Occlusion-Aware Part-Level Supervised Contrastive Learning for Skeleton-Based Gait Recognition

Minas Aslanyan[1] and Levon Aslanyan[1]

Institute for Informatics and Automation Problems, National Academy of Sciences of the Republic of Armenia (IIAP), Yerevan, Armenia
minas.aslanyan@gmail.com, lasl@sci.am

Abstract. Human gait is a behavioral biometric that supports recognition from a distance when face or appearance cues are unavailable. Skeleton-based gait recognition is attractive because 3D pose sequences capture body-joint motion while suppressing background, clothing texture, and illumination variation. However, skeleton quality depends on visible body evidence, and realistic occlusion can produce missing joints, noisy coordinates, inconsistent limb tracks, and unreliable temporal segments. Existing models often degrade under such partial-body evidence. This paper proposes OAP-SupCon, an occlusion-aware part-level supervised contrastive learning framework for robust skeleton-based gait recognition. The method trains an encoder with paired views that simulate joint masking, body-part masking, temporal cropping, coordinate perturbation, view rotation, and speed variation. Unlike generic contrastive learning, the objective models anatomical structure by contrasting complete and partially occluded sequences at both global and body-part levels. It combines supervised contrastive learning, visibility-weighted part representations, temporal consistency, and cross-entropy classification. We also define a reproducible evaluation protocol for clean, synthetically occluded, and real-occlusion settings, including rank-1 accuracy, mean average precision, top-1 transfer accuracy, and robustness drop across increasing occlusion severity. The contribution is methodological: the framework, four-loss objective, and fixed protocol are presented for pre-registered and auditable evaluation, while empirical validation is deferred to a follow-up study.

Keywords: skeleton-based gait recognition · occlusion robustness · contrastive learning · supervised contrastive learning · 3D pose sequences · body-part masking

1 Introduction

Gait is a behavioral biometric that can identify a person from walking patterns at a distance. Compared with appearance-based representations such as RGB frames or silhouettes, skeleton sequences offer a compact and semantically meaningful description of human motion. They encode the trajectories of body joints,

are less dependent on clothing texture or background, and can be processed naturally by graph convolutional networks or transformers. These properties make skeleton-based gait recognition appealing for long-range recognition, robotics, healthcare monitoring, and surveillance. For gait recognition in particular, the periodic movement of the pelvis, knees, ankles, shoulders, and arms can reveal identity-specific walking dynamics that are not obvious from a single image.

The main weakness is that skeletons are only as reliable as the body evidence visible to the pose estimator. In realistic scenes, body parts are frequently hidden by other pedestrians, bags, furniture, clothing, viewpoint changes, camera truncation, and self-occlusion. When this happens, a skeleton sequence can contain missing joints, misplaced limbs, temporally inconsistent tracks, or low-confidence coordinates. Recent gait-recognition surveys identify occlusion as a central open challenge and organize existing methods by occlusion type, datasets, and model design [8]. The ECCV 2024 OccGait/GaitMoE work further emphasizes that occlusion creates both missing and noisy body information, motivating dedicated occlusion-aware recognition strategies [6]. A gait system trained only on clean pose tracks may implicitly learn brittle shortcuts that break when a pedestrian carries a bag, walks behind a barrier, or appears in a crowded scene.

Contrastive learning is a natural tool for this problem because it can teach an encoder to map different views of the same sequence close together while separating different identities or action classes. General methods such as SimCLR, MoCo, and supervised contrastive learning show that carefully designed positive pairs, projection heads, and large negative sets can produce strong representations [1, 5, 7]. However, simply applying a generic contrastive objective to gait data is not sufficiently domain-specific. The augmentation space must reflect the physics and failure modes of human skeleton sequences. In particular, an occluded gait model should learn that the same identity can remain recognizable when the left leg is unreliable, when arms are missing, or when only a short temporal crop is visible. The central design question is therefore not whether contrastive learning can be used, but which positive pairs are scientifically meaningful for occluded human motion. We argue that positive pairs should preserve the same underlying identity while varying the visible body evidence in anatomically plausible ways.

This paper proposes OAP-SupCon, an occlusion-aware part-level supervised contrastive learning framework for skeleton-based gait recognition. Figure 1 summarizes the overall pipeline. The method generates complete, temporally distorted, coordinate-perturbed, and body-part-masked views from the same 3D skeleton sequence. A skeleton encoder maps each view into global and part-level embeddings. The global loss enforces identity consistency between complete and incomplete skeleton sequences, while the part-level loss uses human body partitions to learn stable representations from torso, arms, and legs. Temporal positives are sampled from different crops of the same walking sequence to encourage gait-cycle consistency. The final objective combines supervised contrastive learning with cross-entropy classification. This formulation encourages the encoder to preserve identity-relevant motion patterns even when the available pose evidence

is incomplete. At the same time, the classification term keeps the embedding useful for the downstream recognition objective rather than only optimizing view invariance.

The main research question is:

> Can occlusion-aware contrastive learning improve the robustness of skeleton-based human recognition when body parts are partially occluded?

The paper makes the following contributions:

- **Complete-to-partial positive construction.** We formulate a supervised contrastive objective in which one view of each skeleton sequence is clean and the other is anatomically occluded, so identity alignment between intact and partial body evidence is the explicit learning signal rather than a side effect of generic augmentation.
- **Part-level SupCon with visibility-weighted pooling.** We introduce per-region contrastive heads on a five-part anatomical partition (torso, left/right arm, left/right leg), where each part embedding is pooled only over joints that remain visible according to a binary mask M.
- **Reliability-gated part contrast.** We weight each part's loss contribution by its per-view visibility reliability $\omega_{i,p}^{v}$, so fully masked parts neither dominate gradients nor force the encoder to hallucinate features for missing anatomy.
- **A reproducible robustness protocol.** We specify datasets, three corruption families (random-joint, body-part, temporal) at five severity levels, a robustness-drop metric $\Delta(\rho)$, and a pre-registered ablation grid that isolates the contribution of each loss term.

Scope of this paper. This paper contributes the OAP-SupCon framework and its evaluation protocol; empirical results are deferred to a follow-up study, and every predicted effect in the remainder of the text is stated as a hypothesis to be tested rather than a measured outcome.

2 Related Work

2.1 Contrastive Representation Learning

Contrastive learning trains representations by pulling positive pairs together and pushing negative pairs apart. SimCLR demonstrated that augmentation design, nonlinear projection heads, larger batches, and longer training are important for self-supervised visual representations [1]. MoCo reformulated contrastive learning as dictionary lookup, using a queue and momentum encoder to maintain a large and consistent set of negatives [5]. Supervised contrastive learning extends the contrastive objective to labeled data by treating all samples from the same class as positives and samples from other classes as negatives [7]. These methods motivate our objective, but their standard image augmentations do not directly

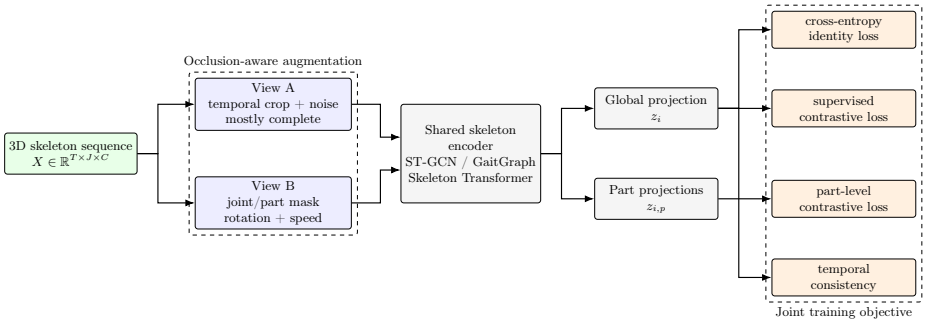

Fig. 1. Overview of OAP-SupCon. Two occlusion-aware views are sampled from the same 3D skeleton sequence. A shared skeleton encoder produces global and part-level features. Supervised contrastive, part-level contrastive, temporal consistency, and classification losses jointly train an occlusion-robust gait representation.

capture skeleton-specific occlusion, joint reliability, or gait-cycle structure. For skeleton gait, the relevant invariances are different from image augmentations: a valid representation should tolerate a missing arm, a noisy ankle track, a shorter walking segment, or a moderate viewpoint change, while still preserving fine-grained identity cues, motivating a domain-specific design rather than reuse of image augmentation recipes.

2.2 Skeleton-Based Contrastive Learning

Skeleton-specific contrastive learning has shown that pose sequences benefit from dedicated spatial and temporal transformations. Skeleton-Contrastive 3D Action Representation Learning learns invariance across multiple skeleton representations and skeleton augmentations for self-supervised action recognition [14]. For person re-identification, skeleton prototype contrastive learning models multi-level skeleton graphs and learns from unlabeled skeletons through prototype-based contrast [11]. Masked-autoencoder pretraining has also been adapted to skeleton sequences: SkeletonMAE reconstructs spatial-temporal joint patches to provide a self-supervised initialization that is complementary to contrastive objectives [15]. These works demonstrate the value of contrastive objectives for skeleton data. Our work differs by targeting occluded gait recognition and by making missing body regions part of the positive-pair construction and loss design. In other words, occlusion is not treated as incidental noise added after training; it is part of the learning signal. The model is repeatedly asked to answer whether two sequences belong to the same identity when one of them contains incomplete anatomical evidence.

2.3 Occlusion in Gait and Human Pose

Occlusion remains a major challenge for gait recognition. Li et al. provide a 2024 survey of gait recognition against occlusion, including occlusion taxonomies,

datasets, and method categories [8]. Huang et al. propose OccGait and GaitMoE, arguing that extensive occlusion introduces missing and noisy information as well as body misalignment [6]. In skeleton-based action recognition, IosPSTL explores self-supervised learning in occluded environments and combines occluded partial spatio-temporal learning with imputation [2]. Contrastive learning has also been applied to occluded humans in related domains, including instance-aware contrastive learning for occluded human mesh reconstruction [4] and exemplar-guided contrastive learning for pedestrian detection [9]. These studies support the broader idea that contrastive objectives can improve robustness under ambiguous or missing human evidence. They also show that occlusion is not a single phenomenon. It may appear as localized limb disappearance, person-to-person overlap, inaccurate joint assignment, temporal interruption, or systematic camera truncation. The proposed framework therefore separates random-joint, body-part, and temporal occlusion during training and evaluation.

2.4 Positioning Against the Closest Neighbors

The two methods that overlap most directly with OAP-SupCon are IosPSTL [2] and OccGait/GaitMoE [6]. IosPSTL targets occluded skeleton-based *action* recognition with a self-supervised partial spatio-temporal objective and an imputation branch that reconstructs missing joints; its supervision signal comes from masked-region reconstruction rather than from cross-view identity alignment. OccGait/GaitMoE targets occluded *gait* but on silhouette and image-level features, using an action-detection viewpoint and a mixture of experts to specialize across occlusion patterns; it does not learn part-level contrastive structure in skeleton space. OAP-SupCon differs from both on three concrete points: (i) the learning signal is *complete-to-partial positive construction* on the same identity, so the encoder is explicitly trained to match a clean and an occluded view rather than to impute or to route through experts; (ii) the objective adds a *part-level supervised contrastive loss with visibility-weighted pooling* that operates on anatomical regions defined over skeleton joints; and (iii) part contributions are *reliability-gated*, so fully masked parts neither dominate the loss nor are forced to align with visible counterparts. Neither IosPSTL nor OccGait/GaitMoE combines these three mechanisms, and neither casts occluded gait as a part-level contrastive problem on 3D skeletons.

2.5 Skeleton-Based Gait Recognition

Graph-based skeleton encoders are a common foundation for pose and gait analysis. ST-GCN models skeleton sequences as spatial-temporal graphs and learns action representations from joint connectivity and temporal edges [16]. GaitGraph adapts graph convolutional networks to skeleton-based gait recognition and argues that pose-based gait can provide a cleaner model-based alternative to silhouettes [13]. OpenGait provides a practical gait-recognition benchmark and codebase for evaluating methods across multiple public gait datasets [3]. Our

framework can use ST-GCN, GaitGraph-style graph encoders, or skeleton transformers as the backbone, with the proposed occlusion-aware contrastive objective added during training. This backbone-agnostic design is important because the contribution lies primarily in the occlusion-aware representation-learning objective. A stronger encoder can be substituted without changing the definitions of positive pairs, body parts, temporal crops, or robustness metrics.

3 Method

3.1 Problem Definition

Let a 3D skeleton sequence be

$$X \in \mathbb{R}^{T \times J \times C}, \tag{1}$$

where T is the number of frames, J is the number of joints, and $C = 3$ denotes 3D coordinates. We optionally concatenate a visibility or confidence channel, producing $\tilde{X} \in \mathbb{R}^{T \times J \times (C+1)}$. For gait recognition, each sequence has an identity label y_i. For action-recognition transfer experiments, y_i can instead denote the action class. The goal is to learn an encoder f_θ that maps a sequence to an embedding $h_i = f_\theta(X_i)$ that remains discriminative when subsets of joints or body regions are missing.

We define a body partition

$$\mathcal{P} = \{\text{torso}, \text{left arm}, \text{right arm}, \text{left leg}, \text{right leg}\}, \tag{2}$$

where each part $p \in \mathcal{P}$ is a subset of skeleton joints. This partition is used both to generate realistic occlusion and to compute part-level embeddings. The partition can be adapted to the skeleton format. For a 25-joint Kinect skeleton, the torso may include spine, neck, shoulders, and hips; leg parts may include hip, knee, ankle, and foot joints; arm parts may include shoulder, elbow, wrist, and hand joints. For 17-joint COCO-style pose, the same idea can be implemented with the corresponding shoulder, elbow, wrist, hip, knee, and ankle joints.

3.2 Design Rationale

The proposed framework follows three design principles. First, an occlusion-robust gait representation should be *view-consistent*: clean and corrupted views of the same walking sequence should remain close in embedding space. Second, it should be *part-aware*: the model should not collapse all body evidence into one global feature that hides which anatomical regions are reliable. Third, it should be *temporally stable*: different visible fragments of the same gait cycle should preserve identity-relevant dynamics.

These principles address common failure modes of skeleton recognition. If only global contrast is used, the model may learn to ignore local limb motion because global identity supervision is easier to satisfy. If only part-level contrast

is used, the model may lose the holistic coordination pattern between upper and lower body. If temporal consistency is omitted, the representation may become sensitive to the exact crop boundary and may fail when surveillance footage contains only a partial gait cycle. OAP-SupCon combines all three constraints so that global structure, local body-part evidence, and temporal motion consistency reinforce each other.

3.3 Core Method Components

The proposed method is implemented as four concrete modules rather than as a generic contrastive-learning wrapper. The first module is an *anatomical occlusion sampler*. For each training sequence, it samples a corruption type from joint masking, body-part masking, temporal interruption, coordinate noise, view rotation, and speed perturbation. For body-part masking, a part $p \in \mathcal{P}$ is sampled and all joints $j \in p$ are removed for either the full clip or a contiguous temporal interval. This simulates common gait failures such as a hidden leg, a missing arm swing, or a partially visible lower body.

The second module is *complete-to-partial positive construction*. View A is usually weakly augmented so that it remains close to the original walking sequence, while View B receives stronger occlusion. Both views keep the same identity label, and the contrastive objective explicitly treats them as positives. This teaches the representation that identity-relevant gait information should be stable when only partial body evidence is available.

The third module is *visibility-aware part pooling*. The encoder produces a feature tensor over time and joints. We denote by $M^v \in \{0,1\}^{T\times J}$ the binary visibility mask of view v, where $M^v_{i,t,j} = 1$ if joint j of frame t of sample i is visible after augmentation and 0 otherwise; we use the same symbol H^v for the per-joint encoder features. For each anatomical region, the part representation is pooled only from the joints that remain visible:

$$h^v_{i,p} = \frac{\sum_{t=1}^{T}\sum_{j\in p} M^v_{i,t,j} H^v_{i,t,j}}{\epsilon + \sum_{t=1}^{T}\sum_{j\in p} M^v_{i,t,j}}, \tag{3}$$

where ϵ prevents division by zero. Thus, a visible torso, left leg, or right arm can contribute a meaningful local descriptor without forcing the model to hallucinate features for fully missing parts.

The fourth module is *reliability-gated part contrast*. Part embeddings are contrasted only in proportion to their visibility reliability. If a body part is fully masked, its loss contribution is reduced; if it is visible, the part-level objective encourages that local motion pattern to remain close to same-identity part embeddings. Together, these modules define the paper's core method: anatomical occlusion simulation, complete-to-partial positives, visibility-aware part pooling, and reliability-gated contrastive learning.

3.4 Occlusion-Aware View Generation

For each training sequence X_i, two stochastic views are sampled:

$$X_i^a = \mathcal{A}_a(X_i), \qquad X_i^b = \mathcal{A}_b(X_i). \tag{4}$$

Unlike generic image augmentations, $\mathcal{A}$ is designed around skeleton occlusion and pose-estimation failures:

- **Joint masking:** randomly remove individual joints according to a masking probability ρ_j.
- **Part masking:** remove all joints belonging to a sampled body part, such as an arm, leg, torso, or lower body.
- **Temporal cropping:** sample different time windows from the same walking sequence.
- **Temporal speed perturbation:** drop or duplicate frames to simulate gait-rate variation and irregular pose tracks.
- **Coordinate perturbation:** add small Gaussian noise to visible coordinates to mimic pose-estimation uncertainty.
- **View rotation:** rotate the skeleton around the vertical axis to simulate camera-view variation.

The masking operation is recorded in a binary visibility tensor $M \in \{0,1\}^{T\times J}$. When a joint is masked, its coordinates are set to zero after normalization and its visibility value is set to zero. This prevents the model from confusing missing joints with valid coordinates. Figure 2 illustrates the generated views. In practice, the augmentation severity can be scheduled during training. Early epochs may use light masking so that the encoder first learns basic skeleton dynamics, while later epochs can introduce stronger part-level occlusion. This curriculum is useful when training from scratch on small gait datasets because very severe occlusion at the beginning can make positive pairs too difficult.

3.5 Skeleton Encoder and Projection Heads

The encoder f_θ can be instantiated with ST-GCN, GaitGraph, or a transformer-based skeleton model. Given an augmented sequence X_i^v, where $v \in \{a, b\}$, the encoder returns a spatio-temporal feature tensor

$$H_i^v = f_\theta(X_i^v). \tag{5}$$

A global pooling head produces a sequence-level representation

$$h_i^v = \mathrm{Pool}(H_i^v), \tag{6}$$

and a projection head g_ϕ maps it to the contrastive space:

$$z_i^v = \frac{g_\phi(h_i^v)}{\|g_\phi(h_i^v)\|_2}. \tag{7}$$

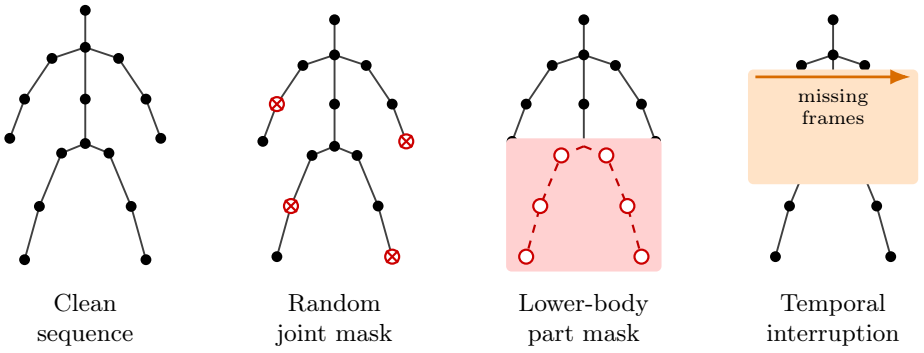

Fig. 2. Occlusion-aware view generation. Starting from the same walking pose, the augmentation sampler produces a clean view, a randomly joint-masked view, a body-part-masked view (here the lower body), and a temporally interrupted view. Masked joints have zero coordinates and a zero entry in the visibility mask M, so the encoder distinguishes missing evidence from valid zero-valued coordinates.

For each body part p, a part pooling operator extracts the corresponding local representation:

$$h_{i,p}^{v} = \mathrm{Pool}_p(H_i^v), \qquad z_{i,p}^{v} = \frac{g_{\phi,p}(h_{i,p}^{v})}{\|g_{\phi,p}(h_{i,p}^{v})\|_2}. \tag{8}$$

Part-level projection heads may share parameters across parts or be independent. We use independent heads when the target dataset is sufficiently large and shared heads when labeled data are limited. The classifier is attached to h_i rather than z_i: the projection space is optimized for contrastive geometry, while the encoder representation h_i is retained for retrieval and classification.

3.6 Global Supervised Contrastive Loss

For a mini-batch of N skeleton sequences, each sequence contributes two augmented views. Let $\mathcal{V}$ be the set of $2N$ views. For an anchor view u, the positive set is

$$P(u) = \{v \in \mathcal{V} \setminus \{u\} : y_v = y_u\}. \tag{9}$$

The supervised contrastive loss is

$$\mathcal{L}_{\mathrm{supcon}} = \sum_{u \in \mathcal{V}} \frac{-1}{|P(u)|} \sum_{v \in P(u)} \log \frac{\exp(\mathrm{sim}(z_u, z_v)/\tau)}{\sum_{k \in \mathcal{V} \setminus \{u\}} \exp(\mathrm{sim}(z_u, z_k)/\tau)}, \tag{10}$$

where $\mathrm{sim}(\cdot,\cdot)$ is cosine similarity and τ is a temperature parameter. This term pulls together views from the same identity even when one view is clean and the other is occluded. When identity labels are unavailable, the same equation can be reduced to self-supervised InfoNCE by treating only the two views of the

same sequence as positives. The supervised version is preferred for gait recognition because multiple walking sequences of the same person may show different clothing, carrying conditions, viewpoints, or gait cycles. Using all same-identity samples as positives encourages the embedding to capture persistent identity cues rather than only sequence-level similarity.

3.7 Part-Level Contrastive Loss

The part-level loss applies the same supervised contrastive principle to body-part embeddings. Because occluded parts may contain little useful information, each part is weighted by its visibility reliability:

$$\omega_{i,p}^{v} = \frac{1}{T|p|} \sum_{t=1}^{T} \sum_{j \in p} M_{i,t,j}^{v}. \tag{11}$$

The reliability-weighted part loss is

$$s_{u,v,p} = \exp(\mathrm{sim}(z_{u,p}, z_{v,p})/\tau_p). \tag{12}$$

$$\mathcal{L}_{\mathrm{part}} = -\sum_{p \in \mathcal{P}} \sum_{u \in \mathcal{V}} \frac{\omega_{u,p}}{|P(u)|} \sum_{v \in P(u)} \omega_{v,p} \log \frac{s_{u,v,p}}{\sum_{k \in \mathcal{V} \setminus \{u\}} s_{u,k,p}}. \tag{13}$$

This loss encourages a stable torso representation to match other torso evidence from the same identity, a stable leg representation to match other leg evidence, and so on. The global loss handles complete-to-incomplete matching, while the part loss prevents the representation from ignoring local gait cues. Reliability weighting is important because a fully masked part should not be forced to match a visible part with the same strength as two informative part observations. The weights therefore reduce the contribution of unreliable local features while still allowing the global loss to learn from the remaining visible body regions.

3.8 Temporal Consistency Loss

Gait recognition often depends on cyclic motion, and a real camera may capture only part of a walking cycle. We therefore create temporal positives from different crops of the same sequence. Let $z_i^{c_1}$ and $z_i^{c_2}$ be embeddings from two temporal crops. The temporal consistency term is

$$\mathcal{L}_{\mathrm{temp}} = -\log \frac{\exp(\mathrm{sim}(z_i^{c_1}, z_i^{c_2})/\tau_t)}{\sum_{k=1}^{N} \exp(\mathrm{sim}(z_i^{c_1}, z_k^{c_2})/\tau_t)}. \tag{14}$$

This term teaches the encoder that different visible segments of a walking sequence should preserve identity-relevant motion. The temporal term is especially relevant for surveillance data, where a person may be visible for only a few frames before being occluded or leaving the field of view. It also discourages overfitting to a fixed clip length, which improves compatibility with datasets that provide gait sequences of different durations.

3.9 Classification and Total Objective

The contrastive encoder is trained jointly with a classifier q_ψ. The cross-entropy term is

$$\mathcal{L}_{\text{ce}} = -\sum_i \log q_\psi(y_i \mid h_i). \tag{15}$$

The full training objective is

$$\mathcal{L} = \mathcal{L}_{\text{supcon}} + \alpha\mathcal{L}_{\text{part}} + \beta\mathcal{L}_{\text{temp}} + \lambda\mathcal{L}_{\text{ce}}, \tag{16}$$

where α, β, and λ control the contribution of part-level consistency, temporal consistency, and classification. Equation 16 combines the global identity-alignment term in Eq. 10, the visibility-weighted part term in Eq. 13, and the temporal-crop term in Eq. 14 into a single objective whose weights are reported alongside any empirical evaluation. At inference time, the projection heads are discarded and retrieval or classification uses the encoder representation h_i.

3.10 Training Procedure

For each mini-batch, the data loader samples identity-balanced skeleton sequences when possible. Each sequence is normalized, two augmented views are generated, and both views are passed through the shared encoder. The training step computes global supervised contrastive loss over all views in the batch, part-level contrastive loss over each anatomical partition, temporal consistency loss over crop pairs, and cross-entropy loss over identity labels. The total loss is back-propagated through the encoder and projection heads. During validation and testing, no synthetic masking is applied unless the purpose of the split is robustness evaluation. This separation ensures that clean accuracy and occlusion robustness can be reported independently.

3.11 Training Overhead

OAP-SupCon adds a moderate, predictable cost on top of a cross-entropy baseline that uses the same backbone. Two augmented views per sample roughly double the forward and backward pass cost; the encoder is shared, so parameter count grows only by the size of the global and per-part projection heads, which are small two-layer MLPs and therefore negligible relative to a graph or transformer encoder. The four-term loss adds one batch-wise similarity matrix for the global SupCon term, one similarity matrix per anatomical part for the part-level term, and one additional crop-pair comparison for the temporal term; each of these is $O(N^2 d)$ in the batch size N and embedding dimension d, which is the same order of magnitude as a single SupCon step. Overall, training time is expected to be approximately $2\times$ that of a cross-entropy baseline at matched batch size, with memory dominated by the doubled activations rather than by the loss machinery. Inference cost is unchanged: the projection heads and contrastive losses are discarded, and only $h_i = \text{Pool}(f_\theta(X_i))$ is used for retrieval or classification, so deployment-time complexity matches the chosen backbone exactly.

Table 1. Suggested datasets and evaluation roles.

Dataset	Task	Occlusion type	Primary metric
CASIA-B	gait ID	synthetic	rank-1
OU-MVLP	gait ID	synthetic	rank-1
Gait3D	gait ID in the wild	natural noise	rank-1, mAP
GREW	gait ID in the wild	natural noise	rank-1, mAP
OccGait / OccCASIA-B	gait ID	real annotated occlusion	rank-1, mAP
NTU RGB+D 60/120	action transfer	synthetic skeleton occlusion	top-1

4 Experimental Design

This section defines the experimental protocol. Table 1 fixes the datasets and metric roles used by the protocol, while Table 2 records the hypothesized qualitative ordering of methods that the protocol is designed to test; no empirical claims are made before those experiments are run. The goal of the experiments is not only to obtain high clean-data accuracy, but to measure whether the learned representation degrades gracefully as body evidence is removed. For that reason, the protocol reports both absolute recognition performance and robustness drop.

4.1 Datasets

For gait recognition, the framework can be evaluated on public benchmarks commonly supported by OpenGait, including CASIA-B, OU-MVLP, Gait3D, and GREW [3]. For real occlusion, OccGait and OccCASIA-B from the Gait-MoE work provide an especially relevant benchmark [6]. For skeleton action-recognition transfer, NTU RGB+D and NTU RGB+D 120 provide 3D skeletons with 60 and 120 action classes, respectively [12, 10]. The official NTU dataset page reports 56,880 samples for NTU RGB+D and 114,480 samples for NTU RGB+D 120, with RGB, depth, skeleton, and infrared modalities.

4.2 Baselines

The following baselines isolate the effect of occlusion-aware contrastive learning:

- **Cross-entropy only:** the same encoder trained with classification loss only.
- **Generic SimCLR/SupCon:** contrastive training with temporal crop, noise, and rotation, but without body-part masking.
- **Joint-dropout training:** random joint masking with cross-entropy, but no contrastive loss.
- **Backbone baselines:** ST-GCN, GaitGraph, and skeleton transformers trained under their standard protocols.

– **OAP-SupCon:** the full proposed framework with global, part-level, temporal, and classification losses.

All baselines should use the same train/test split, skeleton normalization, input length, optimizer, and backbone capacity. This is essential because the expected improvement should come from the occlusion-aware objective rather than from a larger model or a more favorable training schedule.

4.3 Occlusion Protocol

Synthetic occlusion is applied at severity levels $\rho = 0$, 0.1, 0.3, 0.5, and 0.7. At each severity, we evaluate three corruption families:

– **Random-joint occlusion:** joints are independently masked across time.
– **Body-part occlusion:** anatomically meaningful regions are masked, such as left leg, right leg, lower body, upper body, or torso.
– **Temporal occlusion:** a contiguous frame interval is masked or removed to simulate track interruption.

The robustness drop at severity ρ is

$$\Delta(\rho) = A(0) - A(\rho), \tag{17}$$

where $A(0)$ is clean accuracy and $A(\rho)$ is accuracy under occlusion severity ρ. Lower $\Delta(\rho)$ indicates better robustness. For real-occlusion datasets, the same metric can be computed by grouping probe sequences according to annotated occlusion severity or occlusion type. When annotations are unavailable, pose confidence or the percentage of missing joints can be used as a proxy severity measure. Figure 3 summarizes the evaluation matrix spanned by occlusion severity and corruption family.

4.4 Ablation Studies

The ablation study should answer four questions:

1. Does supervised contrastive learning improve over cross-entropy training?
2. Does body-part masking improve robustness beyond generic skeleton augmentations?
3. Does the part-level loss improve severe occlusion performance?
4. Does temporal consistency improve recognition from short or interrupted walking clips?

Each ablation removes exactly one component from the full method while keeping all other settings unchanged. The most important comparison is between generic SupCon and OAP-SupCon, because this directly tests whether domain-specific occlusion design adds value beyond standard contrastive learning.

The hypothesized pattern in Table 2 encodes the scientific hypothesis that the evaluation protocol is designed to test. Generic supervised contrastive learning is

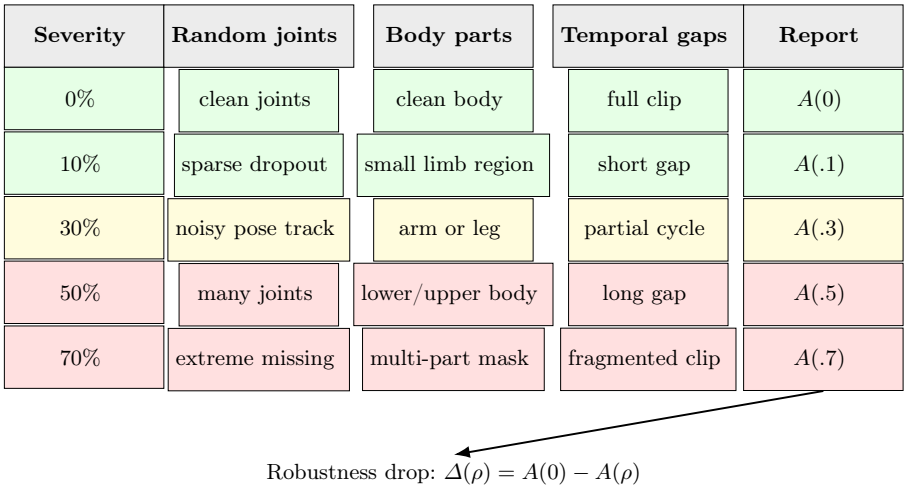

Fig. 3. Evaluation matrix for robustness analysis. Models are trained on clean and occlusion-augmented sequences, then tested across increasing joint, part, and temporal occlusion severity.

expected to improve over cross-entropy training because same-identity samples form a more compact embedding space. Joint masking with cross-entropy should improve robustness under missing joints, but it does not explicitly align complete and occluded representations. Removing the part-level loss is expected to reduce severe-occlusion performance because the model loses explicit anatomical consistency. Removing the temporal loss is expected to mainly hurt temporally interrupted probes. The full method is therefore predicted to produce the smallest rank-1 drop at 50% occlusion. We emphasize that Table 2 records predictions, not measurements, and that any deviation observed during validation would itself be an informative outcome of the protocol.

4.5 Implementation Details

All skeletons should be root-centered at the pelvis or torso joint and normalized by an estimate of body scale. For 3D skeletons, view rotation is sampled around the vertical axis. For 2D pose-derived gait skeletons, rotation should be replaced by horizontal flip and scale perturbation unless camera calibration is available. Training uses two views per sequence, a temperature $\tau \in [0.05, 0.2]$, and an embedding dimension between 128 and 512. A practical starting point is $\alpha = 0.5$, $\beta = 0.2$, and $\lambda = 1.0$, followed by validation on a held-out split. The same encoder architecture and optimizer settings should be used across baselines to make the comparison fair. For retrieval experiments, the gallery should remain clean in one protocol and be occluded in a second protocol. The first protocol measures whether an occluded probe can still match a clean enrolled identity, while the second measures robustness when both enrollment and query skeletons

Table 2. Hypothesized relative ordering of methods under the proposed protocol. Each cell records the predicted change relative to the cross-entropy baseline, where "+" always indicates "better than baseline": higher rank-1 accuracy for the first three columns, and a smaller robustness drop (i.e. more robust) for the last column. "−" denotes no expected change, "+" a small expected gain, "++" a moderate gain, and "+ + +" a large gain. The table records the qualitative pattern the experiments are designed to test; no numerical values are reported because experiments have not yet been run.

Method	Clean	30%	50%	Drop at 50%
CE only (reference)	−	−	−	−
Generic SupCon	+	+	+	+
Joint masking + CE	−	+	+	+
OAP-SupCon without part loss	+	++	+	+
OAP-SupCon without temporal loss	+	++	++	++
OAP-SupCon full	+	++	+ + +	+ + +

are imperfect. Reporting both cases is useful because operational gait systems may face either condition.

4.6 Reproducibility Statement

To make the protocol auditable, we commit to releasing reference code, configuration files, occlusion samplers, evaluation scripts, and all random seeds upon publication. Any empirical study following this protocol should report, at minimum: the contrastive temperatures τ and τ_p (and τ_t if separate); the loss weights α, β, and λ; the embedding dimension and projection-head architecture; the per-part visibility-reliability weights $\omega_{i,p}^{v}$; the batch size, optimizer, learning-rate schedule, weight decay, and number of training epochs; the augmentation severities ρ_j for random-joint masking and the per-part masking probabilities; the temporal-crop length, stride, and speed-perturbation range; the rotation range used for view augmentation; and the backbone architecture and capacity. The same encoder configuration, normalization, and input length must be used across all baselines so that any observed difference is attributable to the occlusion-aware objective rather than to capacity or schedule. The split definitions, occlusion seeds, and evaluation severities should be fixed in advance and released alongside the code so that the protocol can be reproduced bit-for-bit by independent groups.

5 Expected Scientific Claims and Validation Criteria

The central claim to validate is that occlusion-aware positive-pair design improves robustness more than generic contrastive learning. A positive result should satisfy three criteria. First, OAP-SupCon should match or improve clean-data rank-1 accuracy relative to the same backbone trained with cross-entropy.

Second, the robustness drop $\Delta(\rho)$ should be lower under part occlusion, especially at medium and severe levels. Third, ablations should show that body-part masking and part-level contrastive learning each contribute measurable gains.

The representation can also be analyzed visually. UMAP or t-SNE plots should compare embeddings from clean and occluded sequences. If the method works as intended, clean and occluded samples from the same identity should remain close, while different identities remain separated. Part-level attention or reliability scores can further show whether the model relies on visible gait evidence instead of hallucinating missing joints. A strong result would show compact identity clusters across clean and occluded variants, smaller robustness drops than the supervised baseline, and ablation evidence that part masking and part-level loss are both necessary. If the full method improves severe occlusion but slightly reduces clean accuracy, the paper should report that trade-off explicitly and discuss whether robustness is the intended deployment priority.

6 Limitations and Ethics

Method limitations. OAP-SupCon assumes that an upstream pose extractor produces usable skeletons and that at least some body evidence remains visible; when the estimator fails completely, no contrastive objective can recover identity information that is absent from the input. The framework also depends on anatomically meaningful body partitions, so unusual camera views, assistive devices, or non-standard body morphologies may require the partition $\mathcal{P}$ to be redefined. A further limitation is that synthetic occlusion is not equivalent to real pose-estimator errors: real failures are typically structured rather than random, including left/right limb swaps, cross-identity limb assignment in crowded scenes, and temporally smooth but anatomically incorrect joint trajectories. The proposed protocol therefore separates synthetic-occlusion evaluation from real-occlusion evaluation, and we expect future work to combine the contrastive objective with explicit pose imputation or uncertainty-aware graph modeling to address structured-error cases.

Ethics and deployment considerations. Gait recognition is a biometric technology and may be deployed in sensitive surveillance contexts, so any use of this framework must respect local law, institutional review, informed consent, and strict data-governance practices. Evaluation should include demographic-bias audits and disability-related failure modes, in particular performance under mobility aids, varied body shapes, clothing, viewpoints, and recording environments, since occluded gait is the regime in which biased shortcuts are most likely to surface. Claims of recognition reliability should be confined to the populations and conditions actually evaluated, and robustness improvements measured by this protocol should not be treated as a substitute for consent, accountability, or human oversight in deployed systems.

7 Conclusion

This paper presents OAP-SupCon, an occlusion-aware part-level supervised contrastive framework for skeleton-based gait recognition. The contribution is not merely to use contrastive learning, but to redesign contrastive learning around the failure modes of occluded human motion. By generating complete and incomplete skeleton views, contrasting both global and body-part embeddings, and enforcing temporal consistency across gait crops, the framework aims to learn motion representations that preserve identity-relevant structure under missing joints and body regions. The proposed evaluation protocol measures clean accuracy, occlusion robustness, and ablation effects across synthetic and real occlusion settings. This positions the work as a domain-specific contrastive learning contribution for occluded human motion rather than a generic application of contrastive learning to gait data.

References

1. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 1597–1607 (2020), https://arxiv.org/abs/2002.05709
2. Chen, Y., Peng, K., Roitberg, A., Schneider, D., Zhang, J., Zheng, J., Chen, Y., Liu, R., Yang, K., Stiefelhagen, R.: Exploring self-supervised skeleton-based action recognition in occluded environments. arXiv preprint arXiv:2309.12029 (2023), https://arxiv.org/abs/2309.12029
3. Fan, C., Liang, J., Shen, C., Hou, S., Huang, Y., Yu, S.: OpenGait: Revisiting gait recognition toward better practicality. arXiv preprint arXiv:2211.06597 (2022), https://arxiv.org/abs/2211.06597
4. Gwon, M.G., Um, G.M., Cheong, W.S., Kim, W.: Instance-aware contrastive learning for occluded human mesh reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10553–10562 (2024). https://doi.org/10.1109/CVPR52733.2024.01004
5. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9729–9738 (2020), https://arxiv.org/abs/1911.05722
6. Huang, P., Peng, Y., Hou, S., Cao, C., Liu, X., He, Z., Huang, Y.: Occluded gait recognition with mixture of experts: An action detection perspective. In: Proceedings of the European Conference on Computer Vision (2024), https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/1016_ECCV_2024_paper.php
7. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. In: Advances in Neural Information Processing Systems. vol. 33, pp. 18661–18673 (2020), https://arxiv.org/abs/2004.11362
8. Li, T., Ma, W., Zheng, Y., Fan, X., Yang, G., Wang, L., Li, Z.: A survey on gait recognition against occlusion: Taxonomy, dataset and methodology. PeerJ Computer Science **10**, e2602 (2024). https://doi.org/10.7717/peerj-cs.2602

9. Lin, Z., Pei, W., Chen, F., Zhang, D., Lu, G.: Pedestrian detection by exemplar-guided contrastive learning. IEEE Transactions on Image Processing **32**, 2003–2016 (2023). https://doi.org/10.1109/TIP.2022.3189803
10. Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.Y., Kot, A.C.: NTU RGB+D 120: A large-scale benchmark for 3d human activity understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence **42**(10), 2684–2701 (2020). https://doi.org/10.1109/TPAMI.2019.2916873
11. Rao, H., Miao, C.: Skeleton prototype contrastive learning with multi-level graph relation modeling for unsupervised person re-identification. In: Quality, Reliability, Security and Robustness in Heterogeneous Systems. pp. 306–321. Springer (2024). https://doi.org/10.1007/978-3-031-65126-7_19, https://arxiv.org/abs/2208.11814
12. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: NTU RGB+D: A large scale dataset for 3d human activity analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1010–1019 (2016). https://doi.org/10.1109/CVPR.2016.115
13. Teepe, T., Khan, A., Gilg, J., Herzog, F., Hörmann, S., Rigoll, G.: GaitGraph: Graph convolutional network for skeleton-based gait recognition. In: 2021 IEEE International Conference on Image Processing. pp. 2314–2318 (2021). https://doi.org/10.1109/ICIP42928.2021.9506717
14. Thoker, F.M., Doughty, H., Snoek, C.G.M.: Skeleton-contrastive 3d action representation learning. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 1655–1663 (2021). https://doi.org/10.1145/3474085.3475307, https://arxiv.org/abs/2108.03656
15. Wu, W., Hua, Y., Zheng, C., Wu, S., Chen, C., Lu, A.: SkeletonMAE: Spatial-temporal masked autoencoders for self-supervised skeleton action recognition. arXiv preprint arXiv:2209.02399 (2023), https://arxiv.org/abs/2209.02399
16. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018). https://doi.org/10.1609/aaai.v32i1.12328