Gait-X: Exploring X modality for Generalized Gait Recognition Supplementary Material

Zengbin Wang[1] Saihui Hou[2] Junjie Li[1] Xu Liu[3] Chunshui Cao[3]
Yongzhen Huang[2,3] Siye Wang[1] Man Zhang[1,4]*
[1]Beijing University of Posts and Telecommunications [2]Beijing Normal University
[3]WATRIX.AI [4]Qinghai Institute of Technology

Table A1. Ablation of suppression level (γ) in Eq. (3).

Setting	Within-domain (CCPG, Gait)					Cross-domain (to CASIA-B)			
	CL	UP	DN	BG	Mean	NM	BG	CL	Mean
$\gamma = 1$	74.2	81.0	85.7	94.9	84.0	83.7	76.5	26.6	62.3
$\gamma = 1/4$	76.1	81.8	88.0	95.7	85.4	84.4	79.0	26.6	63.3
$\gamma = 1/8$	76.8	83.8	88.2	96.1	86.2	86.2	80.6	25.7	64.2
$\gamma = 1/12$	76.1	83.2	88.8	96.2	86.1	84.9	79.0	26.6	63.5
$\gamma = 1/16$	79.0	84.4	89.0	96.8	**87.3**	85.9	80.0	28.8	**64.9**
$\gamma = 1/20$	77.6	84.3	88.6	96.7	86.8	86.8	81.2	26.1	64.7
$\gamma = 1/24$	77.8	84.8	89.0	96.6	87.1	85.7	80.3	26.7	64.2
$\gamma = 1/32$	77.8	84.4	89.0	96.6	87.0	85.4	80.0	26.6	64.0

A. More Ablations

A.1. Impact of suppression level (γ)

The suppression level (γ) controls the trade-off of high-frequency suppression. As shown in Table A1, the grid search results indicate that γ=$1/16$ is the optimal suppression level for both within-domain and cross-domain evaluations. We set γ=$1/16$ for all gait datasets.

A.2. Ablations beyond standard DCT assumptions

As stated in Sec. 4, we adopt the standard DCT assumptions, *i.e.*, patch size (kernel size: 8) and progressive suppression matrix. The main reason is that standard settings have proven general applicability across various fields, so maintaining consistency is a good choice. Our results in Sec. 5 also demonstrate that this is an initial effective attempt in gait recognition.

Here, we expand our progressive suppression matrix with a learnable one. We find within-domain result remains similar (87.1→87.3, CCPG-Gait, average rank-1), but cross-domain result drops (64.9→62.9, CCPG→CASIA-B, average rank-1), likely because a shared suppression pattern across domains works better. More suppression strategies are worth further exploration.

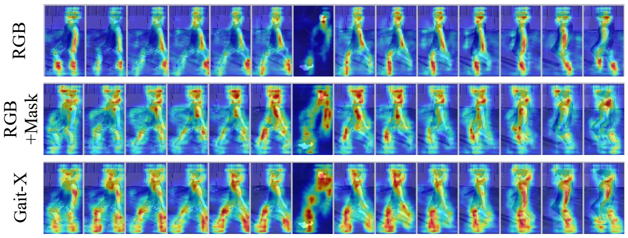

Figure A1. Heatmap visualization across half a gait cycle.

A.3. Computational cost

Regarding GaitBase with RGB as reference (1×, CCPG), the training and inference time of X branch, Decomposition branch, Gait-X are (1.8×, 1.7×, 2.5×) / (1.7×, 1.6×, 3.7×). Training with 8×8×30 frames requires an extra 1.357 GFLOPs.

A.4. Visualization

Figure A1 shows that the RGB demonstrates high responses at key local positions. Including the mask in RGB can spread responses to global information, but weaken the local cues. Our Gait-X achieves both global responses and strong responses at key local positions.

B. More Discussions

B.1. Discussion about the name “X”

In our opinion, “X” indicates something new and powerful to solve specific challenges. Like Marvel’s X-Men, “X” moality can be regarded as a controllable “X” variant of RGB with powerful generalization.

B.2. Discussion about information loss

To a large extent, we remove or suppress unnecessary frequencies based on our empirical analysis and pilot experiments (Fig. 2). The final performance comparison also shows that filtered frequencies contribute less to generaliza-

*Corresponding author (zhangman@bupt.edu.cn).

tion. This is a feasible solution, and I hope it will inspire more exploration.

B.3. Discussion about mask dependence

As stated in Sec. 4.2 and Sec. 4.3, we adopt the mask operation to filter unnecessary background noise in both X-modality and decomposition branches. The reason comes from the fact that the mask is widely used in previous gait research for its background removal and shape variance. Our work follows this setting for a fair comparison.

In some extreme cases, such as the night condition in SUSTech1K, the result is indeed affected by mask quality. However, our X modality can be seen as supplementing rich discriminative information beyond the silhouette. The result in Table 2 shows that it improves performance from 30.3% to 71.9% in the night condition.

B.4. Discussion about the choice of DCT

Other frequency transformation methods (*e.g.*, Fourier and Wavelet) may also work, but we find that: (1) The frequency in the Fourier tends to spread across more coefficients, making it harder to analyze each frequency. (2) Multi-level Wavelet is computationally complex and lacks the expected energy compression. However, the discrete cosine transformation (DCT) offers a good trade-off between energy compaction and computational efficiency. The patch-based DCT facilitates localized frequency analysis, and its progressive suppression strategy is widely used and proven effective in image compression. These properties make DCT well-suited for analyzing the relationship between frequency components and image content.

B.5. Discussion about the expanding capacity to other gait tasks

Our frequency operations stem from empirical studies (Fig. 2), and aim to remove or suppress unnecessary frequencies (or noises) with less information loss for representation learning. This may also benefit other gait tasks with this reduced noise and better representation.

B.6. Discussion about temporal design

As stated in the main paper, we mainly focus on modality exploration. More model structure design (*e.g.*, temporal design) is orthogonal to our goal. We think our spatial-frequency insights can help extend to temporal-frequency modeling, *e.g.*, low-freq captures statics while high-freq indicates motions. Balancing both achieves better temporal modeling. More temporal designs are ongoing.