Supplementary Materials - Watch Where You Head: A View-biased Domain Gap in Gait Recognition and Unsupervised Adaptation

1. Evaluation

1.1. GOUDA with view annotations

In the paper, we present GOUDA results on various target datasets. We provide the results of the entire pipeline, including using estimated views extracted from the View Extraction module. However, there are two indoor datasets, CASIA-B and OU-MVLP, that provide ground-truth view annotations in advance. In order to reduce the noise produced by the View Extraction module and purely assess our Triplet Selection algorithm, we present GOUDA results using the view annotations (see Table 1). It is apparent that the Rank-1 improvements are higher, in this case, indicating the potential of GOUDA if using more accurate views. Particularly, the results on OU-MVLP dataset show a significant change. As an example, there is a substantial increase in Rank-1, from 27.9% with estimated views to 40.5% with annotated views, when adapting GaitSet backbone from CASIA-B (source) to OU-MVLP (target).

In OU-MVLP dataset, RGB frames are not available (compared to CASIA-B dataset), and the view estimation is inferred from the quality of the provided 2D keypoints (Fig. 1a), especially compared to the view estimation of CASIA-B (Fig. 1b). These estimation errors might affect the Triplet Selection algorithm. The positive samples, which belong to the *cross view* set, should indeed have different views than the anchor. However, if their views were estimated incorrectly, they might have entered to this set by mistake, and should have actually belong to the *similar view* set. This unintended situation results in bringing closer samples with similar views, causing to not fulfilling the entire potential of GOUDA.

1.2. Applications

Suppose a real-world scenario of a target dataset with varied views but without true-positives (each person appears once). GOUDA can still be applied as it does not assume any conditions on the target labels. To demonstrate it, we create a subset of OU-MVLP dataset with a single appearance of each person and use it to train GOUDA using GaitGL pretrained on GREW. The Rank-1 accuracy improves from 25.7% to 29.3%. In this experiment we used OU-MVLP ground-truth view annotations.

2. Analysis

In this section, we provide additional analysis aspects, expanding the analysis presented in the paper.

2.1. GREW and Gait3D view estimation

Both GREW [8] and Gait3D [7] were captured in the wild. Therefore, their view annotations are not provided. Fig. 2 presents the histograms of the estimated views for these datasets, using the proposed View Extraction module.

2.2. View analysis

Fig. 3 presents the internal results on OU-MVLP test set by the end of GOUDA training. Two different cases are presented. In the first case, GOUDA was trained using OU-MVLP estimated views. In the second case, ground-truth view annotations were utilized for training. As detailed, the view estimation quality of OU-MVLP dataset is insufficient. Consequently, we observe greater improvements in internal Rank-1 accuracies when utilizing view annotations.

2.3. Target view distribution

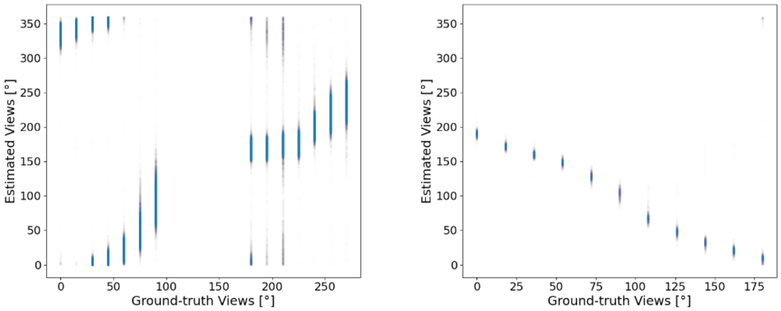
In the paper, we describe a fundamental phenomenon in gait recognition models. As detailed, gait recognition models place a strong emphasis on view-based features in the target domain, exposing the huge gap between source and target domains. The observed behavior is not unique to any particular gait backbone, and is consistent across different backbones that we checked. In this context, we present similar patterns for additional gait recognition models beyond what is covered in the paper (see Figs. 4, 5, and 6). The pattern is less clear (but still exists) when using GaitSet as the backbone (Fig. 5), probably because it perceives the silhouettes as a set of images rather than a temporal sequence, thus assigning less significance to the view information.

2.4. Curriculum Learning

Further ablation experiments on the curriculum learning hyperparameter q (the percentage of top confident valid triplets selected for training in each stage) are reported in

Source Dataset	Backbone	Target Dataset			
		CASIA-B			OU-MVLP
		NM	BG	CL	
CASIA-B	GaitSet		-		**40.5** - 27.9 - 9.6
	GaitPart				**34.9** - 27.5 - 10.8
	GaitGL				**41.4** - 34.0 - 16.2
OU-MVLP	GaitSet	**90.8** - 87.0 - 74.0	**70.4** - 68.0 - 55.5	**29.5** - 27.2 - 16.4	-
	GaitPart	**94.5** - 91.5 - 73.9	**80.6** - 78.4 - 56.9	36.7 - **38.7** - 20.7	
	GaitGL	**94.8** - 92.8 - 81.7	**82.3** - 80.9 - 71.5	38.8 - **44.6** - 28.8	
GREW	GaitSet	**82.1** - 76.9 - 65.6	**61.6** - 56.8 - 44.9	23.9 - **26.0** - 20.8	**49.5** - 38.8 - 21.8
	GaitPart	**85.7** - 81.4 - 69.2	**70.1** - 68.3 - 52.1	33.8 - **33.9** - 25.4	**52.5** - 43.6 - 23.9
	GaitGL	**88.9** - 82.7 - 69.8	**79.3** - 61.1 - 73.2	**45.6** - 44.6 - 31.9	**62.1** - 44.2 - 25.7
Gait3D	GaitSet	**80.2** - 73.4 - 62.8	**58.1** - 51.9 - 45.8	**21.7** - 20.4 - 11.9	**44.5** - 41.9 - 20.8
	GaitPart	**80.1** - 76.7 - 61.8	**65.4** - 61.1 - 47.0	23.0 - **24.5** - 13.7	**37.4** - 32.2 - 19.1
	GaitGL	**82.3** - 79.0 - 63.5	**70.3** - 65.4 - 51.2	**29.6** - 24.9 - 16.3	**40.4** - 37.0 - 23.7

Table 1. Rank-1 accuracy of GOUDA on the target datasets CASIA-B and OU-MVLP [5, 6], compared to direct testing results on the backbones [1], [2], [3]. Here, we present two different setups of GOUDA results, one with estimated views (as presented in the paper), and the other by using ground-truth annotated views. The results of GOUDA with view annotations are reported on the left, GOUDA results with estimated views are in the middle, and the direct testing results are on the right, separated by dashes. In most cases, the best results are achieved by using GOUDA with annotated views, avoiding any noise produced by wrong view estimations or provided 2D keypoints.



(a) OU-MVLP

(b) CASIA-B

Figure 1. View estimation of OU-MVLP [5] and CASIA-B [6] datasets, compared to their ground-truth view annotations provided as meta-data. The view estimation of OU-MVLP is limited due to the inaccurate 2D keypoints provided. Contrary to OU-MVLP, CASIA-B dataset includes RGB images that can be used to estimate accurate 2D keypoints, and therefore its view estimation is better (ambiguity of ±180 degrees is ignored).

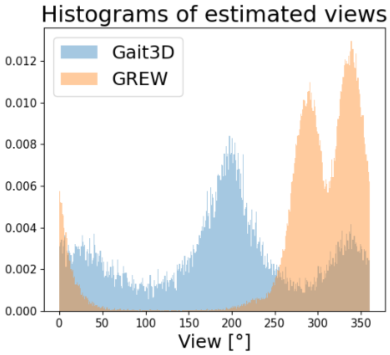


Figure 2. Histograms of estimated views for Gait3D and GREW datasets [7, 8] using the View Extraction module.

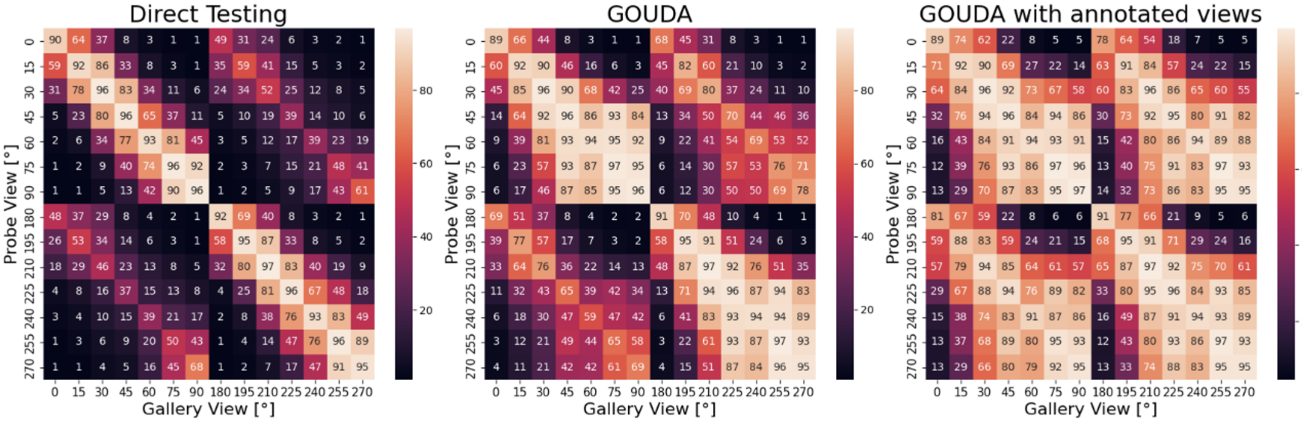


Figure 3. The internal Rank-1 accuracies on the target domain (OU-MVLP) test set. "Direct Testing" is a gait backbone pretrained on the source domain (GREW). Each element in the matrix is the Rank-1 accuracy for the setting in which the probe view is α and the gallery set includes sequences of view β. Brighter color represents higher Rank-1. The final Rank-1 is the average of all Rank-1 accuracies such that $\alpha \neq \beta$. We present higher improvements when using view annotations (on the right) rather than using estimated views (in the middle).

q values [%]	OU-MVLP → CASIA-B		
	NM	BG	CL
5, 25, 70, 100	91.3	76.7	39.7
10, 20, 40, 100	92.4	80.3	47.2
15, 30, 60, 100	90.1	74.1	40.6
20, 50, 75, 100	91.5	76.1	38.1
GOUDA (10, 25, 50, 100)	92.8	80.9	44.6

Table 2. Ablation experiments on the curriculum learning hyperparameter q. In this setting, OU-MVLP is the source domain and CASIA-B is the target domain. Average Rank-1 accuracies across all 11 views, are shown, excluding identical-view cases.

Table 2. These results are inferior or comparable to those achieved with the original q values: 10, 25, 50, and 100.

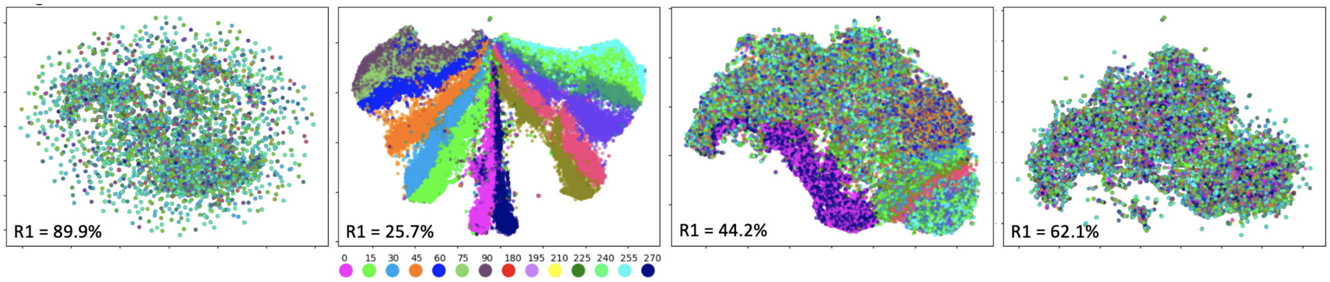
3. Justification

3.1. Positive sample selection

Here, we illustrate the intuition behind our positive sample strategy. Fig. 7 presents a shared visualization of viewing angles (color coded) and identities (shapes) on a subset of the OU-MVLP target dataset. For the sake of visualization, we present only 10 identities. In Fig. 7a, the green dashed curves illustrate the cases in which our hypothesis is satisfied, meaning that the closest samples with different views share the same identity. Conversely, the red-bordered regions showcase "optional" breach of our hypothesis. Our approach effectively functions under "noisy" pseudo labeled data. To this end, our curriculum learning protocol is structured to potentially avoid the choice of anchors from the red-bordered areas. This gradual model enhancement, as advocated by curriculum learning, consequently fosters the selection of anchors from the green areas while contributing to the reduction of red-bordered regions by gradual target adaption. This result is justified by the ramp-up in cross-view R1 accuracy after the adaptation (from 25.7% to 44.2%), and with better identity-based clustering, as demonstrated in Fig. 7b.

3.2. Quantitative evaluation

To further support the triplet selection strategy, we present Table 3. It depicts the percentage of selecting the *correct* triplets, as well as the correct positive samples and negative samples, separately. The positive sample is correct if it shares the same identity as the anchor, whereas the negative sample is correct if it has a different identity. We show two different settings using the initial pre-trained source model (first curriculum learning phase). It is shown that the selection percentage of both positive and negative samples is substantial, even under "noisy" conditions. Moreover, by employing the curriculum learning approach, the model performance on the target domain is improved throughout training, leading to an increase in the proportion of correct triplets (see Fig. 8).



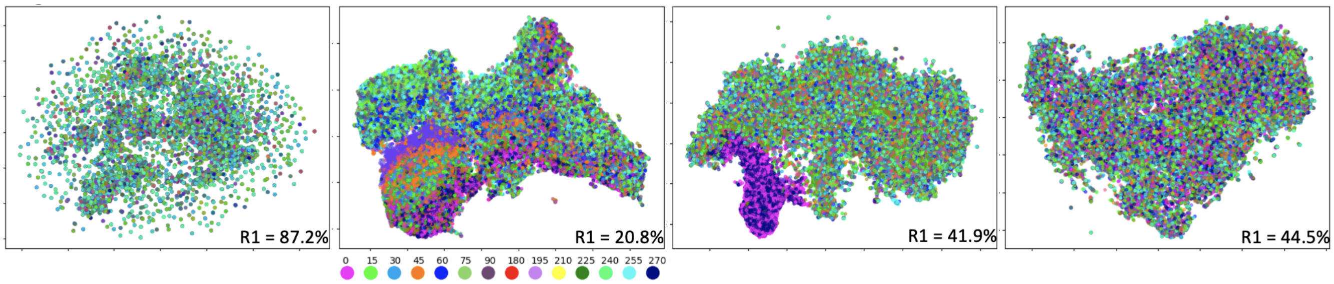
(a) Target = Source

(b) Direct Testing

(c) Direct Testing + GOUDA with estimated views

(d) Direct Testing + GOUDA with annotated views

Figure 4. Gait embedding visualization (UMAP [4]) for domain transfer between GREW as source and OU-MVLP as the target domain using GaitGL. Points are color coded by viewing angle. Fig. 4a presents the single domain scenario, in which the model was trained on OU-MVLP training set. Fig. 4b presents the Direct Testing scenario, in which the model was trained on a different gait dataset (GREW). Fig. 4c presents the Direct Testing scenario after applying GOUDA using estimated views. Fig. 4d presents the Direct Testing scenario after applying GOUDA using the ground-truth annotated views. GOUDA achieves higher improvements when using annotated views.



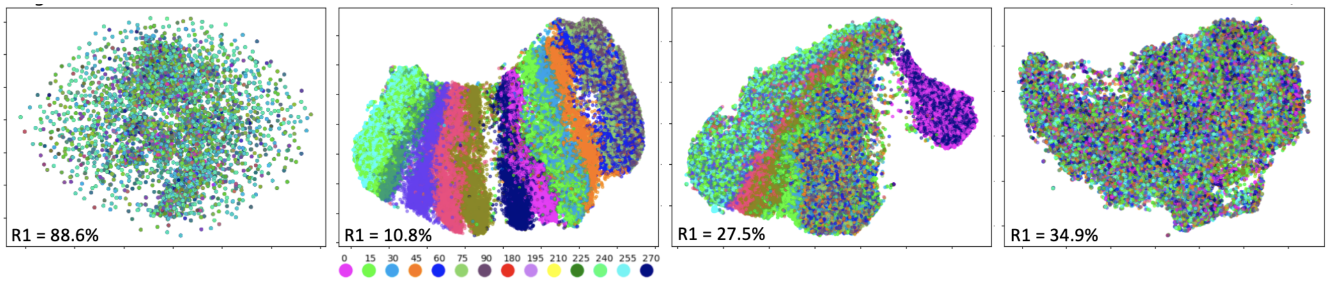
(a) Target = Source

(b) Direct Testing

(c) Direct Testing + GOUDA with estimated views

(d) Direct Testing + GOUDA with annotated views

Figure 5. Gait embedding visualization (UMAP [4]) for domain transfer between Gait3D as source and OU-MVLP as the target domain using GaitSet. Points are color coded by viewing angle. Fig. 5a presents the single domain scenario, in which the model was trained on OU-MVLP training set. Fig. 5b presents the Direct Testing scenario, in which the model was trained on a different gait dataset (Gait3D). Fig. 5c presents the Direct Testing scenario after applying GOUDA using estimated views. Fig. 5d presents the Direct Testing scenario after applying GOUDA using the ground-truth annotated views. GOUDA achieves higher improvements when using annotated views.



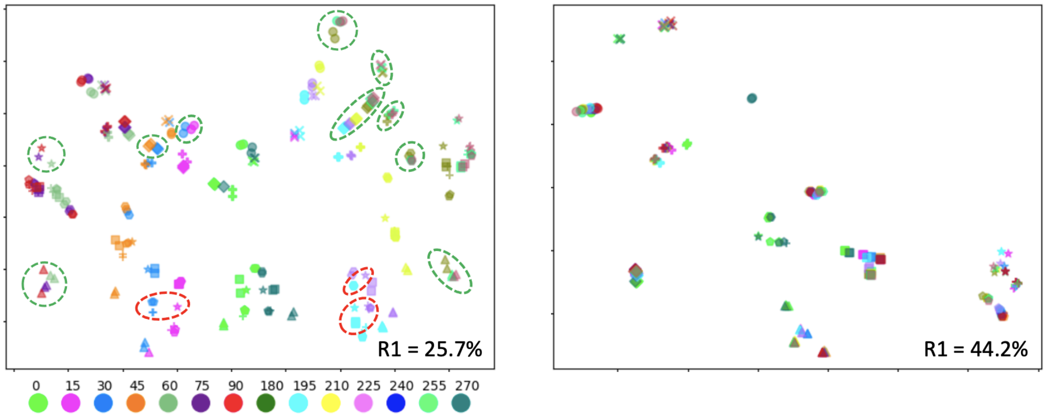
(a) Target = Source

(b) Direct Testing

(c) Direct Testing + GOUDA with estimated views

(d) Direct Testing + GOUDA with annotated views

Figure 6. Gait embedding visualization (UMAP [4]) for domain transfer between CASIA-B as source and OU-MVLP as the target domain using GaitPart. Points are color coded by viewing angle. Fig. 6a presents the single domain scenario, in which the model was trained on OU-MVLP training set. Fig. 6b presents the Direct Testing scenario, in which the model was trained on a different gait dataset (CASIA-B). Fig. 6c presents the Direct Testing scenario after applying GOUDA using estimated views. Fig. 6d presents the Direct Testing scenario after applying GOUDA using the ground-truth annotated views. GOUDA achieves higher improvements when using annotated views.



(a) Direct Testing

(b) Direct Testing + GOUDA

Figure 7. Gait embedding visualization (UMAP [4]) for domain transfer between GREW as source and OU-MVLP as the target domain using GaitGL. Points are color coded by viewing angle while identities are discriminated via the plot shapes. For the sake of visualization, only 10 identities are shown. Fig. 7a presents the Direct Testing scenario, in which the model was trained on a different gait dataset (GREW). The green/red dashed curves present the case in which the closest sample with different view shares the same/different identity, respectively. Fig. 7b presents the distribution after applying GOUDA. Same identities are well clustered, reflected by higher Rank-1 accuracy.

Target Dataset	Triplets	Positives	Negatives
CASIA-B	88.1	98.5	88.9
OU-MVLP	69.4	69.4	100

Table 3. Identity-wise correctness [%] of the selected triplets in the first curriculum learning phase. The correctness of the positive and negative samples is presented as well. A positive sample is correct if it shares the same anchor identity, while a negative sample is correct if it has a different identity. Here, GaitGL is used for domain transfer from GREW (source) to CASIA-B or OU-MVLP (target) using the ground-truth annotated views.

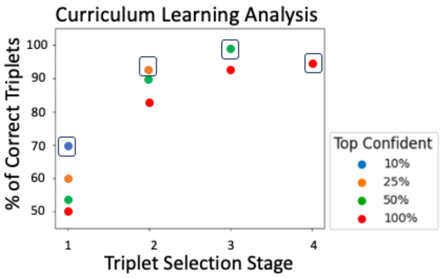


Figure 8. Curriculum Learning Analysis. Triplet Selection is applied 4 times during training (x-axis). Y-axis shows how many of the selected triplets are correct considering person identities. The colors represent the top confident triplets, and in each stage the chosen percentage is marked by a rectangle. Here, GaitGL is used for domain transfer from GREW (source) to OU-MVLP (target) with the ground-truth annotated views.

References

[1] Hanqing Chao, Yiwei He, Junping Zhang, and Jianfeng Feng. Gaitset: Regarding gait as a set for cross-view gait recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8126–8133, 2019. 2

[2] Chao Fan, Yunjie Peng, Chunshui Cao, Xu Liu, Saihui Hou, Jiannan Chi, Yongzhen Huang, Qing Li, and Zhiqiang He. Gaitpart: Temporal part-based model for gait recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14225–14233, 2020. 2

[3] Beibei Lin, Shunli Zhang, and Xin Yu. Gait recognition via effective global-local feature representation and local temporal aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14648–14656, 2021. 2

[4] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 4, 5

[5] Noriko Takemura, Yasushi Makihara, Daigo Muramatsu, Tomio Echigo, and Yasushi Yagi. 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, 2018. 2

[6] Shiqi Yu, Daoliang Tan, and Tieniu Tan. A framework for evaluating the effect of view angle, clothing and car-

rying condition on gait recognition. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 4, pages 441–444, 2006. 2

[7] Jinkai Zheng, Xinchen Liu, Wu Liu, Lingxiao He, Chenggang Yan, and Tao Mei. 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 (CVPR)*, pages 20228–20237, June 2022. 1, 2

[8] Zheng Zhu, Xianda Guo, Tian Yang, Junjie Huang, Jiankang Deng, Guan Huang, Dalong Du, Jiwen Lu, and Jie Zhou. Gait recognition in the wild: A benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14789–14799, October 2021. 1, 2