WaveLoss: An Adaptive Dynamic Loss for Deep Gait Recognition

Zicheng Wang, Qiuxia Wu*

South China University of Technology
smallsquirrel007@gmail.com, qxwu@scut.edu.cn

Abstract

Designing an appropriate loss function can enhance the discriminative power on gait recognition. However, previous research focuses on improving network structure and enriching input modalities but overlooks the loss functions. Although transferring loss functions from face recognition can address sample-level loss, additional design is needed for part-level loss. Therefore, we have designed a new loss function called Waveloss, aimed at adaptively and dynamically changing the preference for parts of different difficulties. First, the previous method treats the loss of different parts equally, which brings the problems of difficult convergence or susceptibility to noise interference, so we propose norm-fusion to adaptively learn samples of different difficulties. Additionally, since we find the exponential value represents preference for learning different samples, we introduce the Dynamic Learning Process, which dynamically adjusts the exponential value during iteration to focus on samples of varying difficulties at different training stages. Finally, as the changes of the exponential value leads to significant fluctuations in the gradient, we introduce the gradient truncation and normalization to avoid getting trapped in local optima and gradient vanishing or exploding by adaptively adjusting the gradient. Experimental results demonstrate that our proposed Waveloss achieves state-of-the-art performance on various gait recognition datasets and can improve the performance of different backbones as well.

Introduction

In recent years, there has been rapid development in deep gait recognition, with numerous methods (Chao et al. 2019; Hou et al. 2020; Lin et al. 2021; Dou et al. 2022; Fan et al. 2023a) being proposed. Some (Lin et al. 2021; Fan et al. 2023a; Dou et al. 2023a; Ma et al. 2023) focus on improving network structures, while others (Meng et al. 2020; Zheng et al. 2022a, 2023; Shen et al. 2023; Cui et al. 2023; Fan et al. 2024) concentrate on enriching input modalities. These approaches enable deep learning models to extract discriminative gait features better, thereby increasing intra-class similarity and reducing inter-class similarity.

As mentioned above, the recent deep gait recognition methods have demonstrated commendable performance across various datasets, but there has been a notable lack of research on loss functions. They simply combine triplet loss and softmax loss to pull intra-class samples and push inter-class samples. However, triplet loss and softmax loss suffer from some drawbacks. For instance, Arcface (Deng et al. 2019) highlighted the issue of combinatorial explosion with triplet loss, leading to slow convergence, while softmax loss is not only challenging to open-set recognition but also leads to weak feature discriminability. Simply adding them together may not necessarily be complementary. Therefore, we aim to propose a more suitable loss function for deep gait recognition.

Directly transferring the loss functions from face recognition may be a feasible solution. Through this approach, sample-level losses can be integrated. However, the horizontal pooling (Chao et al. 2019) before the loss function in gait recognition results in calculating intra-class and inter-class similarities for each different part of different samples separately. Therefore, we also need to integrate the losses of different parts in part-level.

The traditional part fusion approach(Chao et al. 2019; Lin et al. 2021; Fan et al. 2023a) adopts an ostrich strategy, wherein after obtaining losses from different parts, they are directly averaged through pooling, referred to as post-fusion. Figure1 illustrates the gradients of each part. As shown in Figure1 (c)(d), in the later stages of training, most parts reside in regions of zero gradients, while a few challenging parts are located in gradient regions. Despite the model's efforts to learn from difficult parts, noisy parts make it hard to learn useful features. Typically, the solution to distinguishing between low-quality samples and difficult samples involves image quality assessment. However, defining image quality for binary gait images is difficult compared to RGB images, and introducing complex image quality assessment makes the loss function more complex.

Another intuitive and straightforward method is to perform fusion pooling on the parts obtained from horizontal pooling before calculating the loss. This helps remove additional part dimensions generated by horizontal pooling. Subsequently, the loss is directly obtained through the loss function, referred to as pre-fusion. As shown in Figure 1 (a)(b), pre-fusion aggregates all parts before computing the loss, so the gradient of all parts is the gradient of the centroid. The advantage of pre-fusion is that it enhances robustness against

*Corresponding author.

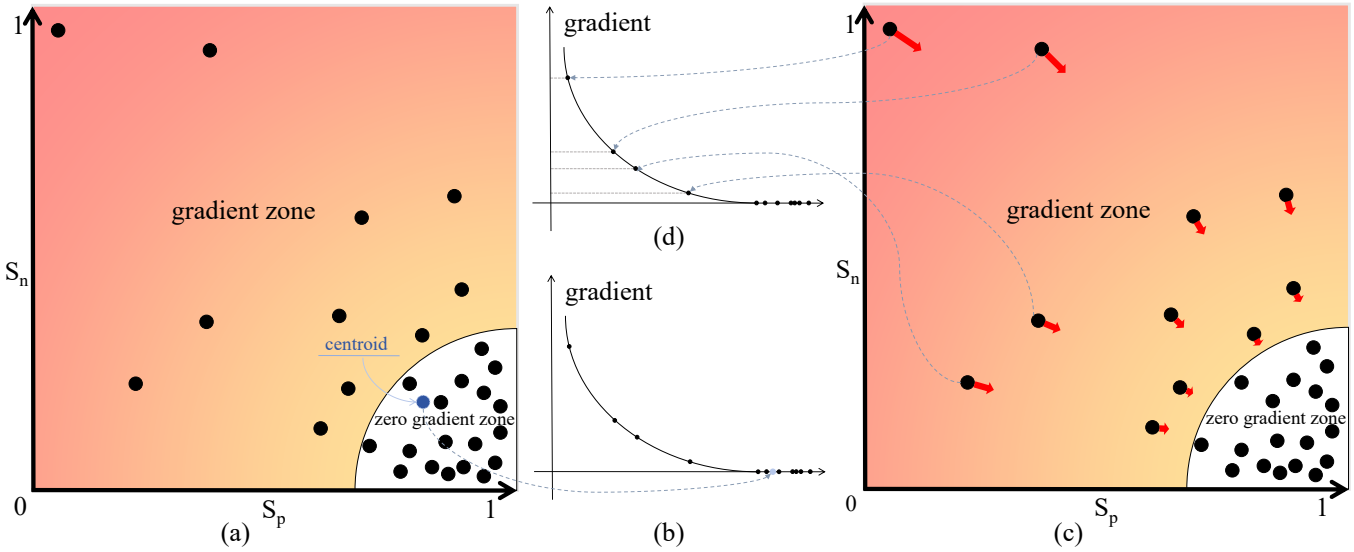


Figure 1: The toy scenario (only a single S_n and S_p of each part) of the gradient of each part (black points) during the backpropagation for pre-fusion and post-fusion. S_n and S_p represent the intra-class similarity and inter-class similarity of parts respectively. (a)(b) Pre-fusion aggregates the gradients of all part points' centroids (blue point) and equally backpropagates them to all part points. This causes the model to get trapped in local optima and fail to converge in the later stages of training. (c)(d) Post-fusion aggregates the gradients (red arrows) of all parts before backpropagation, focusing on the samples that are not converged. Additionally, gradients become larger farther from the convergence region, making the model more susceptible to unidentifiable samples.

noisy samples. However, due to the margin of the loss function, it is easy to be trapped in local optima during the later stages of training.

Therefore, a natural idea is to combine the advantages of pre-fusion and post-fusion to create a new loss function. As shown in Formula 3, we find the two ends of it precisely correspond to pre-fusion and post-fusion. Inspired by it, we propose norm-fusion as an intermediate form of two expressions. Norm-fusion obtains the base weights by applying a nonlinear transformation followed by l_2-norm on the loss of each part. We take l as the exponent of the base weight. The norm loss can be obtained by calculating the weighted sum of the weight and the original part's loss. When $l = 0$, it represents the form of post-fusion. When $l < 0$, it reduces the gradients of samples with high loss to avoid interference from noise, thus achieving the advantages of pre-fusion.

Furthermore, based on the viewpoint of dynamic learning(Bengio et al. 2009; Dou et al. 2022), learning different preferences for samples of varying difficulty during different training stage can effectively enhance model efficiency and accuracy. To this end, we propose the Dynamic Learning Process, where the exponential l is adjusted during training to enable the model to learn more effectively. Since the base weight is sensitive to l, inspired by ISR (Dou et al. 2023b), we introduce gradient normalization to compensate for the scaling effect of different exponential l on the loss function. We also introduce the gradient truncation to ensure convergence. Experimental results demonstrate that our proposed loss function achieves state-of-the-art performance on current benchmark datasets. Furthermore, our loss function brings improvements on different backbones as well.

In summary, the contributions of this paper include:

- To combine the advantages of part-level pre-fusion and post-fusion, we propose norm-fusion as an intermediate form between the post-fusion and pre-fusion. This allows us to obtain a function whose exponential l can be adjusted to adaptively focus on parts with varying levels of difficulty.
- To enable the model to focus on parts of varying difficulty at different stages, we introduce the Dynamic Learning Process, which allows to dynamically adjust its preference for different parts by change the exponential l, thereby enhancing learning efficiency and performance.
- To prevent the gradient explosion or vanishing and to avoid getting stuck in local optima, we propose gradient normalization and truncation to reduce the sensitivity of the loss to the exponential l.
- Experimental results demonstrate that our proposed Waveloss achieves state-of-the-art performance on different gait datasets. Moreover, integrating Waveloss into other backbones can also improve performance.

Related Work

Loss function in metric learning. In recent years, many loss functions in metric learning have been proposed, such as CosFace(Wang et al. 2018), ArcFace(Deng et al. 2019), AdaFace(Kim et al. 2022), CircleLoss(Sun et al. 2020). These loss functions are designed to increase intra-class similarity and decrease inter-class similarity through hard ex-

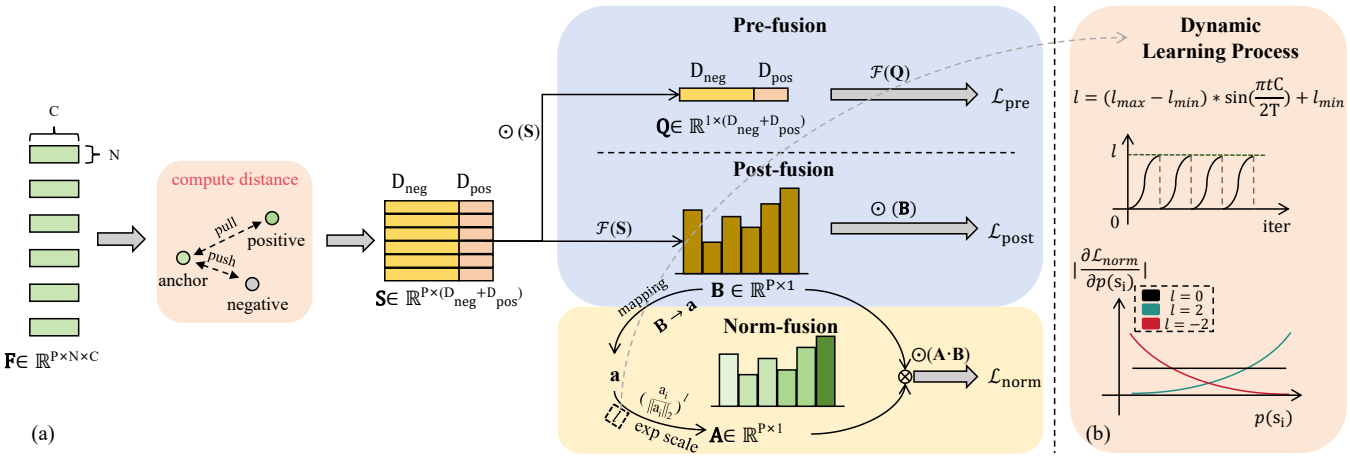


Figure 2: (a) Diagram of different fusion methods for parts. After horizontal pooling, each batch will obtain the part-level features $\mathbf{F}$. By calculating the distance between each sample, the distance vector $\mathbf{S}$ can be obtained for each part. The Pre-fusion and Post-fusion adopt different methods to obtain the loss $\mathfrak{L}_{pre}$ and $\mathfrak{L}_{post}$. Our Waveloss is calculated by weighting the part-level vector $\mathbf{B}$ using the exponent value l to obtain the final loss $\mathfrak{L}_{norm}$. (b) Diagram of the Dynamic Learning Process. The exponent value l is changed through the sine strategy during iteration. The change in exponent value l affects the preference of the model for parts of different difficulty.

ample mining or changing the distance metric. Despite their prevalence in related fields, none of them have been explored for use in gait recognition. This could be attributed to the relatively scarce research in gait recognition. Besides, the final horizontal pooling operation in gait recognition brings about unique challenges when it comes to incorporating such loss functions. This is, the introduction of extra dimensions in gait recognition makes it arduous to transplant loss functions that have been successful in face recognition. This paper aims to introduce a part-level loss function based on the sample-level loss function from face or gait recognition and improve it according to the characteristics of gait recognition.

Deep gait recognition. Since the proposal of GaitSet(Chao et al. 2019), many deep gait recognition representation methods(Lin et al. 2021; Dou et al. 2022; Fan et al. 2023a) have been introduced, which have made many improvements in spatiotemporal representation. For example, from aspects such as local and global (Lin et al. 2021), long-term and short-term (Huang et al. 2021), dynamic and static (Wang et al. 2023), they have achieved remarkable results. However, these methods basically use triplet loss and cross-entropy as loss functions, and they do not focus on studying metric loss functions. Due to the long development of loss functions in face recognition, this paper draw on the loss functions in face recognition and gradually refine them into loss functions tailored for gait recognition.

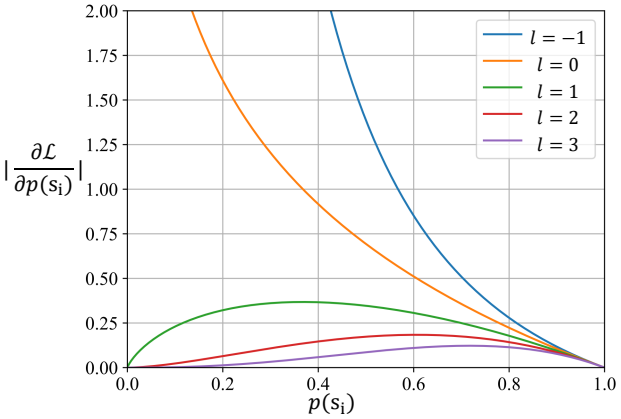
Dynamic learning method. The dynamic learning method can be divided into curriculum learning (Bengio et al. 2009) and hard sample mining (Shrivastava et al. 2016), focusing respectively on the learning processes of starting with easy and progressing to difficult tasks and starting with difficult and progressing to easy tasks. GaitMPL (Dou et al. 2022) and CurriculumFace (Huang et al. 2020) propose methods based on curriculum learning to converge models better, applied respectively to gait recognition and face recognition. Hard sample mining implies that difficult samples are rich in information. As a result, methods for hard sample mining lay more emphasis on difficult samples to extract more discriminative information. Nevertheless, this often brings about the problem that distinguishing between hard samples and low-quality samples is difficult and often requires the introduction of additional image quality assessment. This paper simply defines the difficulty level of samples based on the magnitude of the loss values. By adaptively and dynamically adjusting the model's preference for samples of different difficult, it ingeniously combines the advantages of curriculum learning and hard sample mining.

Proposed Approach

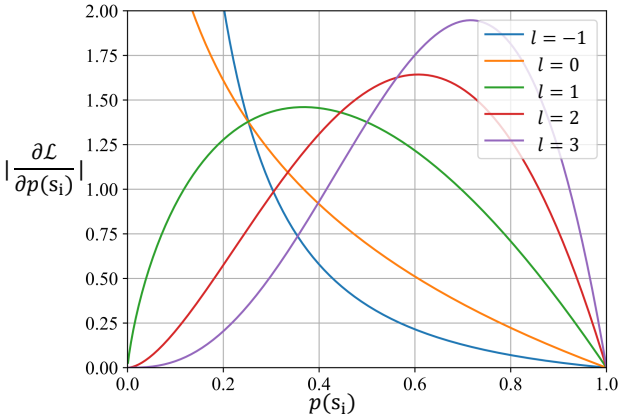
Revisit Part Fusion on Deep Gait Recognition

As shown in Figure 2, in gait recognition, input gait sequences are processed through a backbone network and horizontal pooling (Chao et al. 2019), resulting in part-level features $\mathbf{F} \in \mathbb{R}^{\mathrm{P}\times\mathrm{N}\times\mathrm{C}}$, where P, N, and C represent the number of parts, the batch size and the channel dimension, respectively. $\mathbf{F}$ are used to calculate the distance $\mathbf{S} = \{\mathbf{s_1}, \mathbf{s_2}, ..., \mathbf{s_P}\} \in \mathbb{R}^{\mathrm{P}\times(\mathrm{D_{pos}}+\mathrm{D_{neg}})}$ for each pair of samplers, where $\mathrm{D_{pos}}$ and $\mathrm{D_{neg}}$ represent the number of distances for positive and negative sample pairs, respectively. Previous gait recognition methods employ the post-fusion strategy for $\mathbf{S}$. In this approach, the distance obtained for each part is first pooled using a distance fusion function $\mathcal{F}(\cdot)$. The part features $\mathbf{B} \in \mathbb{R}^{\mathrm{P}\times 1}$ are aggregated by summing or averaging function $\odot(\cdot)$ to gain the final loss $\mathfrak{L}_{post}$. The formula is as follows:

$$\mathfrak{L}_{post} = \odot(\mathcal{F}(\mathbf{S})). \tag{1}$$



(a) The curve of the $\mathfrak{L}_{norm}$ without the gradient normalization.



(b) The curve of the $\mathfrak{L}_{norm}$ with the gradient normalization.

Figure 3: (a) The loss without the standard strategy. During the Dynamic Learning Process, the norm value l is adjusted, causing significant fluctuations in the loss function values. When $l \leq 0$, the loss function values for low-confidence samples are very high, and when $l > 0$, the loss function values for low-confidence samples are very low. (b) Scaling the loss function by standard strategy, which does not fluctuate dramatically with change in the l.

Another method of fusing parts, as shown in Figure 2, is called pre-fusion. The calculated distance S for sampled pairs undergoes part pooling before passing through the distance fusion function $\mathcal{F}$, ultimately yielding the pre-fusion loss $\mathfrak{L}_{pre}$. The formula is as follows:

$$\mathfrak{L}_{pre} = \mathcal{F}(\odot(\mathbf{S})). \tag{2}$$

The two fusion methods exactly correspond to the two ends of Formula 3 when the number x_i is replaced with the distance $\mathbf{s_i}$, where $i \in \{1, 2, ..., \mathrm{P}\}$ and the function f is replaced with the distance fusion function $\mathcal{F}$.

Jensen's inequality. For a real convex function f, numbers $x_1, x_2, ..., x_n$ in its domain, Jensen's inequality can be stated as:

$$f(\frac{\sum_i x_i}{n}) \leq \frac{\sum_i f(x_i)}{n}. \tag{3}$$

For the pre-fusion, which is left of the Formula 3, it computes the average of the feature of all part points to obtain the gradient. As illustrated in Figure 1 (a)(b), each part's point in the gradient space corresponds to a gradient value. Pre-fusion calculates the centroid of all points and updates parameters based on the centroid's gradient. Although it can avoid the influence of outliers, for margin-based loss functions, once the majority of part points fall into the zero-gradient region, the centroid will also fall into that, resulting in zero gradient and preventing parameter updates.

The post-fusion, the right end of the Formula 3, continues to learn hard samples even in the later stages, as shown in Figure 1 (c)(d), making it more discriminative compared to the pre-fusion. However, as the existence of low-quality samples, the model struggles to learn useful information, leading to unstable convergence in the later stages of training. Therefore, we aim to integrate the advantages of both ends of the Formula 3 (*i.e.* pre-fusion and post-fusion) to find a more discriminative loss function. To address this issue, we propose part norm-fusion.

Part Norm Fusion

We define the norm-fusion function as:

$$\mathcal{G}_{norm}(\mathbf{s_1}, \mathbf{s_2}, ..., \mathbf{s_P}, l) = \sum_{i}^{\mathrm{P}} (\frac{a_i}{||a_i||_2})^l \mathcal{F}(\mathbf{s_i}), \tag{4}$$

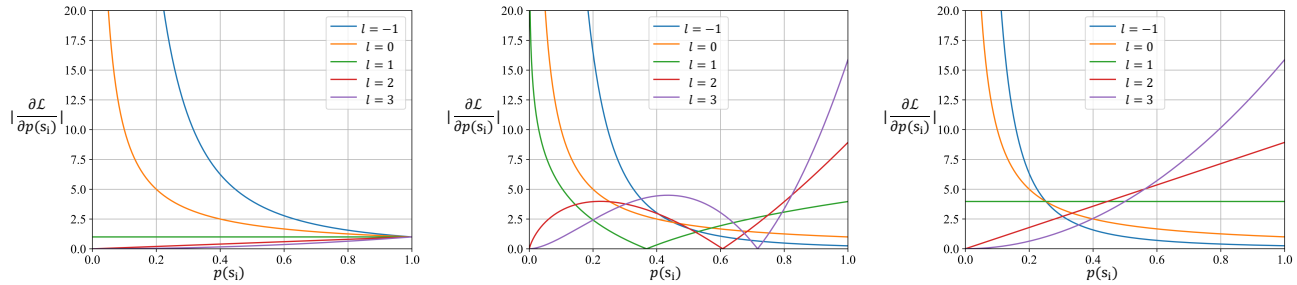
where l is the exponential value, $||\mathbf{a_i}||_2$ represents the normalization of weight a_i by l_2-norm and $\mathbf{s_i} \in \mathbb{R}^{(\mathrm{S_{pos}}+\mathrm{S_{neg}})}$ is the distance vector of each part.

Through the Formula 4, we can adjust the value of the weight to influence the learning of the model. The application of l_2-norm to a_i serves the purpose of restricting excessive fluctuations in its value, which in turn facilitates more efficient model training. When $l > 0$, the magnitude relationship among each weight will not change. However, when $l < 0$, it will be reversed. Consequently, we can achieve the effect of controlling weights by changing l.

Furthermore, we impose restrictions on a_i. In order to perform weighted calculations adaptively, we define $a_i = \exp(\frac{\mathcal{F}(\mathbf{s_i})}{||\mathcal{F}(\mathbf{s_i})||_2})$. Similarly, the l_2-norm is for the sake of training stability, and the exponentiation with e is to increase the difference in weights. We define $p(\mathbf{s_i}) = \frac{1}{a_i^{||\mathcal{F}(\mathbf{s_i})||_2}}$, where $p(\mathbf{s_i}) \in (0, 1)$ obviously, as a difficult sample judgment function because it is negatively correlated with the part level loss function $\mathcal{F}$, indicating that the larger the loss, the smaller its value, and the more difficult the sample is.

With this definition of a_i, we can adaptively obtain the value of a_i. And when $l = 0$, Formula 4 is equivalent to post-fusion. When $l < 0$, the weight is more inclined to the parts with smaller loss values, which can reduce the interference of the noisy part. In addition, by adjusting l, we can make the model focus on parts with different difficulty levels. As l increases, the model will pay more attention to difficult parts, that is, parts with higher loss values, and vice versa.

However, it is difficult for us to possess the advantages of both pre-fusion and post-fusion simultaneously. Therefore,



(a) The gradient curve without the standardization.

(b) The gradient curve without the truncation.

(c) The gradient curve with the standardization and truncation.

Figure 4: (a) Without the standardization, the gradients will fluctuate dramatically with changes in the exponent l. For low-confidence samples, this can result in gradient explosion and vanishing. (b) Without truncation, the gradient is not monotonic, which becomes zero at certain points. This can easily lead to getting stuck in local optima, affecting convergence. (c) With normalization and truncation, modifying the exponent l can adjust the preference of parts with different difficulties.

we introduce the Dynamic Learning Process so that the loss function can have different advantages at different training stages.

Dynamic Learning Process

Inspired by curriculum learning and hard sample mining, learning samples with different difficulty levels at different stages can effectively improve efficiency and performance. Therefore, we propose the Dynamic Learning Process that dynamically adjusts the exponent l. By scaling the weights, the model can focus on parts of different difficulty levels at different stages.

We define l_{min} as the minimum exponent value and l_{max} as the maximum exponent value. As the number of iterations increases, the exponent value l changes between l_{min} and l_{max}. The formula is as follows:

$$l = (l_{max} - l_{min})\sin(\frac{\pi t \mathrm{C}}{2\mathrm{T}}) - l_{min}, \tag{5}$$

where t is the current iteration, T is the total iteration and C represents the number of cycles. As shown in Figure 2 (b), l exhibits periodic changes during iterations, allowing the model to switch between difficult sample preferences and simple sample preferences multiple times, which is more conducive to its learning.

However, the Dynamic Learning Process introduces that the value of the loss is highly sensitive to the change of l, as shown in Figure 3 (a). The gradient curve without the normalization is shown in Figure 4 (a), extremely large or small gradients can easily cause gradient explosion or vanishing. Besides, as shown in Figure 4 (b), it is also easy to fall into a local optimum, which is unfavorable for model training.

To address this, inspired by ISR (Dou et al. 2023b), we first introduce the gradient truncation strategy, where the gradients of the weights are excluded from backpropagation. Comparing Figure 4 (b) and 4 (c), after applying this strategy, the gradients become smoother, and there will no longer be extreme points, ensuring that the weights do not affect the convergence of the model. The formula is as follows:

$$\mathfrak{L}_t = -(\frac{a_i}{||a_i||_2})^l_{\not\leftarrow} \log(p(\mathbf{s_i})), \tag{6}$$

where $(\cdot)_{\not\leftarrow}$ represents the gradient truncation applied to $(\cdot)$, which does not participate in backpropagation.

Subsequently, we introduce the gradient normalization strategy. By normalizing and scaling the loss function values, we can reduce the variance of gradient values across different parts and facilitate controlling the step size of gradient descent with a single learning rate. The formula for loss function value normalization is as follows:

$$\mathfrak{L}_{norm} = \frac{m}{\mathrm{N}} \sum_{i=0}^{\mathrm{N}} \mathfrak{L}_t(\mathbf{s_i}), \tag{7}$$

where $m = (\sum_{i=0}^{\mathrm{N}} \mathfrak{L}_{l=0}(\mathbf{s_i})) / \sum_{i=0}^{\mathrm{N}} \mathfrak{L}_t(\mathbf{s_i}))$ .This helps us adaptively calibrate the gradient step size so that the different exponent values l affect only the preference for parts of different difficulties, rather than the step size of the gradients. As shown in Figure 3 (b), Because the shape of the loss function changes like waves when the exponent l changes, we named our loss function Waveloss.

Experiments

Dataset

To assess the effectiveness of the Waveloss, we will evaluate it on three popular gait datasets, including two indoor datasets, CASIA-B (Yu et al. 2006) and OU-MVLP (Takemura et al. 2018), and one outdoor dataset, GAIT-3D (Zheng et al. 2022b).

CASIA-B. The CASIA-B dataset stands as one of the most widely used repositories for gait analysis, comprising data from 124 subjects. Each subject's data in the CASIA-B dataset consists of 10 sets of gait sequences, captured under three different conditions: normal walking (NM), walking with a bag (BG), and walking in coats (CL). These sequences are labeled NM#01-06, BG#01-02, and CL#01-02, respectively. Furthermore, each set contains 11 videos, each filmed from various view angles ranging from 0°to 180°.

OUMVLP. The OUMVLP dataset stands as one of the most extensive repositories for gait recognition, boasting a vast collection of data from 10,307 subjects. Each subject's data comprises two sets of gait sequences labeled Seq#00 and Seq#01. Within each set, gait sequences are captured from

Method	Probe View														Mean
	0°	15°	30°	45°	60°	75°	90°	180°	195°	210°	225°	240°	255°	270°	
GaitSet	79.5	87.9	89.9	90.2	88.1	88.7	87.8	81.7	86.7	89.0	89.3	87.2	87.8	86.2	87.1
GaitGL	84.9	90.2	91.1	91.5	91.1	90.8	90.3	88.5	88.6	90.3	90.4	89.6	89.5	88.5	89.7
DANet	87.7	91.3	91.6	91.8	91.7	91.4	91.1	90.4	90.3	90.7	90.9	90.5	90.3	89.9	90.7
DeepGaitv2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	91.9
Ours	**89.7**	**92.9**	**92.8**	**93.0**	**94.0**	**92.1**	**91.9**	**92.0**	**92.5**	**92.2**	**92.3**	**92.2**	**91.3**	**91.0**	**92.1**

Table 1: Rank-1 accuracy on OUMVLP under all view angles, excluding the identical-views cases.

	Method	Prob View											Mean
		0°	18°	36°	54°	72°	90°	108°	126°	134°	162°	180°	
NM	GaitSet	90.8	97.9	99.4	96.9	93.6	91.7	95.0	97.8	98.9	96.8	85.8	95.0
	GaitGL	96.0	98.3	99.0	97.9	96.9	95.4	97.0	98.9	99.3	98.8	94.0	97.4
	GaitMPL	97.2	98.3	**99.5**	98.3	96.5	95.2	97.2	98.5	99.1	98.5	93.8	97.5
	DyGait	97.4	98.9	99.2	98.3	97.7	96.8	98.2	99.3	99.3	99.2	97.6	98.4
	Ours	**98.5**	**99.2**	99.3	**98.7**	**98.4**	**98.1**	**98.9**	**99.4**	**99.4**	**99.7**	**97.7**	**98.7**
BG	GaitSet	83.8	91.2	91.8	88.8	83.3	81.0	84.1	90.0	92.2	94.4	79.0	87.2
	GaitGL	92.6	96.6	96.8	95.5	93.5	89.3	92.2	96.5	98.2	96.9	91.5	94.5
	GaitMPL	92.3	96.8	95.8	95.4	93.8	89.5	92.5	95.8	98.4	97.1	91.8	94.5
	DyGait	94.5	96.9	97.4	96.1	**95.4**	**94.0**	94.8	97.6	98.5	97.7	**94.9**	96.2
	Ours	**95.2**	**97.4**	**97.9**	**96.7**	94.9	93.6	**95.1**	**98.5**	**98.9**	**98.0**	94.0	**96.4**
CL	GaitSet	61.4	75.4	80.7	77.3	72.1	70.1	71.5	73.5	73.5	68.4	50.0	70.4
	GaitGL	76.7	90.0	90.3	87.1	84.5	79.0	84.1	87.0	87.3	84.4	69.5	83.6
	GaitMPL	**83.8**	85.5	96.3	**93.0**	**89.8**	83.2	87.3	**92.3**	91.4	88.7	76.7	88.0
	DyGait	82.2	93.0	**95.2**	91.6	87.1	**87.1**	87.2	90.1	**92.4**	88.2	75.8	87.8
	Ours	83.3	**93.1**	95.0	91.0	89.5	86.4	**88.4**	89.7	**92.4**	**91.7**	**78.7**	**89.0**

Table 2: Rank-1 accuracy on CASIA-B under all view angles, excluding the identical-views cases.

14 different angles, ranging from 0°to 90°and from 180°to 270°, with intervals of 15°.

Gait3D. The Gait3D dataset is a substantial repository comprising data from 4,000 subjects, with over 25,000 sequences captured from an unconstrained indoor scene using 39 cameras. Notably, it offers 3D Skinned Multi-Person Linear(SMPL) models recovered from video footage. To ensure compatibility with other algorithms, the dataset is partitioned into a training set containing 3,000 subjects and a test set comprising 1,000 subjects.

Implementation Details

We preprocess the original gait sequences by normalizing the size of each frame to 64×44 for CASIA-B, OU-MVLP, and Gait3D datasets. Unless otherwise specified, our experimental settings follow the settings of the OpenGait (Fan et al. 2023b). All of our experiments were conducted on the GeForce RTX 2080 Ti or 2080.

On the CASIA-B dataset, we use GaitGL as the backbone network with channel sizes set to [64, 128, 256] and circle loss as the base loss function. The number of iterations is set to 10,000, with a learning rate of 1e-4. We employ the MultiStepLR strategy, reducing the learning rate by a factor of 10 at iterations 80,000 and 95,000. We utilize the Adam optimizer with a weight decay of 5e-4. Data augmentation techniques such as random flipping and occlusion are also applied. Cosine similarity is used for probe matching.

On the OUVMLP dataset, we employ DeepGaitV2 as the backbone network. A triplet loss with a margin of 0.2 and cross-entropy are used as the base loss functions. The number of iterations is 150,000. The learning rate is set to 0.1, with a weight decay of 5e-4. We set the batch size to [32, 8]. Data augmentation strategies include perspective transformations, flipping, and rotation. Probe matching is performed using Euclidean distance.

On the Gait3D dataset, we utilize DeepGaitV2 as the backbone network. We employ a triplet loss with a margin of 0.2 and cross-entropy as the base loss functions. The learning rate is set to 0.1, with a weight decay of 5e-4. The total number of iterations is 120,000, and we adopt the MultiStepLR strategy, where the learning rate is reduced by a factor of 1/10 at iterations 40,000, 80,000, and 100,000. Data augmentation strategies include perspective transformations, flipping, and rotation. Probe matching is performed using Euclidean distance.

Comparison with State-of-the-art Methods

In this section, we will compare our Waveloss with other state-of-the-art methods on different datasets.

Evaluation on OU-MVLP. Table 1 shows the experimental results of our Waveloss on the OUMVLP indoor dataset. Our method achieves state-of-the-art (SOTA) performance on this dataset, demonstrating its excellent capability in cross-view feature extraction on large-scale datasets. Compared

Methods	Rank-1	Rank-5	mAP	mINP
GaitSet	36.7	58.3	30.0	17.3
GaitGL	29.7	48.5	22.3	13.3
SMPLGait	46.3	64.5	37.2	22.2
DyGait	66.3	80.8	56.4	37.3
DeepGaitv2	74.4	88.0	65.8	-
Ours	**75.6**	**88.4**	**66.5**	**46.5**

Table 3: Rank-1 accuracy (%), Rank-5 accuracy (%), mAP (%) and mINP on the Gait3D dataset.

	l	NM	BG	CL	Mean
Static	-1	98.7	96.1	88.3	94.4
	0	98.6	96.1	87.9	94.2
	1	98.6	96.2	88.4	94.4
Dynamic	[-1,0]	98.4	96.3	**89.0**	94.6
	[0,1]	98.2	96.0	88.9	94.4
	[-1,1]	**98.7**	**96.4**	**89.0**	**94.7**

Table 4: Ablation study on the exponent value l.

with the baseline method DeepGaitV2, our Waveloss also achieves an improvement of 0.2%.

Evaluation on CASIA-B. Table 2 shows the experimental results of our Waveloss on the CASIA-B dataset. It can be observed that compared to other models, ours achieve state-of-the-art performance. Specifically, we achieve accuracy rates of 98.7%, 96.4%, and 89.0% on NM, BG, and CL, respectively, with an average accuracy of 94.7%. Besides, our Waveloss achieves SOTA, with higher accuracy on challenging samples (CL), which demonstrates that our loss function can better increase intra-class similarity and reduce inter-class similarity. Our proposed norm-fusion strategy makes Waveloss can better extract features from difficult samples.

Evaluation on Gait3D. As shown in Table 3, compared to other methods, our Waveloss has achieved the state-of-the-art. Specifically, we achieve 75.6%, 88.4%, 66.5%, and 46.5% on Rank-1, Rank-5, mAP, and mINP, respectively. This is attributed to the adaptability of our loss function in distinguishing parts. For low-quality outdoor datasets, different parts exhibit varying quality differences. Our Waveloss can dynamically and adaptively focus on important parts. Therefore, Our Waveloss can lead to greater improvements on the low-quality dataset.

Ablation Study

Effect of Hyperparamemter l. To thoroughly investigate the effects of the Dynamic Learning Process, we conducted comparative experiments with different parameter values for l, as shown in Table 4, dividing them into static and dynamic groups. We found that when l is fixed at values such as -1, 0, or 1, the performance is not as good as with the dynamic approach (94.4%, 94.2%, and 94.4% v.s. 94.6%, 94.4%, and 94.7%). This is because the model focuses only on samples of the same difficulty level during training, whereas the dynamic process can adapt to samples of varying difficulty levels, thereby better capturing distinguishable features and resulting in improved performance.

Methods	NM	BG	CL	Mean
Pre-fusion	92.7	84.7	63.7	80.4
Post-fusion	98.6	96.1	87.9	94.2
Norm-fusion	**98.7**	**96.4**	**89.0**	**94.7**

Table 5: Comparison of different fusion ways on CASIA-B dataset.

Methods	NM	BG	CL	Mean
Softmax Loss	98.1	93.9	80.6	90.9
Triplet Loss	**98.7**	**96.9**	86.8	94.1
Circleloss	98.6	96.1	87.9	94.2
Circleloss+Norm	98.8	96.3	88.1	94.4
Circleloss+DLP	98.7	96.2	88.3	94.4
Ours	**98.7**	96.4	**89.0**	**94.7**

Table 6: Ablation study on the loss function.

Effect of Part Fusion Method. As shown in Table 5, we explored the impact of different fusion methods on performance and found that using the pre-fusion strategy led to gradients becoming small in the later stages of training. This impeded the model from learning the features of hard samples, and the average accuracy on the CASIA dataset was only 80.4Although the effect on simple samples (NM) is excellent, reaching 92.7%, the effect on difficult samples is rather poor, only 63.7%. Pre-fusion has difficulty learning the features of difficult parts, and this phenomenon is in line with our previous judgment. Norm fusion led to a 0.5% improvement compared to post-fusion and increased accuracy on challenging samples (CL) by 1.1%. This indicates that the reweighting approach of norm fusion significantly enhances the model's ability to discriminate difficult samples.

Effect of Norm Fusion and Dynamic Learning Process. As shown in Table 6, we conducted ablation experiments comparing our proposed norm fusion and the Dynamic Learning Process to validate the effectiveness of the two modules. By adding our proposed modules to Circleloss, the average accuracy on the CASIA-B dataset increased by 0.2% and 0.2%, respectively. This demonstrates the effectiveness of our modules. Additionally, the accuracy on challenging samples (CL) improved by 1.1%, indicating that our proposed modules can better handle difficult samples.

Conclusions

This paper introduces an adaptive dynamically loss function for part-level fusion called Waveloss. Unlike traditional sample-level loss functions, our Waveloss fills the gap in the research of part-level loss fusion. Compared to traditional post-fusion and pre-fusion, our Waveloss has several advantages in terms of performance without additional training parameters. It shows improvements under different datasets and different backbone networks. We hope that our proposed loss function can provide a new perspective for gait recognition and garner more attention to the research of the loss function.

References

Bengio, Y.; Louradour, J.; Collobert, R.; and Weston, J. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, 41–48.

Chao, H.; He, Y.; Zhang, J.; and Feng, J. 2019. Gaitset: Regarding gait as a set for cross-view gait recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 8126–8133.

Cui, Y.; and Kang, Y. 2023. Multi-modal gait recognition via effective spatial-temporal feature fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17949–17957.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4690–4699.

Dou, H.; Zhang, P.; Su, W.; Yu, Y.; Lin, Y.; and Li, X. 2023a. Gaitgci: Generative counterfactual intervention for gait recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5578–5588.

Dou, H.; Zhang, P.; Zhao, Y.; Dong, L.; Qin, Z.; and Li, X. 2022. Gaitmpl: Gait recognition with memory-augmented progressive learning. *IEEE Transactions on Image Processing*.

Dou, Z.; Wang, Z.; Li, Y.; and Wang, S. 2023b. Identity-seeking self-supervised representation learning for generalizable person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15847–15858.

Fan, C.; Hou, S.; Huang, Y.; and Yu, S. 2023a. Exploring deep models for practical gait recognition. *arXiv preprint arXiv:2303.03301*.

Fan, C.; Liang, J.; Shen, C.; Hou, S.; Huang, Y.; and Yu, S. 2023b. Opengait: Revisiting gait recognition towards better practicality. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9707–9716.

Fan, C.; Ma, J.; Jin, D.; Shen, C.; and Yu, S. 2024. SkeletonGait: Gait Recognition Using Skeleton Maps. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 1662–1669.

Hou, S.; Cao, C.; Liu, X.; and Huang, Y. 2020. Gait lateral network: Learning discriminative and compact representations for gait recognition. In *European conference on computer vision*, 382–398. Springer.

Huang, X.; Zhu, D.; Wang, H.; Wang, X.; Yang, B.; He, B.; Liu, W.; and Feng, B. 2021. Context-sensitive temporal feature learning for gait recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12909–12918.

Huang, Y.; Wang, Y.; Tai, Y.; Liu, X.; Shen, P.; Li, S.; Li, J.; and Huang, F. 2020. Curricularface: adaptive curriculum learning loss for deep face recognition. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5901–5910.

Kim, M.; Jain, A. K.; and Liu, X. 2022. Adaface: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18750–18759.

Lin, B.; Zhang, S.; and Yu, X. 2021. Gait recognition via effective global-local feature representation and local temporal aggregation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 14648–14656.

Ma, K.; Fu, Y.; Zheng, D.; Cao, C.; Hu, X.; and Huang, Y. 2023. Dynamic aggregated network for gait recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22076–22085.

Meng, Z.; Fu, S.; Yan, J.; Liang, H.; Zhou, A.; Zhu, S.; Ma, H.; Liu, J.; and Yang, N. 2020. Gait recognition for co-existing multiple people using millimeter wave sensing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 849–856.

Shen, C.; Fan, C.; Wu, W.; Wang, R.; Huang, G. Q.; and Yu, S. 2023. Lidargait: Benchmarking 3d gait recognition with point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1054–1063.

Shrivastava, A.; Gupta, A.; and Girshick, R. 2016. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 761–769.

Sun, Y.; Cheng, C.; Zhang, Y.; Zhang, C.; Zheng, L.; Wang, Z.; and Wei, Y. 2020. Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6398–6407.

Takemura, N.; Makihara, Y.; Muramatsu, D.; Echigo, T.; and Yagi, Y. 2018. Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. *IPSJ transactions on Computer Vision and Applications*, 10: 1–14.

Wang, H.; Wang, Y.; Zhou, Z.; Ji, X.; Gong, D.; Zhou, J.; Li, Z.; and Liu, W. 2018. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5265–5274.

Wang, M.; Guo, X.; Lin, B.; Yang, T.; Zhu, Z.; Li, L.; Zhang, S.; and Yu, X. 2023. DyGait: Exploiting dynamic representations for high-performance gait recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13424–13433.

Yu, S.; Tan, D.; and Tan, T. 2006. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *18th international conference on pattern recognition (ICPR'06)*, volume 4, 441–444. IEEE.

Zheng, J.; Liu, X.; Liu, W.; He, L.; Yan, C.; and Mei, T. 2022a. Gait recognition in the wild with dense 3d representations and a benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20228–20237.

Zheng, J.; Liu, X.; Liu, W.; He, L.; Yan, C.; and Mei, T. 2022b. Gait recognition in the wild with dense 3d representations and a benchmark. In *Proceedings of the IEEE/CVF*

conference on computer vision and pattern recognition, 20228–20237.

Zheng, J.; Liu, X.; Wang, S.; Wang, L.; Yan, C.; and Liu, W. 2023. Parsing is all you need for accurate gait recognition in the wild. In *Proceedings of the 31st ACM International Conference on Multimedia*, 116–124.