Spatial Topology Repair and Semantic Geometry Alignment Framework for Occluded Person Re-Identification

Anonymous Author(s)

***Abstract*—Identifying pedestrians under heavy occlusion (Occluded Re-ID) remains highly challenging, primarily because obstacles inevitably corrupt human structural integrity and induce severe spatial-semantic mismatching. Current approaches either struggle to recover fragmented topological features or blindly trust fragile pose estimators, making them highly vulnerable to complex background interference. To overcome these bottlenecks, we present SSGA, a unified multi-modal enhancement framework that seamlessly couples topology restoration, cross-modal feature calibration, and semantic-driven decoding. Specifically, a Spatial Guided Graph Convolutional Network (SG-GCN) is first formulated to repair corrupted local structures by embedding physical spatial constraints into visual patch representations. Moreover, to tackle cross-modal mismatching, we propose the Spatio-Semantic Dual-Metric Greedy Alignment (SSDA) strategy. By anchoring visual embeddings to reliable skeletal cues under strict geometric boundaries, SSDA effectively eliminates semantic ambiguity such as symmetrical limb confusion. Furthermore, a Geometry-Aware Semantic Matching (GASM) module is designed to employ learnable semantic queries for dynamically extracting part-level features, which forces the network to highlight visible body regions and filter out occlusion noise. Comprehensive evaluations across five standard benchmarks validate the superiority of our SSGA framework, which establishes new state-of-the-art results and yields substantial improvements particularly on the severely occluded Occluded-Duke and Occluded-ReID datasets.**

***Index Terms*—Person Re-identification, Occlusion Handling, Graph Convolutional Networks, Cross-Modal Alignment, Semantic Decoding, Vision Transformer.**

I. Introduction

Person Re-identification (Re-ID) aims to retrieve images of a specific person across non-overlapping camera views. Despite achieving remarkable success on holistic datasets, real-world surveillance scenarios inevitably involve complex occlusions caused by obstacles such as vehicles or pedestrians. These occlusions severely compromise the integrity of human features, leading to drastic performance degradation in existing models. The primary challenges in occluded Re-ID [1] are twofold: (1) Structural Destruction and Noise: Occlusions disrupt the topological integrity of the human body while introducing significant background noise, making it difficult for standard encoders to extract contiguous features. (2) Semantic and Spatial Misalignment: In crowded or complex pose scenarios, the limbs of non-target pedestrians are frequently misidentified as parts of the query target, and symmetric body parts (e.g., left vs. right arm) are easily confused due to the lack of explicit spatial guidance.

Existing solutions generally fall into three categories: Local feature-based methods [2], [3] attempt to mitigate occlusion via hard partitioning but often lose contextual information in non-occluded regions due to rigid grid division. Pose-guided alignment methods [2], [3] represent another mainstream direction. As generically illustrated in Figure 1, these methods typically adopt a "Visual-Skeleton Dual-stream" paradigm, where an auxiliary pose branch is employed to guide the visual encoder in localizing visible regions. While this dual-stream architecture provides explicit structural priors, existing implementations are heavily reliant on the accuracy of upstream pose estimators. They tend to blindly trust the skeletal coordinates, leading to cascading failures when estimation errors occur (e.g., in crowded or low-light scenarios). To mitigate this dependency, we introduce a dual-metric alignment strategy (SSDA) that does not blindly trust the pose estimator. Instead, it enforces geometric consistency only within a dynamically gated local field, allowing the model to correct misalignment even when the pose estimation is noisy. Attention-based methods [4], [5] offer flexibility, yet without explicit structural guidance, they often struggle to distinguish valid human parts from occlusions, rendering them susceptible to overfitting background noise.

To address these limitations, this paper proposes a novel multi-modal synergistic enhancement framework. We construct a closed-loop system encompassing Topology Repair, Cross-Modal Alignment, and Semantic Decoding. Our main contributions are summarized as follows:

1) Spatial Guided Graph Convolutional Network (SG-GCN): Unlike standard ViTs that lack inductive biases for local proximity, we propose an SG-GCN module that integrates dynamic graph construction with Laplacian Positional Encoding. By fusing physical spatial priors with semantic affinities, this module effectively reconstructs compromised topological structures and adapts to diverse pose variations.
2) Spatio-Semantic Dual-Metric Greedy Alignment (SSDA): To resolve the spatial misalignment between visual patches and skeletal keypoints, we design a dual-metric alignment strategy. By introducing a Bounded Linear Decay Field as a geometric constraint alongside semantic similarity, we eliminate ambiguities in visual matching (e.g., symmetric limbs) and inject robust anatomical semantics into visual features.
3) Geometry-Aware Semantic Decoder with GASM: We introduce a decoding paradigm that decouples high-level semantics (e.g., head, torso) via learnable queries. The proposed Geometry-Aware Semantic Synergistic Matching (GASM) mechanism utilizes dynamic centroids to accurately map these semantic views back to visual features, ensuring precise focusing on valid human parts even under severe occlusion.

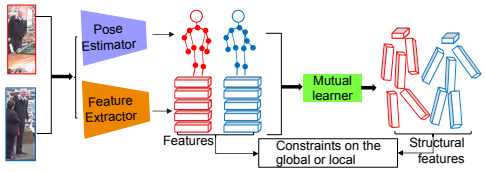

Fig. 1. A typical Visual-Skeleton Cross-modal Dual-stream Framework employed in pose-guided Re-ID. It generally consists of a visual encoder and a pose estimator, where skeletal keypoints are used to align or attend to visual features. Our proposed method improves upon this paradigm by introducing topological repair and uncertainty-aware alignment.

4) State-of-the-Art Performance: We achieve results comparable to or exceeding state-of-the-art (SOTA) performance on both occluded and holistic benchmarks. Furthermore, we provide a detailed adaptability strategy, demonstrating how our uncertainty modeling components effectively balance performance between clean and occluded data.

II. Related Work

A. Vision Transformer

Following the significant success of the Transformer architecture in Natural Language Processing (NLP), its self-attention mechanism has been adapted to the Computer Vision (CV) domain, leading to the emergence of the Vision Transformer (ViT) [6]. By partitioning images into sequences of patches and employing self-attention, ViT effectively transcends the limitations of Convolutional Neural Networks (CNNs) in modeling long-range dependencies. However, standard ViTs inherently lack the inductive bias for local spatial structures. Consequently, they often overlook crucial physical proximity information when processing image patches arranged in a regular grid layout. To address this deficiency, we introduce a Graph Convolutional Network (GCN) module to explicitly compensate for this limitation.

B. Graph Convolutional Network (GCN)

When dealing with non-Euclidean data such as molecules and point clouds, CNN and ViT architectures, which rely on regular grids, often exhibit limited adaptability. As a cornerstone of the Graph Neural Network (GNN) paradigm, the Graph Convolutional Network (GCN) [7] successfully extends the convolution operator to irregular graph domains. By updating node features via the aggregation of neighborhood information based on topological adjacency, this mechanism not only inherits the local perception characteristic of convolution but also transcends the limitations of grid structures, thereby enabling effective modeling and feature fusion for irregular data. The GNN landscape has witnessed continuous evolution in node aggregation mechanisms, ranging from the early GCN that simplified computation, to GAT [8] which introduced attention mechanisms, and GraphSAGE [9] which enabled large-scale sampling. More recently, advanced variants have emerged to address complex visual tasks. For instance, Sun et al. [10] proposed a Wavelet GCN to filter non-stationary noise in corrupted VI-ReID scenes. Nevertheless, conventional GCNs and even these recent variants typically assume a fixed and complete topological structure, resulting in a lack of flexibility when processing data with compromised topology, such as occluded pedestrians. In contrast, the SG-GCN module proposed in this paper employs a dynamic graph construction strategy. By adaptively generating graph structures based on feature similarity and spatial distance, it effectively achieves a synergy between local geometric constraints and global semantic correlations.

C. Pose Estimation

Pose estimation has witnessed an evolution from early direct regression methods to encoder-decoder based architectures, such as Stacked Hourglass [11] and SimpleBaselines [12], and more recently, discrete token-based compositional representations [13]. Despite these advancements, conventional heatmap-based approaches still dominate practical applications, yet they often result in a loss of fine-grained spatial details due to inherent downsampling operations. Addressing this bottleneck, HRNet [14] maintains high-resolution representations throughout the entire feature extraction process, thereby significantly enhancing the precision of keypoint localization. In the context of occluded person Re-ID, purely visual features are highly susceptible to interference from occlusion noise. Conversely, skeletal keypoints offer robust priors based on rigid body structures. Consequently, we employ HRNet to extract high-fidelity skeletal features to serve as semantic guidance. By injecting explicit anatomical semantics into visual representations via cross-modal alignment, our approach effectively suppresses background noise and focuses attention on valid human regions.

D. Occluded Person Re-identification

Existing methodologies for addressing occlusion challenges have primarily crystallized into three mainstream paradigms: strategies based on feature enhancement, structure-guided alignment, and attention-based mining. Local Feature and Enhancement Methods focus on mitigating occlusion effects through operations at the feature level. These methods typically employ strategies such as feature erasing, diffusion, or local partitioning to recover or reinforce human features within the feature space. For instance, Somers et al. [3] proposed a learning scheme for body part representations that is robust to occlusion and non-discriminative local appearances; Wang et al. [2] designed the Feature Erasing and Diffusion network (FED) to improve model robustness against feature absence by simulating occlusions. However, when dealing with severe occlusion, these methods often struggle to maintain complete contextual consistency in non-occluded regions due to the lack of explicit semantic guidance. Structure-Guided Alignment Methods utilize external priors, such as human parsing or keypoints, to explicitly localize visible human regions. Dou et al. [15] introduced a co-parsing mechanism to guide feature alignment, aiming to filter out background noise; Somers et al. [16] further explored keypoint-based prompt learning to enhance retrieval capabilities for specific body parts. Although incorporating structural priors effectively mitigates background interference, this category of methods is highly sensitive to the accuracy of upstream estimation models, where

deviations in pose estimation can easily lead to a cascading failure of feature alignment. Attention-based Methods leverage the data-driven characteristics of Transformers or attention modules to automatically mine highly discriminative regions. Li et al. [4] proposed an Occlusion-Aware Transformer (OAT), employing a second-order attention mechanism to capture complex feature correlations; Ren et al. [5] enhanced feature representation by leveraging occlusion attributes. While these methods possess high flexibility in feature extraction, in the absence of explicit topological constraints, models are prone to overfitting to specific background noise, and the interpretability of the decision-making process is relatively weak. The framework proposed in this paper aims to integrate the advantages of the aforementioned paradigms. We utilize SG-GCN to repair damaged local topology (aligning with local feature enhancement), employ skeletal keypoint features as rigid structural guidance (aligning with the robustness of structure alignment), and achieve adaptive feature focusing via semantic views (aligning with the flexibility of attention mechanisms). This synergistic approach realizes a unification of structural integrity and semantic accuracy in complex occlusion scenarios.

III. Methodology

To address the dual challenges of structural destruction and semantic ambiguity in occluded Re-ID, we propose a multi-modal synergistic enhancement system. The detailed architecture is depicted in Figure 2. This system is designed as a closed-loop framework: 1. Topology Repair: The SG-GCN (Sec. 3.1) re-establishes local structural continuity within visual patches to mitigate initial feature corruption. 2. Cross-Modal Alignment: The SSDA strategy (Sec. 3.2) anchors these repaired visual features to robust skeletal priors via spatio-semantic dual-metrics, resolving misalignment and limb confusion. 3. Semantic Decoding: The GASM-based decoder (Sec. 3.3) utilizes learnable queries to focus on valid human parts while adaptively suppressing background and occlusion noise.

The synergy between these components ensures that the model transitions from low-level topology repair to high-level semantic reasoning, forming a robust representation for complex environments.

A. Visual Encoder and Topology Enhancement

1) Basic Feature Formulation: Following the standard Vision Transformer (ViT) paradigm, the initial step in visual processing involves discretizing the image into a sequence of patches. Given an input image $I \in \mathbb{R}^{H\times W\times C}$, we employ a sliding window strategy to partition it into N fixed-size patches. The number of patches N depends on the stride S and patch size P, calculated as follows:

$$N = \left\lfloor \frac{H+S-P}{S} \right\rfloor \times \left\lfloor \frac{W+S-P}{S} \right\rfloor \tag{1}$$

When $S < P$, the generated patches overlap, which helps mitigate the loss of spatial neighborhood information caused by rigid partitioning. Subsequently, we apply a trainable linear projection $f(\cdot)$ to the flattened patches, mapping them to D-dimensional embeddings $E_i = f(x_i)$. To preserve essential spatial structure and viewpoint information during serialization, we incorporate learnable positional encodings P_E and camera view encodings C_{id} into the patch embeddings. Additionally, a learnable [class] token is prepended to the sequence to aggregate global features. The final input sequence E_{input} is defined as:

$$E_{input} = \{x_{class}; E_i\} + P_E + \lambda_{cm} C_{id} \tag{2}$$

where λ_{cm} is a hyperparameter balancing the weight of the camera encoding. After interaction through m layers of the Transformer encoder, we obtain the feature sequence f_{VB}.

2) Spatial Guided Graph Convolutional Network (SG-GCN): While the self-attention mechanism of standard Transformers excels at capturing long-range dependencies, it lacks an inductive bias for local spatial proximity. To explicitly enhance the local topological representation of features f_{VB}, we propose the Spatial Guided Graph Convolutional Network (SG-GCN) module. Specifically, considering that the global token f_{gb} (contained within f_{VB}) lacks explicit spatial coordinates, we decouple it from the sequence and construct dynamic graphs exclusively on the remaining N patch tokens. To embed absolute spatial relationships within this patch feature space, we construct a static adjacency matrix A_{tp} and a binary truncation matrix M_{sp} based on the physical grid coordinates. Let $\mathbf{p}_i$ and $\mathbf{p}_j$ denote the spatial coordinates of the i-th and j-th patches, respectively. We utilize a Gaussian kernel function to define the continuous spatial affinity A_{tp}:

$$A_{tp}(i,j) = \exp\left(-\frac{\|\mathbf{p}_i - \mathbf{p}_j\|_2^2}{2\sigma^2}\right) \tag{3}$$

Simultaneously, to explicitly delineate the scope of the local neighborhood, we define the binary matrix M_{sp} as a hard spatial constraint:

$$M_{sp}(i,j) = \begin{cases} 1, & \text{if } \|\mathbf{p}_i - \mathbf{p}_j\|_2 \le \tau \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

Here, $\|\cdot\|_2$ represents the Euclidean distance,and σ controls the width of the Gaussian distribution. τ is a predefined spatial truncation threshold, which is empirically set to preserve local neighbors within a radius of 2 grid units. While A_{tp} provides the model with smooth distance awareness, M_{sp} is employed during the dynamic graph construction phase to explicitly decouple local neighborhoods from long-range semantic associations.

Simultaneously, we introduce Laplacian Positional Encoding (LPE). By performing eigendecomposition on the normalized Laplacian matrix $\mathbf{L}$, we project the eigenvectors $\mathbf{U}_k$ corresponding to the k smallest non-zero eigenvalues into the feature dimension, thereby endowing each node with coordinate awareness within the global topology. Considering that static graphs struggle to adapt to pose variations, we further design a dynamic graph construction mechanism. This mechanism integrates semantic affinity with spatial priors, defining the dynamic adjacency matrix $\tilde{A}$ as:

$$\hat{A} = \alpha\cdot(A_{sem}\odot A_{tp}\odot M_{sp}) + (1-\alpha)\cdot(A_{sem}\odot A_{tp}\odot(1-M_{sp})) \tag{5}$$

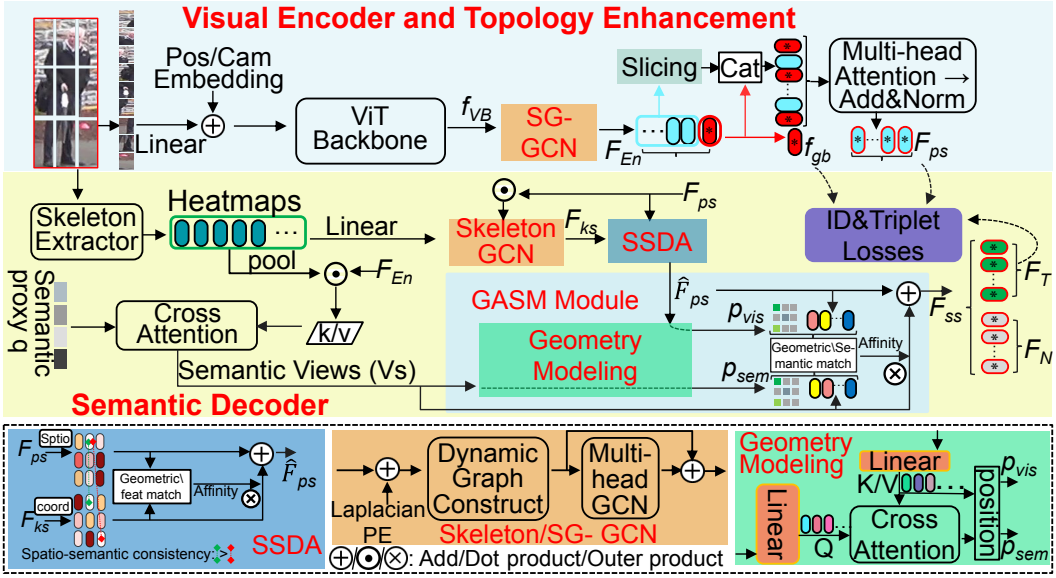

Fig. 2. Overview of the proposed Spatial Topology Repair and Semantic Geometry Alignment (SSGA) framework. It consists of three main stages: (a) Visual encoding with SG-GCN for topology repair; (b) Cross-modal feature alignment via SSDA to bridge the gap between visual patches and skeletal keypoints; and (c) Geometry-aware semantic decoding using GASM for precise part-level feature extraction under occlusion. Visually, the diagram is organized into three tiers: the top tier for Visual Encoder and Topology Enhancement, the middle tier for the Semantic Decoder, and the bottom tier detailing sub-module architectures and symbols.

where A_{sem} is derived from feature similarity, A_{tp} is derived from topology, M_{sp} is a spatial mask, and α is a balancing factor that decays as training progresses (0.9999→0.9(epoch=20)). This design enables the model to rely on local spatial continuity to cluster objects in the early stages, and gradually transcend spatial limitations to associate disjoint regions of the same class (separated by occlusion) using semantic similarity as training deepens. Finally, we employ a multi-head graph convolution mechanism to update node features in parallel across different subspaces:

$$H^{(l+1)} = \text{Concat}\left(\hat{A}H_1^{(l)}W_1^{(l)}, \ldots, \hat{A}H_h^{(l)}W_h^{(l)}\right)W_O \quad (6)$$

where W_i denotes the learnable weight matrix for each head. This multi-head design enables the model to simultaneously capture local geometric details and long-range semantic relationships across different subspaces. The number of GCN layers is set to 1, while the hidden dimension d is determined based on ablation studies. Finally, the output features are fused with the original input via a residual connection to yield the final spatially enhanced features F_{En}.

Visual Group Features (F_{ps}): We partition the spatially enhanced patch features F_{En} into $M = 17$ local groups. Note that the global token f_{gb} (derived from the ViT backbone) is preserved and concatenated to the end of each group sequence to provide global context. The resulting concatenated sequences are then fed into a shared Transformer layer to facilitate intra-group interaction. Finally, we select the last token of each output sequence as the representative feature for that group, denoted as $F_{ps} \in \mathbb{R}^{M\times C}$.

3) Encoder Loss: The training of the encoder employs a joint supervision strategy. For the M elements of F_{ps}, we denote as $F_i, i \in [1, M]$ and apply Cross-Entropy classification loss and an Patch-wise Triplet loss (metric loss), respectively.

$$L_{cls} = \frac{1}{M}\sum_{i=1}^{M} L_{id}(CLS_i(F_i)) \quad (7)$$

$$L_{metr} = \frac{1}{M}\sum_{i=1}^{M} L_{tri}(F_i) \quad (8)$$

B. Cross-Modal Feature Alignment

1) Feature Grouping and Preprocessing: To bridge the heterogeneity between visual and skeletal modalities and establish a canonical feature space, we initially perform Topology-Enhanced operations on the encoder outputs. Topology-Enhanced Skeleton Features (F_{ks}): we utilize HRNet to extract features for the $M = 17$ corresponding skeletal keypoints and project them to dimension C via a linear mapping. To incorporate visual semantics, we perform an element-wise product between these features and F_{ps} to obtain initial embeddings. Subsequently, a Graph Convolutional Network (GCN) based on the natural human topology is constructed to process these embeddings, capturing structural dependencies to yield the topology-enhanced skeleton features $F_{ks} \in \mathbb{R}^{M\times C}$. The implementation details of this GCN align closely with those in Section III-A2, with the primary distinction being that the initial adjacency matrix is derived from the human skeletal topology.

2) Spatio-Semantic Dual-Metric Greedy Alignment (SSDA): Although the visual features F_{ps} and skeletal features F_{ks} are aligned in the feature dimension, they are constructed based on regular grids and human topology, respectively, resulting in inherent spatial misalignment. This issue is particularly pronounced under complex poses, where reliance on appearance features alone often leads to semantic confusion between symmetric body parts. To address this, we propose the Spatio-Semantic Dual-Metric Greedy Alignment (SSDA) strategy.

We posit that robust cross-modal matching should exhibit consistency across both the feature manifold and Euclidean space. Therefore, we construct a comprehensive similarity matrix $\mathbf{S}_{total} \in \mathbb{R}^{M\times M}$, which is a weighted combination of Feature Similarity $\mathbf{S}_{feat}$ and Spatial Similarity $\mathbf{S}_{geo}$.

a) Skeleton Coordinate Extraction and Spatial Normalization: Before defining the spatial metric, we explicitly formulate the source of the skeletal coordinates. Let I be the input image. The HRNet backbone generates a set of M keypoint heatmaps $\mathcal{H} = \{\mathbf{H}_j\}_{j=1}^{M}$, where $\mathbf{H}_j \in \mathbb{R}^{H'\times W'}$. The raw coordinate $\hat{p}_{skl}^{(j)}$ of the j-th skeletal point is obtained by locating the peak response in the heatmap:

$$\hat{p}_{skl}^{(j)} = \arg\max_{(x,y)} \mathbf{H}_j(x,y) \tag{9}$$

To construct a unified metric space, we map both the center coordinates of visual patches $p_{vis}^{(i)}$ and the skeletal coordinates $\hat{p}_{skl}^{(j)}$ into a normalized unit space $\Omega \in [0,1]^2$. The normalized skeletal coordinate $p_{skl}^{(j)}$ is defined as:

$$p_{skl}^{(j)} = \left(\frac{\hat{p}_{skl,x}^{(j)}}{W'}, \frac{\hat{p}_{skl,y}^{(j)}}{H'} \right) \tag{10}$$

where W' and H' denote the width and height of the heatmap, respectively.

b) Spatio-Semantic Dual Similarity Modeling: Based on the above definitions, we impose two layers of matching constraints:

- Feature Similarity ($\mathbf{S}_{feat}$): We employ cosine similarity to capture the consistency of appearance semantics:

$$\mathbf{S}_{feat}(i,j) = \frac{F_{ps}^{(i)} \cdot (F_{ks}^{(j)})^{\top}}{\|F_{ps}^{(i)}\|_2 \|F_{ks}^{(j)}\|_2} \tag{11}$$

- Spatial Similarity ($\mathbf{S}_{geo}$): We introduce a Bounded Linear Decay Field as a geometric constraint. Linear decay provides a hard boundary for spatial gating, effectively cutting off long-range noise, whereas Gaussian decay might still assign non-zero weights to distant background clutter. We calculate the Euclidean distance between visual node i and skeletal node j in the normalized space and define the affinity function as follows:

$$\mathbf{S}_{geo}(i,j) = \max\left(0, 1 - \|p_{vis}^{(i)} - p_{skl}^{(j)}\|_2\right) \tag{12}$$

Interpretation: Mathematically, Equation (12) defines a local response field with a radius of $R = 1$ centered at the visual node $p_{vis}^{(i)}$ (derived from the spatial grid coordinates of the i-th ViT patch). This field exhibits an isotropic distance decay property: the closer the physical distance, the higher the spatial prior weight. Once the distance exceeds the threshold, the similarity is hard-thresholded to 0. This mechanism acts as a Spatial Gating, forcing the model to disregard long-range erroneous matches that share similar appearance but violate human anatomical constraints.

c) Fusion and Alignment: Finally, the fusion matrix $\mathbf{S}_{total} = \beta\mathbf{S}_{feat} + (1-\beta)\mathbf{S}_{geo}$ guides the feature aggregation. We adopt a unidirectional greedy strategy to retrieve the optimal skeletal anchor $j^* = \arg\max_j \mathbf{S}_{total}(i,j)$ for each visual group i. Explicit semantic information is then injected into the visual features via a residual connection:

$$\hat{F}_{ps}^{(i)} = F_{ps}^{(i)} + \lambda_{align} F_{ks}^{(j^*)} \tag{13}$$

where the alignment weight λ_{align} is set to 1. This integration enforces rigorous spatio-semantic consistency, thereby resolving semantic ambiguity.

C. Geometry-Aware Semantic Decoder

1) Learnable Semantic View Construction: To decouple high-level semantics (e.g., head, torso) from the global representation, we define a set of learnable semantic queries $Z \in \mathbb{R}^{N_v\times D}$. In the cross-attention layer, Z acts as the Query, while the encoder features F_{En}, weighted by pose heatmaps, serve as the Key and Value. This interaction generates a set of part-level views $\mathcal{V}$ rich in semantic information.

2) Geometry-Aware Semantic Matching (GASM): Although the semantic view set $\mathcal{V}$ contains pure high-level semantic information, it lacks specific spatial details. The GASM module aims to accurately inject these semantic views into the visual features $\hat{F}_{ps}$ via a fusion mechanism. This process similarly relies on dual constraints of geometry and semantics. Unlike fixed grid anchors, our dynamic centroids adaptively shift towards the most relevant semantic regions (e.g., shifting the "head" centroid downwards if the person is crouching). To obtain the physical location of the semantic views, we first utilize the decoder's cross-attention map $\mathcal{A} \in \mathbb{R}^{N_v\times N_{patch}}$ as a probability density distribution to compute the Dynamic Centroid $p_{sem}^{(j)}$ for the j-th semantic view:

$$p_{sem}^{(j)} = \frac{\sum_{k=1}^{N_{patch}} \mathcal{A}_{j,k} \cdot \text{Coord}(k)}{\sum_{k=1}^{N_{patch}} \mathcal{A}_{j,k}} \tag{14}$$

where $\text{Coord}(k)$ represents the center coordinate of the k-th patch. This allows the semantic coordinates to adaptively track changes in human pose. Subsequently, to achieve precise matching between semantic views and visual features, we construct a dual-metric matrix $\mathbf{M}_{total} \in \mathbb{R}^{N_{patch}\times N_v}$ in the decoding stage. Similar to SSDA, this matrix comprises both semantic and geometric matching terms:

- Semantic Matching ($\mathbf{M}_{sem}$): We compute the cosine similarity between the visual feature $\hat{F}_{ps}^{(i)}$ and the semantic view v_j:

$$\mathbf{M}_{sem}(i,j) = \frac{\hat{F}_{ps}^{(i)} \cdot v_j^{\top}}{\|\hat{F}_{ps}^{(i)}\|_2 \|v_j\|_2} \tag{15}$$

- Geometric Matching ($\mathbf{M}_{geo}$): We calculate the spatial affinity based on the visual coordinate $p_{vis}^{(i)}$ and the dynamic semantic centroid $p_{sem}^{(j)}$:

$$\mathbf{M}_{geo}(i,j) = \max\left(0, 1 - \|p_{vis}^{(i)} - p_{sem}^{(j)}\|_2\right) \tag{16}$$

Finally, the fusion matrix $\mathbf{M}_{total} = \gamma\mathbf{M}_{sem} + (1-\gamma)\mathbf{M}_{geo}$ guides the feature aggregation. We employ a greedy matching

strategy for hard assignment. Specifically, for each visual patch i, we identify the optimal semantic view index k^* corresponding to the maximum combined similarity: $k^* = \arg\max_j \mathbf{S}_{total}(i,j)$. Subsequently, the retrieved semantic feature v_{k^*} is fused into the original visual representation via a residual connection. This process injects explicit semantic interpretability and discriminability into the underlying visual patches, yielding the final feature F_{ss}. Furthermore, leveraging the confidence scores from HRNet, we segregate the features into high-confidence human features F_T and low-confidence background features F_N.

3) Decoder Loss: The overall decoder loss function comprises the Cross-Entropy Classification Loss $\mathcal{L}_{id}$ and the Triplet Loss $\mathcal{L}_{tri}$. Generally, both objective functions are formulated based on the final decoded representations F_T to simultaneously optimize discriminative power and metric constraints.

4) Inference: For retrieval, we utilize a multi-view representation. Specifically, the encoder's global feature f_{gb} and local topological feature F_{ps}, along with the decoder's semantic feature F_T, are concatenated to serve as the final inference features.

IV. Experiments

A. Datasets and Evaluation Metrics

We evaluate our proposed framework on five mainstream benchmarks, comprising two occlusion-specific datasets (Occluded-Duke, Occluded-REID) and three holistic person Re-ID datasets (Market-1501, DukeMTMC-reID, MSMT17). To quantitatively assess performance, we employ the standard metrics: mean Average Precision (mAP) and Cumulative Matching Characteristics (CMC) at Rank-1.

B. Implementation Details

The model is built upon the ViT-Base architecture (hidden dimension of 768), utilizing a pre-trained HRNet-W32 as the skeleton extraction network. Input images are resized to 256×128. During training, the batch size is set to 16, comprising 4 instances per identity. We employ the SGD optimizer with an initial learning rate of 0.008, coupled with a cosine learning rate decay strategy. For coordinate normalization, skeletal coordinates are divided by the width and height of the feature map to map them into the $[0, 1]$ interval.

Adaptability Strategy for Occluded and Holistic Scenarios: To maximize adaptability across both occluded and holistic datasets, we conceptually categorize the Topology-Enhanced Features (F_{ks}) (Section III-B1) and the Decoder (Section III-C) as Uncertainty Modeling components (tailored for occlusion), while treating the remaining modules as General Modeling components. The rationale is twofold:

- Graph Aggregation Trade-off: F_{ks} encodes human graph structures where each node aggregates information from its neighbors. In occlusion scenarios, this neighbor aggregation effectively mitigates the impact of missing information. However, for fully visible (non-occluded) data, excessive aggregation may inadvertently dilute the discriminability of the original local features.

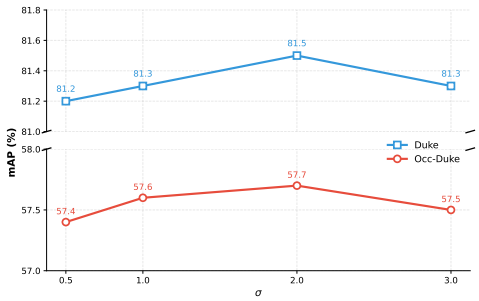

Fig. 3. Ablation studies on Gaussian standard deviation σ in SG-GCN.

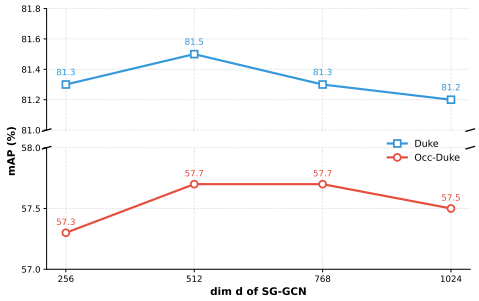

Fig. 4. Ablation studies on hidden dim d of SG-GCN.

- Semantic Matching Ambiguity: The decoder introduces learnable semantic blocks to match with $\hat{F}_{ps}$. For clean data where local human information is already distinct, this fusion might introduce unnecessary semantic ambiguity. Conversely, in occlusion scenarios, this mechanism is crucial for learning precise semantic part matching, thereby effectively attenuating the influence of occluded regions.

C. Ablation studies

For SG-GCN, we conduct ablation studies on the Gaussian standard deviation σ and the hidden layer dimension d. As shown in Figure 3, the performance improves as σ increases, peaking at $\sigma = 2.0$. This indicates that a moderate receptive field is optimal for aggregating local neighbors; too small limits context, while too large introduces noise. Similarly, Figure 4 illustrates that a hidden dimension of 512 yields the best trade-off between capacity and overfitting.

Analysis of SSDA Strategy: We explicitly investigate the hyperparameter β, which regulates the trade-off between semantic (S_{feat}) and spatial (S_{geo}) similarities. As plotted in Figure 5, the model achieves optimal performance at $\beta = 0.8 \sim 0.9$. Deviating from this balance—specifically, removing the spatial term entirely (equivalent to $\beta = 1.0$)—results in a distinct performance drop exceeding 0.5% in mAP under complex pose conditions. This observation yields a dual conclusion: while feature-based semantics remain the dominant factor for identity matching, the auxiliary geometric constraint is indispensable for filtering outliers and resolving structural ambiguities.

Decoder Components: 1) We examine the influence of the balancing parameter γ, which regulates the trade-off between

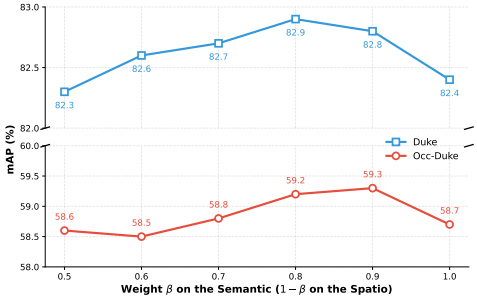

Fig. 5. Parameter analysis of SSDA.

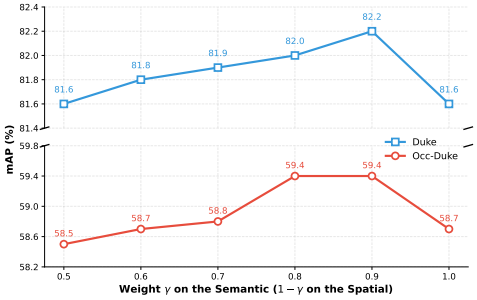

Fig. 6. Parameter analysis of GASM.

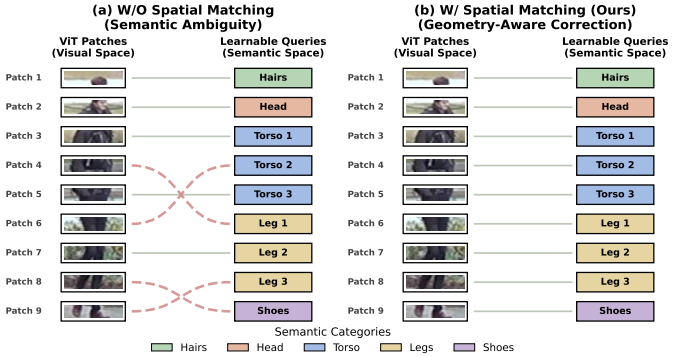

Fig. 7. Visualization of the bipartite matching between visual patches and semantic queries. (a) Relying solely on feature similarity without spatial guidance leads to severe misalignment and symmetric ambiguity (e.g., confusing left and right limbs, indicated by red dashed lines). (b) By incorporating spatial constraints, our Geometry-Aware Semantic Matching (GASM) effectively rectifies these errors (green solid lines), ensuring anatomically logical alignment.

the semantic and spatial terms. The results in Figure 6 show a similar trend to SSDA, where $\gamma = 0.9$ yields the best mAP. This confirms that the geometric term in the decoder serves as a robust "guide" rather than a primary feature, effectively preventing the attention mechanism from drifting to background regions. 2) Visualization results of the decoder's attention heatmaps demonstrate that the model can accurately focus on non-occluded human regions, retaining strong responsiveness to key body parts even in scenarios with severe occlusion. As illustrated in Figure 8, the visualization results are presented in image pairs, where each original input is immediately followed by its corresponding attention heatmap generated by our GASM module. Taking the scenario where the pedestrian's lower body is severely occluded by a car as an example, the heatmap demonstrates that our model effectively resists interference from the vehicle's metallic texture. Guided by the "head" and "torso" semantic queries, the high-response regions are precisely concentrated on the valid, visible parts of the target pedestrian, proving the module's capability to suppress occlusion noise. 3) The proposed GASM module effectively resolves semantic ambiguity during part-level alignment. Figure 7 visualizes the bipartite matching between visual patches and semantic queries. Without spatial guidance (Fig. 7(a)), relying solely on appearance similarity leads to severe semantic ambiguity. The network frequently confuses symmetric parts (e.g., left vs. right leg) or clothing with background, resulting in chaotic cross-matching (red dashed lines). In contrast, our geometry-aware matching (Fig. 7(b)) effectively resolves this. By anchoring queries with dynamic centroids and penalizing unreasonable spatial distances, GASM cleanly rectifies the tangled relationships into a logical topological mapping (green solid lines). This validates that geometric constraints are essential for robust part-level decoding under severe occlusions. 4) Additionally, we perform ablation experiments on the number (N_v) of learnable semantic queries ($Z \in R^{N_v \times D}$). Our findings show that this number should neither be excessively small (which leads to insufficient semantic expressiveness) nor overly large (which introduces noise due to semantic redundancy). As shown in Figure 9, the performance improves steadily as the number of semantic queries (N_v) increases from 8 to 27. This suggests that a sufficient number of queries is necessary to capture fine-grained body parts (e.g., distinguishing upper-arm from lower-arm). However, performance saturates and slightly drops when $N_v > 32$. We hypothesize that excessive queries introduce semantic redundancy, leading the decoder to overfit to trivial details rather than generalizable part features. Beyond this point, the performance plateaued or slightly decreased, suggesting that excessive semantic queries might introduce redundancy and noise.

As shown in Table I, the baseline ViT achieves 53% mAP on Occluded-Duke. The integration of SG-GCN yields a notable improvement of +4.7% mAP, verifying that reconstructing local topology is crucial for recovering occluded features. Furthermore, the inclusion of SSDA boosts performance to 59.2% mAP, demonstrating that resolving spatial misalignment significantly enhances the model's robustness against limb confusion caused by obstacles. The GASM module improves performance by +6.4%, proving the effectiveness of semantic-aware decoding. The full configuration achieves the best performance, validating that spatial-semantic alignment and geometric decoding are complementary rather than redundant.

Complexity Analysis: We further evaluate the computational efficiency of our framework. Although our method introduces additional GCN layers and cross-modal attention, the overhead remains manageable. Compared to the standard ViT-Base backbone (86.52M parameters), our proposed modules (SG-GCN, SSDA, GASM) add only 17.63M parameters (a ˜20% increase).

Fig. 8. Decoder's Attention Heatmap Visualization.

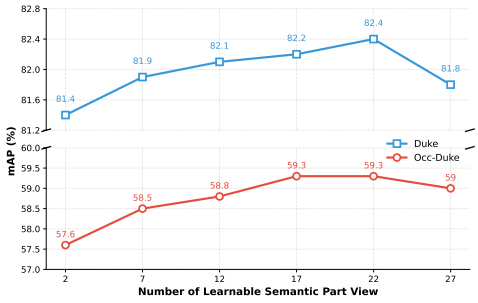

Fig. 9. Parameter analysis on the number (N_v) of learnable semantic queries.

D. Comparison with State-of-the-Art Methods

1) Performance on Occluded Datasets: As presented in Table II, our framework demonstrates superior performance against state-of-the-art competitors. Notably, our standard ViT-based model (SSGA) achieves **61% mAP** and **68.5% Rank-1** accuracy on the challenging Occluded-Duke dataset. Furthermore, when equipped with the CLIP backbone, our method (denoted as **SSGA**†‡) achieves **66.7% mAP** and **75.5% Rank-1**, securing the second-best performance and significantly outperforming most existing baselines.

- **Comparison with ViT-based Methods:** Compared to TransReID [17], which employs a pure Transformer architecture (55.7% mAP), our SSGA achieves a remarkable gain of **+5.3% mAP**. This improvement verifies that while standard ViTs excel at global context, they typically lack structural inductive biases. Our framework effectively bridges this gap through the **synergy of SG-GCN's topology repair and GASM's semantic decoding**, enabling robust feature extraction even under severe occlusion.
- **Comparison with Pose-Guided Methods:** Unlike PFD [18] (60.1% mAP), which relies on simple feature concatenation, our SSGA outperforms it by **+0.9% mAP**. This advantage stems from our **SSDA strategy**, which enforces rigorous alignment in both geometric and semantic spaces rather than blindly trusting noisy pose estimation. By dynamically integrating skeletal priors with visual semantics, our method offers superior interpretability and resilience to pose errors.
- **Impact of CLIP-based SOTA Performance:** As shown in the SSGA†‡ entries of Table II, incorporating the CLIP backbone further elevates our performance to a state-of-the-art level (e.g., **93.2% Rank-1** on Occluded-ReID). This validates that our proposed modules (SG-GCN, SSDA, GASM) are highly compatible with pre-trained vision-language models, effectively unleashing their potential in handling complex occlusion scenarios through explicit structural and semantic modeling.

TABLE I
COMPREHENSIVE ABLATION STUDIES ON OCCLUDED-DUKE/DUKEMTMCREID.

SG-GCN	SSDA	GASM	Occluded-Duke R1	Occluded-Duke mAP	DukeMTMCReID R1	DukeMTMCReID mAP
			60.6	53	88.7	79.3
✓			65.2	57.7	90.2	81.5
	✓		66.8	59.2	90.6	82.9
		✓	66.5	59.4	90.4	82.2
✓	✓		67.6	59.8	91	83
✓		✓	68.2	60	90.6	82.9
	✓	✓	68.1	60.2	90.6	83
✓	✓	✓	68.5	61	91.1	83.4

2) Performance on Holistic Datasets: A common drawback of occlusion-specialized models is performance degradation on clean datasets due to over-designing for missing parts, as rigid part-matching or excessive background suppression mechanisms often dilute global discriminative features. However, our SSGA framework effectively overcomes this dilemma. This robustness is attributed to our adaptability strategy, where the graph aggregation intensity is dynamically adjusted, ensuring that global discriminative features are preserved even when occlusion is absent.

As shown in Table III, our method maintains high performance on Market-1501 (**95.7% Rank-1**) and DukeMTMC-reID (**91.1% Rank-1**). Notably, when adopting the CLIP model as the backbone, our approach achieves state-of-the-art (SOTA) performance on these clean datasets.

To further verify the generalization capability of our framework under more complex holistic scenarios, we evaluate SSGA on MSMT17, one of the most challenging large-scale datasets with severe illumination and viewpoint variations. As shown in Table IV, we compare our framework against three categories of state-of-the-art methodologies:Comparison with Part-based Methods:

- Compared to classic part-based architectures that rely solely on visual patches (e.g., TransReID), our baseline SSGA achieves competitive performance (83.7% Rank-1, 65.2% mAP). With further enhancements (SSGA†‡), it reaches 85.2% Rank-1 and 72.5% mAP, outperforming recent advanced models such as CAM2Former and MLRAT. This demonstrates that our Geometry-Aware Semantic Decoder adaptively focuses on discriminative regions without destroying the global context in holistic images.
- Comparison with Auxiliary-based Methods: While our framework, similar to MTIPE and PFD, leverages auxiliary tools (e.g., pose estimators) to extract structural priors, it effectively overcomes their common vulnerability to upstream estimation errors. Instead of blindly trusting the extracted skeletal coordinates—which often leads to cascading misalignment when the pose estimator fails in complex scenarios—SSGA introduces a dual-metric matching strategy to dynamically verify and rectify spatial-semantic alignments. Consequently, our SSGA†‡ significantly outperforms the pose-guided PFD by +1.4%

in Rank-1 and +8.1% in mAP, demonstrating superior robustness against auxiliary noise.
- Comparison with CLIP-based Methods: Leveraging the cross-modal pre-training prior, models like CLIP-ReID and SGFNet achieve top-tier performance on MSMT17. While our framework is primarily optimized to resolve severe occlusions and topology destruction, our full version (SSGA†‡) still narrows the gap significantly, achieving a highly competitive 72.5% mAP.

In summary, the results across Market-1501, DukeMTMC-reID, and MSMT17 validate the excellent adaptability of the SSGA framework. It proves that the incorporation of SG-GCN, SSDA, and GASM not only dominates in occluded scenarios but also preserves robust global discriminability in holistic multi-camera networks.

3) Critical Analysis: Advantages and Limitations: **Advantages:** The primary strength of our framework lies in its *Structural-Semantic Synergy*. Unlike attention-based methods (e.g., FED) that implicitly learn to ignore occlusion, our method explicitly models human topology via SG-GCN and aligns it with semantic queries. This makes the model highly robust to non-target noise (e.g., obstacles overlapping with the person) and provides better interpretability via the generated semantic views.

TABLE II
COMPARISON WITH STATE-OF-THE-ART (SOTA) METHODS ON OCCLUDED-REID/OCCLUDED-DUKE DATASETS: * DENOTES MODELS WITH TRANSFORMER BACKBONE, † INDICATES ENCODERS WITH A SMALL-STEP SLIDING-WINDOW SETTING, AND ‡ REPRESENTS OUR FINAL VERSION WITH CLIP BACKBONE. BOLD VALUES = BEST PERFORMANCE, UNDERLINED VALUES = SECOND-BEST.

Methods	Occluded-ReID		Occluded-Duke	
	R1	mAP	R1	mAP
PVPM [19]	66.8	59.5	-	-
CAAO [20]	65.2	61.1	67.8	55.8
CAAO*	87.1	83.4	68.5	59.5
PFD* [18]	79.8	81.3	67.7	60.1
TransReID* [17]	70.2	67.3	64.2	55.7
TransReID*†	-	-	66.4	59.2
PAT* [21]	81.6	72.1	64.5	53.6
FED* [2]	86.3	79.3	68.1	56.4
DPM* [22]	80.2	75.2	66.7	57.2
SGFNet* [23]	<u>93</u>	**90.3**	**76.5**	**67.2**
THCB-Net* [24]	87.3	84.5	72.3	62.6
RGANet* [25]	86.4	80	71.6	62.4
ProFD* [26]	92.3	**90.3**	70.6	63.1
SPT* [27]	87.8	81.1	74.7	63
MLRAT* [28]	88.9	83.6	73.3	63.1
SSGA(Ours)	81	82.4	68.5	61
SSGA†	82.2	83.7	70	62.4
SSGA†‡	**93.2**	<u>89.1</u>	<u>75.5</u>	<u>66.7</u>

Limitations: Despite the promising results, our framework has two primary limitations. **First**, it relies on the keypoints generated by the upstream HRNet. In extreme scenarios, such as low-light environments or heavy motion blur, where the pose estimator fails to detect any valid keypoints, the efficacy of our SSDA module may degrade. **Second**, regarding computational complexity, compared to single-stream baselines (e.g., Baseline ViT), our multi-modal architecture introduces additional overhead. This cost primarily stems from the dual-stream processing and the matrix operations required for dynamic graph construction. While accurate, this heavy pipeline poses challenges for resource-constrained edge devices. **Future work** will explore implicitly embedding structural priors into the visual encoder to reduce the dependency on external estimators, alongside investigating knowledge distillation to further compress the model for real-time applications.

TABLE III
COMPARISON WITH STATE-OF-THE-ART (SOTA) METHODS ON MARKET1501/DUKEMTMCREID

Methods	Market1501		DukeMTMC	
	R1	mAP	R1	mAP
CAAO [20]	95.1	87.3	88.9	77.5
CAAO*	95.3	88	89.8	80.9
PFD* [18]	95.5	89.6	90.6	82.2
TransReID* [17]	95	88.2	89.6	80.6
PAT* [21]	95.4	88	88.8	78.2
FED* [2]	95	86.3	89.4	78
SGFNet* [23]	<u>96.2</u>	<u>91.1</u>	**92.8**	<u>84.1</u>
THCB-Net* [24]	<u>96.2</u>	90.6	91.7	83.5
RGANet* [25]	95.5	89.8	-	-
ProFD* [26]	95.6	90.8	92.1	84
SPT* [27]	95.5	89.4	91.1	82.4
MLRAT* [28]	95.8	89.9	-	-
SSGA(Ours)	95.7	89.9	91.1	83.4
SSGA†	96	90.2	91.5	83.5
SSGA†‡	**96.5**	**92.7**	<u>92.2</u>	**84.4**

TABLE IV
COMPARISON WITH STATE-OF-THE-ART (SOTA) METHODS ON MSMT17

Methods	MSMT17	
	R1	mAP
RFCnet* [29] (TPAMI21)	82.2	60
SSPEM* [30] (ESWA24)	84	64.6
TransReID* [17] (ICCV21)	83.3	64.9
MAHATMA* [31] (TCSVT25)	85.6	68.1
MVI2P* [32] (IF24)	83.9	61.4
MLRAT* [28] (NN24)	85.3	67.6
CAM2Former* [33] (IF25)	85.8	68.1
MTIPE* [34] (PR25)	80.4	61.1
PFD* [18] (AAAI22)	83.8	64.4
CLIP-ReID* [35] (AAAI23)	<u>89.7</u>	**75.8**
RGANet* [25] (TIFS23)	88.1	72.3
SGFNet* [23] (TIFS25)	**90**	<u>74.4</u>
SSGA(Ours)	83.7	65.2
SSGA†	84.4	65
SSGA†‡	85.2	72.5

V. CONCLUSION

In this paper, we present SSGA, a unified Spatial Topology Repair and Semantic Geometry Alignment framework designed to tackle the formidable challenges of Occluded Person Re-Identification. To address the inherent structural destruction caused by obstacles, we first introduce the SG-GCN module, which effectively restores compromised local topology by embedding physical spatial priors into visual representations. More importantly, we decouple the cross-modal alignment and feature decoding processes into two distinct synergistic mechanisms. On one hand, the proposed SSDA strategy establishes a rigorous spatio-semantic dual-metric

constraint, successfully rectifying cross-modal misalignment and eliminating symmetrical limb confusion without blindly trusting noisy pose estimators. On the other hand, the GASM module utilizes learnable queries anchored by geometric centroids to dynamically decode part-level semantics, compelling the network to precisely focus on visible body regions while filtering out occlusion artifacts. Extensive evaluations across multiple benchmarks validate that our SSGA framework not only establishes new state-of-the-art performance in severe occlusion scenarios but also preserves robust global discriminability on holistic datasets, offering a highly adaptable and interpretable solution for real-world retrieval applications.

REFERENCES

[1] C. Zhao, Z. Qu, X. Jiang, Y. Tu, and X. Bai, "Content-adaptive auto-occlusion network for occluded person re-identification," *IEEE Transactions on Image Processing*, vol. 32, pp. 4223–4236, 2023.

[2] Z. Wang, F. Zhu, S. Tang, R. Zhao, L. He, and J. Song, "Feature erasing and diffusion network for occluded person re-identification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 4754–4763.

[3] V. Somers, C. De Vleeschouwer, and A. Alahi, "Body part-based representation learning for occluded person re-identification," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2023, Conference Proceedings, pp. 1613–1623.

[4] Y. Li, Y. Liu, H. Zhang, C. Zhao, Z. Wei, and D. Miao, "Occlusion-aware transformer with second-order attention for person re-identification," *IEEE Transactions on Image Processing*, vol. 33, pp. 3200–3211, 2024.

[5] T. Ren, Q. Lian, and J. Chen, "Boosting occluded person re-identification by leveraging occlusion attributes," *Information Sciences*, vol. 701, p. 121866, 2025.

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[7] T. Kipf, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[8] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio *et al.*, "Graph attention networks," *stat*, vol. 1050, no. 20, pp. 10–48 550, 2017.

[9] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," *Advances in neural information processing systems*, vol. 30, 2017.

[10] R. Sun, L. Chen, L. Zhang, R. Xie, and J. Gao, "Robust visible-infrared person re-identification based on polymorphic mask and wavelet graph convolutional network," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 2800–2813, 2024.

[11] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *European conference on computer vision*. Springer, 2016, pp. 483–499.

[12] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 466–481.

[13] Z. Geng, C. Wang, Y. Wei, Z. Liu, H. Li, and H. Hu, "Human pose as compositional tokens," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 660–671.

[14] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, and X. Wang, "Deep high-resolution representation learning for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3349–3364, 2020.

[15] S. Dou, C. Zhao, X. Jiang, S. Zhang, W.-S. Zheng, and W. Zuo, "Human co-parsing guided alignment for occluded person re-identification," *IEEE Transactions on Image Processing*, vol. 32, pp. 458–470, 2022.

[16] V. Somers, A. Alahi, and C. D. Vleeschouwer, "Keypoint promptable re-identification," in *European Conference on Computer Vision*. Springer, 2024, Conference Proceedings, pp. 216–233.

[17] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "Transreid: Transformer-based object re-identification," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 15 013–15 022.

[18] T. Wang, H. Liu, P. Song, T. Guo, and W. Shi, "Pose-guided feature disentangling for occluded person re-identification based on transformer," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, no. 3, 2022, pp. 2540–2549.

[19] S. Gao, J. Wang, H. Lu, and Z. Liu, "Pose-guided visible part matching for occluded person reid," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 744–11 752.

[20] C. Zhao, Z. Qu, X. Jiang, Y. Tu, and X. Bai, "Content-adaptive auto-occlusion network for occluded person re-identification," *IEEE Transactions on Image Processing*, vol. 32, pp. 4223–4236, 2023.

[21] Y. Li, J. He, T. Zhang, X. Liu, Y. Zhang, and F. Wu, "Diverse part discovery: Occluded person re-identification with part-aware transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 2898–2907.

[22] L. Tan, P. Dai, R. Ji, and Y. Wu, "Dynamic prototype mask for occluded person re-identification," in *Proceedings of the 30th ACM international conference on multimedia*, 2022, pp. 531–540.

[23] G. Lin, S. Yang, W.-S. Zheng, Z. Li, and Z. Huang, "A semantically guided and focused network for occluded person re-identification," *IEEE Transactions on Information Forensics and Security*, 2025.

[24] C. Wang, S. He, M. Wu, S.-K. Lam, P. Tiwari, and X. Gao, "Looking clearer with text: A hierarchical context blending network for occluded person re-identification," *IEEE Transactions on Information Forensics and Security*, 2025.

[25] S. He, W. Chen, K. Wang, H. Luo, F. Wang, W. Jiang, and H. Ding, "Region generation and assessment network for occluded person re-identification," *IEEE transactions on information forensics and security*, vol. 19, pp. 120–132, 2023.

[26] C. Cui, S. Huang, W. Song, P. Ding, M. Zhang, and D. Wang, "Profd: Prompt-guided feature disentangling for occluded person re-identification," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 1583–1592.

[27] L. Tan, J. Xia, W. Liu, P. Dai, Y. Wu, and L. Cao, "Occluded person re-identification via saliency-guided patch transfer," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, no. 5, 2024, pp. 5070–5078.

[28] G. Lin, Z. Bao, Z. Huang, Z. Li, W.-s. Zheng, and Y. Chen, "A multi-level relation-aware transformer model for occluded person re-identification," *Neural Networks*, vol. 177, p. 106382, 2024.

[29] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, "Feature completion for occluded person re-identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 4894–4912, 2021.

[30] Y. Pang, H. Zhang, L. Zhu, D. Liu, and L. Liu, "Self-similarity guided probabilistic embedding matching based on transformer for occluded person re-identification," *Expert Systems with Applications*, vol. 237, p. 121504, 2024.

[31] G. Zhang, Y. Yang, Y. Zheng, G. Martin, and R. Wang, "Mask-aware hierarchical aggregation transformer for occluded person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 35, no. 6, pp. 5821–5832, 2025.

[32] N. Dong, S. Yan, H. Tang, J. Tang, and L. Zhang, "Multi-view information integration and propagation for occluded person re-identification," *Information Fusion*, vol. 104, p. 102201, 2024.

[33] Z. Tan, G. Zhang, Z. Tan, P. Tiwari, Y. Wang, and Y. Yang, "Cam2former: Fusion of camera-specific class activation map matters for occluded person re-identification," *Information Fusion*, vol. 120, p. 103011, 2025.

[34] T.-T. Yuan, Q.-L. Shu, S.-B. Chen, L.-L. Huang, and B. Luo, "Instant pose extraction based on mask transformer for occluded person re-identification," *Pattern Recognition*, vol. 159, p. 111082, 2025.

[35] S. Li, L. Sun, and Q. Li, "Clip-reid: exploiting vision-language model for image re-identification without concrete text labels," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, 2023, Conference Proceedings, pp. 1405–1413.