.71	0.61	0.66	0.98	1.03	0.84	0.89	0.91	0.84	0.87	0.85 ± 0.126
BHCn (2×10^{-4})	**0.77**	0.62	**0.43**	0.49	**0.44**	0.39	**0.40**	**0.58**	0.66	0.53	0.57	**0.55**	**0.52**	**0.52**	**0.53 ± 0.100**

TABLE 6. EERs averaged over all view pairs for CASIA-B using GEI+ CNN.

Setting Loss (lr)	ST(100)	MT(62)	LT(52)
Cn (10^{-5})	10.48 ± 1.16	6.76 ± 1.14	6.01 ± 1.21
BATr (10^{-4})	9.80 ± 1.08	6.53 ± 1.08	5.49 ± 1.08
BACn (10^{-5})	10.93 ± 1.20	7.26 ± 1.18	6.54 ± 1.27
BACn2 (10^{-5})	11.06 ± 1.20	7.85 ± 1.25	6.98 ± 1.33
BHCn (10^{-5})	**8.66 ± 1.06**	**5.28 ± 0.99**	**4.46 ± 1.02**

TABLE 7. EERs averaged over all view pairs for CASIA-B using GaitSet.

Setting Loss (lr)	ST(100)	MT(62)	LT(52)
Cn (10^{-3})	4.95 ± 0.80	2.50 ± 0.68	2.01 ± 0.66
BATr (10^{-4})	5.42 ± 0.83	2.95 ± 0.78	2.60 ± 0.77
BACn (10^{-3})	5.84 ± 0.89	2.60 ± 0.69	2.14 ± 0.69
BACn2(10^{-3})	**4.58 ± 0.77**	**2.22 ± 0.64**	**1.87 ± 0.63**
BHCn (10^{-3})	4.90 ± 0.80	2.33 ± 0.64	1.91 ± 0.64

TABLE 8. EERs averaged over all view pairs for CASIA-B using GaitGL.

Setting Loss (lr)	ST(100)	MT(62)	LT(52)
Cn (10^{-3})	4.77 ± 0.76	2.76 ± 0.74	2.25 ± 0.75
BATr (10^{-4})	6.67 ± 0.96	3.33 ± 0.85	2.84 ± 0.89
BATr_CE (10^{-4})	5.78 ± 0.85	3.14 ± 0.81	2.80 ± 0.88
BACn (10^{-3})	4.90 ± 0.78	**2.63 ± 0.72**	**2.07 ± 0.72**
BACn2 (10^{-3})	5.11 ± 0.78	3.44 ± 0.82	2.74 ± 0.83
BHCn (10^{-3})	**4.56 ± 0.75**	2.72 ± 0.73	2.20 ± 0.73

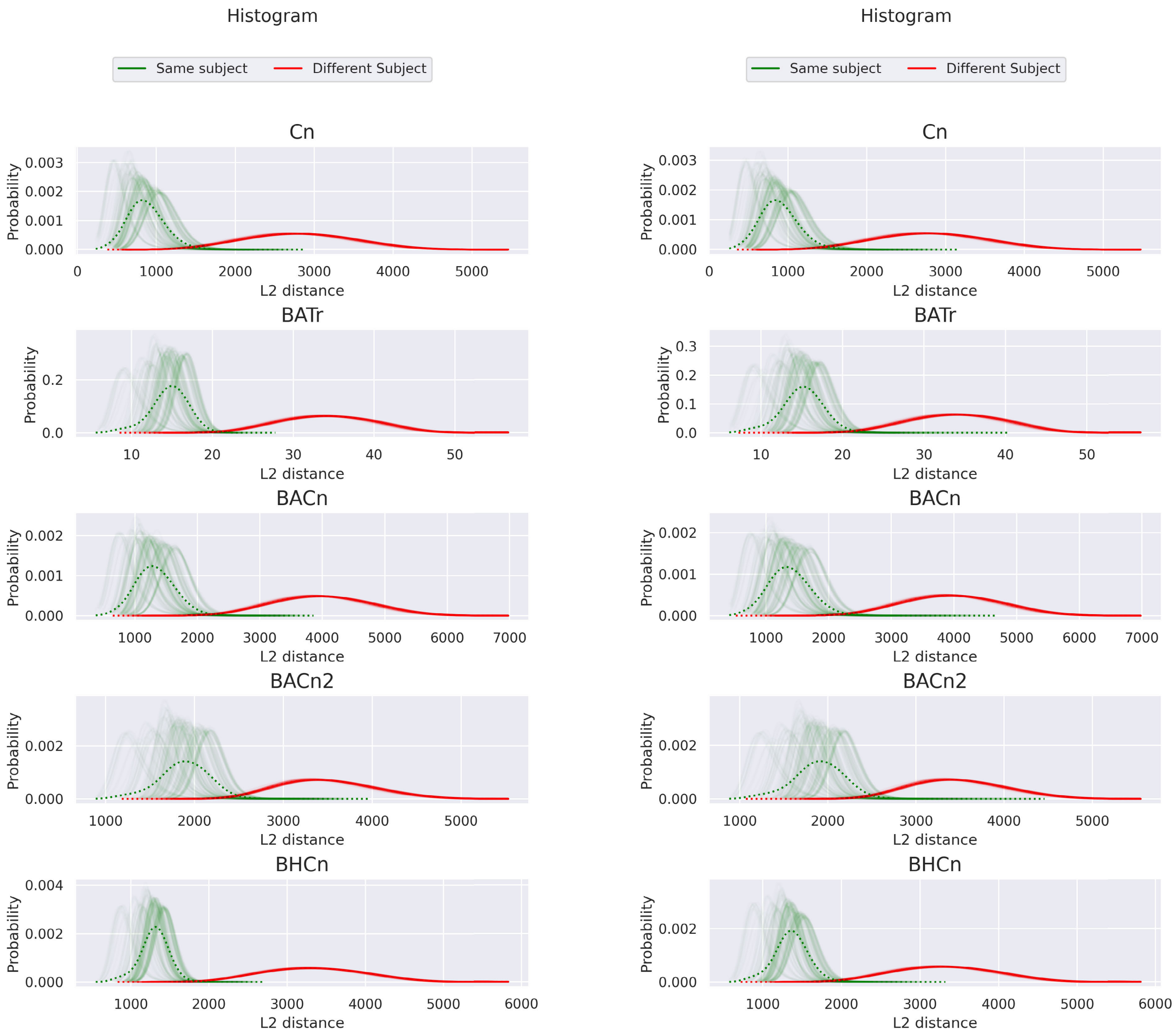
TABLE 9. Ablation studies of sampling methods for contrastive-type loss. Averaged EERs [%] on OU-MVLP are reported.

Method \ Network	GEI + CNN	GaitSet
Sample-based sampling	0.80	0.60
Identity-based sampling (proposed)	**0.66**	**0.50**

the best accuracy for learning rates between 10^{-6} and 10^{-5}. Although the BATr loss seems less sensitive to the learning rate than the proposed BHCn, the BHCn still performs better for learning rates between 10^{-6} and 10^{-5}.

I. DISTANCE DISTRIBUTIONS

To evaluate how well each loss trains the network to be robust against view angles, we compared the L2 distance distribution for each view–angle pair on the training data of the OU-MVLP dataset using the GaitSet network, as shown in Fig. 4a. As a result, the distributions for the proposed BHCn are more consistent among view angle pairs (i.e., the distributions overlap well) than the other loss (i.e., the distributions are diverse). Moreover, to determine how well the embedding can generalize to unseen subjects, we show the L2 distance distributions on the testing data in Fig. 4b. As a result,



(a) Training data

(b) Test data

FIGURE 4. Distance distributions of positive and negative pairs. For the positive pairs, semi-transparent solid lines indicate the distribution for each view angle pair, whereas the dashed lines indicate the distribution averaged over view angle pairs (a) Training data. (b) Test data.

TABLE 10. Sensitivity analysis of the batch size on EERs (mean ± SD) using GEI + CNN on the OU-MVLP dataset. *P* and *K* indicate the numbers of subjects and samples per subject in a batch, respectively. For each *K*, the number *P* of identities is doubled and quadrupled for *PK* = 512 and *PK* = 1, 024 as *PK* = 256.

#Samples per identity K	#Total samples PK		
	256	512	1,024
2	0.90 ± 0.12	0.85 ± 0.12	0.80 ± 0.12
4	$\mathbf{0.80} \pm 0.12$	0.76 ± 0.11	0.70 ± 0.11
8	0.82 ± 0.12	$\mathbf{0.72} \pm 0.11$	$\mathbf{0.67} \pm 0.11$
16	0.94 ± 0.13	0.77 ± 0.11	0.71 ± 0.11
32*	1.23 ± 0.14	1.08 ± 0.14	0.92 ± 0.12

similar to the training data, the proposed BHCn exhibits better consistency among view angle pairs than the other losses.

V. DISCUSSION

This section discusses the advantage of the BHCn over its batch all counterpart.

A. CONTRASTIVE LEARNING: BATCH ALL

At the start of training, the embeddings lie randomly in the embedding space, and any loss can bring improvements overall. As the learning continues, most training data will be clustered appropriately, and thus their contrastive loss will be insignificant. The positive pairs will contribute the most to the training loss in the BACn – fewer negative pairs contribute to the loss because of the hinge. Thus, the batch all approach has an imbalance between positive and negative cases, which leads to rigid updates – the loss ensures the stability of the positive pairs and refuses to change for any negative ones.

B. CONTRASTIVE LEARNING: BATCH HARD

By contrast, batch hard sampling reduces rigidness by reducing constraints. The hard case has a more considerable loss for the negative pairs than the positive pairs; it is enough to push an embedding out of a stale configuration.

There is a trade-off between the rigid but stable batch all sampling and the flexible but less stable batch hard sampling. The batch all method requires a good starting point, making it less likely to become stuck in a local minimum. In contrast, the batch hard method requires much care in parameter selection, such as the number of identities per batch and number of samples per identity. Most important of all is the learning rate, as the learning rate controls the strength of the deformation

VI. CONCLUSION

In this paper, we proposed investing more research effort into gait-based verification scenarios because of their equal importance to identification scenarios in security applications. Owing to the massive interest in identification in the research community, high performance is attainable by modifying the components so that they are more suitable for verification. Because the loss function is tightly coupled with the target task, we proposed BHCn to train deep embedding networks on verification tasks. This loss can replace the batch all (or hard) triplet loss not just on gait recognition, but on any task. We applied our proposed loss to cross-view gait recognition using the OU-MVLP and CASIA datasets. The experimental results demonstrated our proposal's superior performance compared with that of traditional verification loss, and we achieved state-of-the-art performance using GaitSet.

In this paper, we focused on the cross-view challenge of gait recognition because it impacts other challenges. However, many other challenges exist, such as clothing, carrying status, and walking speed. In addition, state-of-the-art networks such as GaitSet are designed with the identification task in mind. Redesigning such a network for verification, coupled with our loss design, is a promising direction for future work.

ACKNOWLEDGMENT

The authors would like to thank Kimberly Moravec, Ph.D., from Edanz (https://jp.edanz.com/ac) for editing a draft of this manuscript.

REFERENCES

[1] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *J. Mach. Learn. Res.*, vol. 10, no. 2, pp. 207–244, Jun. 2009.

[2] N. Takemura, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "On input/output architectures for convolutional neural network-based cross-view gait recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 9, pp. 2708–2719, Sep. 2019, doi: 10.1109/TCSVT.2017.2760835.

[3] H. Chao, K. Wang, Y. He, J. Zhang, and J. Feng, "GaitSet: Cross-view gait recognition through utilizing gait as a deep set," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3467–3478, Jul. 2022.

[4] B. Lin, S. Zhang, and X. Yu, "Gait recognition via effective global-local feature representation and local temporal aggregation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 14648–14656.

[5] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," 2017, *arXiv:1703.07737*.

[6] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: A deep quadruplet network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 1320–1329.

[7] C.-Y. Wu, R. Manmatha, A. J. Smola, and P. Krahenbuhl, "Sampling matters in deep embedding learning," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 2859–2867.

[8] H. O. Song, Y. Xiang, S. Jegelka, and S. Savarese, "Deep metric learning via lifted structured feature embedding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 4004–4012.

[9] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, New York, NY, USA, Jun. 2006, pp. 1735–1742.

[10] S. Li and H. Ma, "A Siamese inception architecture network for person re-identification," *Mach. Vis. Appl.*, vol. 28, no. 7, pp. 725–736, Oct. 2017.

[11] Y. Yuan, K. Yang, and C. Zhang, "Hard-aware deeply cascaded embedding," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 814–823.

[12] R. Gao, F. Yang, W. Yang, and Q. Liao, "Margin loss: Making faces more separable," *IEEE Signal Process. Lett.*, vol. 25, no. 2, pp. 308–312, Feb. 2018.

[13] J. Hu, J. Lu, and Y.-P. Tan, "Discriminative deep metric learning for face verification in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 1875–1882.

[14] J. Žbontar and Y. LeCun, "Stereo matching by training a convolutional neural network to compare image patches," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2287–2318, Jan. 2018.

[15] C. Bailer, K. Varanasi, and D. Stricker, "CNN-based patch matching for optical flow with thresholded Hinge embedding loss," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 2710–2719.

[16] D. Zhong and J. Zhu, "Centralized large margin cosine loss for open-set deep palmprint recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 6, pp. 1559–1568, Jun. 2020.

[17] F. Radenović, G. Tolias, and O. Chum, "CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples," in *Proc. Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, 2016, pp. 3–20.

[18] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua, and F. Moreno-Noguer, "Discriminative learning of deep convolutional feature point descriptors," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 118–126.

[19] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 815–823.

[20] H. Shi, Y. Yang, X. Zhu, S. Liao, Z. Lei, W. Zheng, and S. Z. Li, "Embedding deep metric for person re-identification: A study against large variations," in *Proc. Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, 2016, pp. 732–748.

[21] S. Kong and C. Fowlkes, "Recurrent pixel embedding for instance grouping," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 9018–9028.

[22] A. Newell, Z. Huang, and J. Deng, "Associative embedding: End-to-end learning for joint detection and grouping," in *Proc. Adv. Neural Inf. Process. Syst.,* Long Beach, CA, USA, 2017, pp. 1–11.

[23] Z. Gao, Y. Song, I. McLoughlin, P. Li, Y. Jiang, and L.-R. Dai, "Improving aggregation and loss function for better embedding learning in end-to-end speaker verification system," in *Proc. Interspeech*, Graz, Austria, Sep. 2019, pp. 361–365.

[24] K. Sohn, "Improved deep metric learning with multi-class N-pair loss objective," in *Proc. Adv. Neural Inf. Process. Syst.*, Barcelona, Spain, 2016, pp. 1–9.

[25] N. Takemura, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition," *IPSJ Trans. Comput. Vis. Appl.*, vol. 10, no. 1, pp. 11–14, Feb. 2018.

[26] S. Yu, D. Tan, and T. Tan, "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, Hong Kong, 2006, pp. 441–444.

[27] C. Yuan, J. Guo, P. Feng, Z. Zhao, Y. Luo, C. Xu, T. Wang, and K. Duan, "Learning deep embedding with mini-cluster loss for person re-identification," *Multimedia Tools Appl.*, vol. 78, no. 15, pp. 21145–21166, Aug. 2019.

[28] J. Lezama, Q. Qiu, P. Muse, and G. Sapiro, "OLE: Orthogonal low-rank embedding, a plug and play geometric loss for deep learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 8109–8118.

[29] F. Wang, X. Xiang, J. Cheng, and A. L. Yuille, "NormFace: L_2 hypersphere embedding for face verification," in *Proc. 25th ACM Int. Conf. Multimedia*, Mountain View, CA, USA, Oct. 2017, pp. 1041–1049.

[30] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "SphereFace: Deep hypersphere embedding for face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 6738–6746.

[31] W. Liu, Y. Wen, Z. Yu, and M. Yang, "Large-margin softmax loss for convolutional neural networks," in *Proc. Conf. Mach. Learn.*, New York, NY, USA, 2016, pp. 507–516.

[32] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive margin softmax for face verification," *IEEE Signal Process. Lett.*, vol. 25, no. 7, pp. 926–930, Jul. 2018.

[33] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "CosFace: Large margin cosine loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 5265–5274.

[34] J. Deng, J. Guo, J. Yang, N. Xue, I. Kotsia, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 5962–5979, Oct. 2022.

[35] J. Man and B. Bhanu, "Individual recognition using gait energy image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 2, pp. 316–322, Feb. 2006.

[36] K. Shiraga, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "GEINet: View-invariant gait recognition using a convolutional neural network," in *Proc. Int. Conf. Biometrics (ICB)*, Halmstad, Sweden, Jun. 2016, pp. 1–8.

[37] C. Fan, Y. Peng, C. Cao, X. Liu, S. Hou, J. Chi, Y. Huang, Q. Li, and Z. He, "GaitPart: Temporal part-based model for gait recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14213–14221.

[38] Y. Fu, Y. Wei, Y. Zhou, H. Shi, G. Huang, X. Wang, Z. Yao, and T. Huang, "Horizontal pyramid matching for person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, Honolulu, HA, USA, 2019, pp. 8295–8302.

[39] *Information Technology—Biometric Performance Testing and Reporting—Part 1: Principles and Framework*, Standard ISO/IEC 19795-1:2006, 2006.

[40] A. J. Mansfield and J. L. Wayman, "Best practices in testing and reporting performance of biometric devices," Nat. Phys. Lab., London, U.K., Tech. Rep. CMSC 14/02, Aug. 2002.

MOHAMAD AMMAR ALSHERFAWI ALJAZAERLY received the B.S. degree in information technology engineering from Damascus University, Syria, in 2016, and the M.S. degree in computer science from Osaka University, Japan, in 2020, where he is currently pursuing the Ph.D. degree in computer science. His research interests include computer vision, image processing, deep learning, and gait recognition.

YASUSHI MAKIHARA received the B.S., M.S., and Ph.D. degrees in engineering from Osaka University, in 2001, 2002, and 2005, respectively. He was a specially appointed Assistant Professor (full-time), an Assistant Professor, and an Associate Professor with the Institute of Scientific and Industrial Research, Osaka University, in 2005, 2006, and 2014, respectively. He is currently a Professor with the Institute for Advanced Co-Creation Studies, Osaka University. His research interests include computer vision, pattern recognition, and image processing, including gait recognition, pedestrian detection, morphing, and temporal super-resolution. He is a member of IPSJ, IEICE, RSJ, and JSME. He received several honors and awards, including the 2nd International Workshop on Biometrics and Forensics (IWBF 2014), the IAPR Best Paper Award, the 9th IAPR International Conference on Biometrics (ICB 2016), the Honorable Mention Paper Award, the 28th British Machine Vision Conference (BMVC 2017), Outstanding Reviewers, the 11th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2015), Outstanding Reviewers, the 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017), Outstanding Reviewers, and the Commendation for Science and Technology by the Minister of Education, Culture, Sports, Science and Technology, Prizes for Science and Technology, Research Category, in 2014. He served as a Program Co-Chair for the 4th Asian Conference on Pattern Recognition (ACPR 2017) and an Area Chair for the ICCV 2019, the CVPR 2020, and the ECCV 2020. He served as a Reviewer for journals, such as IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Transactions on Image Processing, IEEE Transactions on Circuits and Systems for Video Technology, IEEE Transactions on Information Forensics and Security, *International Journal of Computer Vision*, and *Pattern Recognition*, and international conferences, such as CVPR, ICCV, ECCV, ACCV, ICPR, and FG. He has served as the Associate Editor-in-Chief for *IEICE Transactions on Information and Systems* and an Associate Editor for *IPSJ Transactions on Computer Vision and Applications* (CVA).

DAIGO MURAMATSU (Member, IEEE) received the B.S., M.E., and Ph.D. degrees in electrical, electronics, and computer engineering from Waseda University, Tokyo, Japan, in 1997, 1999, and 2006, respectively. He is currently a Professor with the Faculty of Science and Technology, Seikei University. His current research interests include gait recognition, signature verification, and biometric fusion. He is a member of IEICE and IPSJ.

YASUSHI YAGI (Senior Member, IEEE) received the Ph.D. degree from Osaka University, in 1991. In 1985, he joined the Product Development Laboratory, Mitsubishi Electric Corporation, where he worked on robotics and inspections. He became a Research Associate, in 1990, a Lecturer, in 1993, an Associate Professor, in 1996, and a Professor, in 2003, with Osaka University. He was also the Director of the Institute of Scientific and Industrial Research, Osaka University, from 2012 to 2015, and the Executive Vice President of Osaka University, from 2015 to 2019. He is currently a Professor with the Institute of Scientific and Industrial Research, Osaka University. His research interests include computer vision, medical engineering, and robotics. He is a fellow of IPSJ and a member of IEICE and RSJ. He received the ACM VRST 2003 Honorable Mention Award, the IEEE ROBIO 2006 Finalist for T. J. Tan Best Paper in Robotics, the IEEE ICRA 2008 Finalist for Best Vision Paper, the MIRU 2008 Nagao Award, and the PSIVT 2010 Best Paper Award. He has served as the Chair for many international conferences, such as the Financial Chair for FG 1998 and PSVIT 2009, an Organizing Chair for OMINVIS 2003, the Program Co-Chair for ROBIO 2006 and ACPR 2011, the Program Chair for ACCV 2007, the Technical Visit Chair for ICRA 2009, and the General Chair for ACCV 2009 and ACPR 2013. He has also served as the Editor for the IEEE ICRA Conference Editorial Board, from 2007 to 2011. He is an Editorial Member of *IJCV* and the Editor-in-Chief of *IPSJ Transactions on Computer Vision and Applications*.