RETHINKING DISTILLATION IN PEDESTRIAN RE-IDENTIFICATION: A SPATIAL AGGREGATION PERSPECTIVE

Anonymous authors
Paper under double-blind review

ABSTRACT

Pedestrian re-identification (Re-ID) is an open-set retrieval task in which practical systems often require either improving retrieval performance under fixed computational constraints or reducing inference cost with limited accuracy degradation. To address these requirements through knowledge distillation, we revisit which teacher signal should be transferred for Re-ID. We find that large changes in individual descriptor values do not necessarily imply comparable degradation of retrieval-relevant structure, motivating supervision of how descriptors are formed rather than their exact values. In Vision Transformer (ViT)-based Re-ID models, late-block `[CLS]`-to-patch attention provides an observable proxy for how spatial evidence is aggregated into the global representation. We study spatial aggregation as an effective teacher signal for open-set Re-ID, focusing on how the retrieval representation is formed rather than on exact agreement with teacher outputs. Across multiple ViT-based Re-ID frameworks and benchmarks, spatial aggregation transfer consistently improves both same-capacity refinement and lightweight student distillation, requiring only a teacher model and standard Re-ID training data, without additional annotations or inference-time components.

1 INTRODUCTION

Pedestrian re-identification (Re-ID) is an open-set retrieval task that aims to retrieve images of the same person from a large gallery despite variations in camera, viewpoint, pose, and time (Zheng et al., 2015; Wei et al., 2018; Ye et al., 2022). In deployment settings, model updates typically aim either *to improve retrieval accuracy under a fixed inference budget* or *to reduce computation with limited performance degradation*. Knowledge distillation (KD) provides a natural framework for both scenarios, where a deployed model serves as the teacher and a new model is trained as the student (Hinton et al., 2015; Gou et al., 2020). A central question, then, is what aspect of the teacher should be distilled for open-set Re-ID.

Existing KD methods typically transfer teacher predictions or representations, for example through logits or final features (Hinton et al., 2015; Romero et al., 2015; Heo et al., 2019; Zhao et al., 2022). However, Re-ID is an open-set retrieval task in which training and test identities are disjoint (Zheng et al., 2016; Ye et al., 2022). Retrieval therefore depends on the relative geometry and ranking induced by the learned descriptors, rather than on reproducing particular feature values or fixed identity prototypes. This motivates us to ask whether supervision should target how a descriptor is formed, rather than its exact values.

To investigate this hypothesis, we conduct an intervention study on TransReID, a representative ViT-based Re-ID architecture that progressively aggregates image information into the `[CLS]` token (Dosovitskiy et al., 2021; He et al., 2021; Luo et al., 2021). We replace the `[CLS]` token at different depths during inference and examine the resulting retrieval behavior. Interestingly, late-block interventions retain substantially higher retrieval performance than early-block interventions, consistent with better preservation of the retrieval-relevant inter-image structure. As shown in Fig. 1a, however, this difference cannot be explained by individual descriptor changes alone, since substantial feature shifts occur even when retrieval performance remains high. In contrast, Fig. 1b and Fig. 1c show that retrieval performance more closely tracks changes in `[CLS]` attention patterns.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

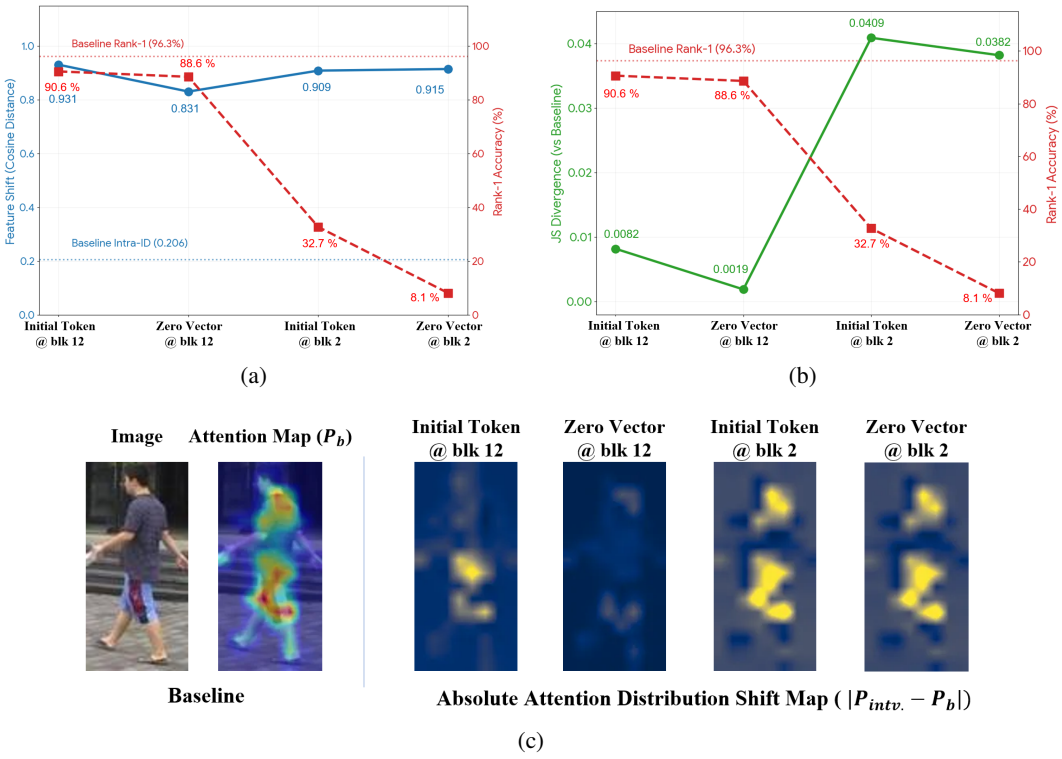

Figure 1: **`[CLS]` token intervention study** on TransReID trained on Market-1501. At inference time, the `[CLS]` token is replaced at different depths with either the initial `[CLS]` token or a zero vector. The red line indicates Rank-1 accuracy. (a) Feature distance between original and intervened descriptors for the same image. (b) Jensen–Shannon divergence between original and intervened attention maps for the same image (Lin, 1991). (c) Representative attention-difference maps. Brighter colors indicate larger attention-distribution differences.

Appendix A.1 provides additional random-vector and intermediate-block interventions, along with quantitative measures of retrieval-relevant inter-image geometry.

These observations motivate a *spatial aggregation* view of Re-ID distillation. For ViT-based Re-ID models, the aggregation of spatial evidence into the `[CLS]` representation is naturally exposed through `[CLS]`-to-patch attention, providing an architecture-native observable proxy (Dosovitskiy et al., 2021; He et al., 2021; Li et al., 2024; Park & Kim, 2022). We therefore instantiate spatial aggregation transfer by distilling re-normalized late-block `[CLS]`-to-patch attention, without directly matching logits or descriptor values. Our main contributions are summarized as follows:

- **A spatial aggregation perspective on ViT-based Re-ID:** We show that large changes in individual descriptor values do not necessarily imply comparable degradation of retrieval structure, and provide evidence that late `[CLS]` attention patterns are closely associated with retrieval-relevant geometry.

- **`[CLS]` attention as an instantiation of spatial aggregation transfer:** Motivated by this analysis, we use re-normalized late-block `[CLS]`-to-patch attention as an observable spatial aggregation signal for guiding students without matching logits or representation values.

- **Validation across self-contained deployment-oriented settings:** We evaluate spatial aggregation transfer across multiple ViT-based Re-ID frameworks, covering both same-architecture refinement and lightweight student distillation using only teacher weights and standard Re-ID training data, without external data or additional supervision.

2 Related Work

2.1 Deployment-Oriented ViT-based Re-ID

ViTs have become strong backbones for pedestrian Re-ID, with representative methods such as TransReID adapting ViTs to retrieval through Re-ID-specific designs including camera-aware embeddings, patch-level augmentation, and part-aware representations (Dosovitskiy et al., 2021; He et al., 2021; Ni et al., 2023; Ye et al., 2025). Subsequent works further improve performance through stronger pretraining, token refinement, or enhanced training strategies (Luo et al., 2021; Fu et al., 2021; Hu et al., 2025; He et al., 2022; Chen et al., 2021; Caron et al., 2021). These directions are effective and continue to advance the upper bound of Re-ID performance. In contrast, this work focuses on a complementary deployment-oriented setting, where an already trained Re-ID model and the standard training data are available. In this setting, the goal is either to improve retrieval accuracy under the same inference budget or to train a compact student with limited performance degradation (Gou et al., 2020). We address this setting through spatial aggregation transfer, enabling both same-capacity refinement and model compression without external data, annotations, or auxiliary supervision.

2.2 Attention Maps as a Carrier of Model Knowledge for Open-Set Re-ID

Prior KD methods have explored teacher-derived supervision from logits, features, relational structures, similarity-preserving representations, contrastive representations, and activation-based attention (Hinton et al., 2015; Romero et al., 2015; Park et al., 2019; Tung & Mori, 2019; Tian et al., 2020; Zagoruyko & Komodakis, 2017). Recent work suggests that ViT attention maps carry substantial information about inter-token relationships and can explain a large part of the gains obtained through fine-tuning (Li et al., 2024; Park & Kim, 2022; Park et al., 2023; Raghu et al., 2021). While this finding shows that attention can reveal knowledge acquired during fine-tuning, our focus is different: we ask what teacher behavior is most relevant to open-set retrieval. In Re-ID, training and test identities are disjoint, so the final feature values learned on training identities are not necessarily the most appropriate transfer target (Zheng et al., 2016; Ye et al., 2022; Kornblith et al., 2019). Instead, our analysis suggests that retrieval performance is closely tied to how identity-relevant spatial evidence is aggregated into the global representation. From this perspective, attention is useful not simply as another intermediate representation to match, but as an observable signal of spatial aggregation behavior (Li et al., 2024; Jelassi et al., 2022). We therefore focus on late-block `[CLS]`-to-patch attention, which provides an observable view of how spatial evidence is routed into the global representation (Dosovitskiy et al., 2021; He et al., 2021; Li et al., 2024).

3 Spatial Aggregation Transfer

3.1 Background and Problem Setup

3.1.1 Vision Transformer (ViT) Preliminaries

We consider ViT-based Re-ID models that use the `[CLS]` token as the global image representation (Dosovitskiy et al., 2021; He et al., 2021). Given an image $x \in \mathbb{R}^{H\times W\times C}$ divided into N patches, a learnable `[CLS]` token is prepended to the patch embeddings, and we denote the input and block-ℓ token sequences as:

$$Z^0 = [z^0_{\text{cls}}; z^0_1; \dots; z^0_N], \qquad Z^\ell = [z^\ell_{\text{cls}}; z^\ell_1; \dots; z^\ell_N] \in \mathbb{R}^{(N+1)\times D}. \tag{1}$$

Within each block, self-attention performs inter-token interaction by projecting $Z^{\ell-1}$ into query, key, and value matrices Q^ℓ, K^ℓ, and V^ℓ (Vaswani et al., 2017; Dosovitskiy et al., 2021; Li et al., 2024):

$$\text{Self-Attention}(Q^\ell, K^\ell, V^\ell) = A^\ell V^\ell, \quad \text{where} \quad A^\ell = \text{softmax}\left(\frac{Q^\ell (K^\ell)^\top}{\sqrt{d}}\right). \tag{2}$$

Here, $A^\ell \in \mathbb{R}^{(N+1)\times(N+1)}$ denotes the attention map, with each row forming a probability distribution over tokens. The `[CLS]` attention pattern therefore provides an observable view of how spatial evidence is routed into the global representation.

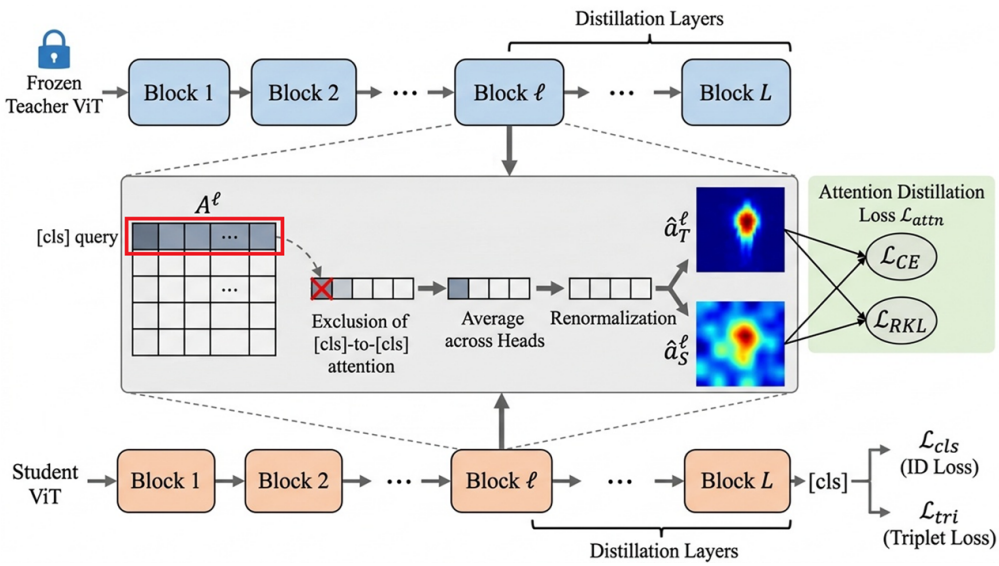

Figure 2: **Overview of the spatial aggregation transfer framework.** The [CLS] query row is extracted from late-block attention maps, the self-attention entry is excluded, attention values are averaged across heads, and the resulting distribution is re-normalized. The student is supervised to preserve teacher-relevant spatial aggregation patterns using cross-entropy and reverse KL divergence.

3.1.2 Distillation Setting

We consider a teacher–student scenario aimed at improving either inference efficiency or discriminative performance. Let f^T denote a teacher model pretrained on a large-scale source dataset $\mathcal{D}_{\text{src}}$ and fine-tuned on the target Re-ID dataset $\mathcal{D}_{\text{tar}}$. We use f^T as the trained reference model under the corresponding framework, serving as the deployment-ready reference. The student f^S is initialized using the corresponding pretrained initialization under the same framework and is trained on $\mathcal{D}_{\text{tar}}$ under the guidance of f^T.

Notably, this distillation framework is entirely self-contained: it requires no additional large-scale models or external datasets beyond those already utilized for the teacher's development. The student simply leverages the task-adapted knowledge acquired by the teacher during its prior fine-tuning on the target domain. Following standard practice in deployment-oriented Re-ID, we adopt ViT-B as our default reference architecture for the teacher and baseline models (He et al., 2021; Luo et al., 2021; Hu et al., 2025).

3.2 Transfer Signal and Objective

3.2.1 [CLS] Token Attention as Spatial Aggregation Supervision

As illustrated in Fig. 2, we instantiate spatial aggregation transfer using the self-attention behavior of the [CLS] token. During training, the teacher model is frozen, and both the teacher and student process the same input image. Let A_T^ℓ and A_S^ℓ denote the teacher and student self-attention maps at transformer block ℓ, respectively.

For each selected block, we extract the attention row corresponding to the [CLS] query from both A_T^ℓ and A_S^ℓ. We keep only the patch-token entries in this row by excluding the attention entry corresponding to the [CLS] key, and re-normalize the remaining values to form a valid [CLS]-to-patch attention distribution. This distribution serves as an observable proxy for spatial aggregation and is used as our distillation target. The detailed block, head alignment strategy, and loss formulation are presented in the next subsection.

3.2.2 Spatial Aggregation Transfer Loss

We apply attention distillation only to the later transformer blocks. As observed empirically and illustrated in Appendix A.2, the [CLS] token attention patterns in early blocks are less informative for supervision, whereas later blocks capture more human-related and task-relevant semantics. Therefore, by default, we use only the last half of the transformer blocks for distillation; in ViT-B and ViT-S, this corresponds to the final six blocks. We focus on depth-aligned ViT teacher–student pairs and apply the loss to corresponding late blocks; a preliminary heterogeneous-depth extension is provided in Appendix A.6.

For each selected block, we use the [CLS]-to-patch attention distribution defined in the previous subsection. Since the attention entry corresponding to the [CLS] token itself is excluded, the remaining patch-token attention values are re-normalized to form a valid probability distribution. In addition, we average the attention maps across heads before computing the distillation loss, rather than enforcing head-wise correspondence. This gives a single target distribution per block and allows the same formulation to be used for both same-architecture and cross-architecture teacher–student pairs, including cases with different numbers of attention heads. We use this head-averaged target as a simple and consistent representation of [CLS] spatial aggregation.

Let $\hat{a}_T^\ell$ and $\hat{a}_S^\ell$ denote the re-normalized teacher and student [CLS]-to-patch attention distributions at block ℓ, respectively. We match these distributions using a combination of cross-entropy ($\mathcal{L}_{\mathrm{CE}}$) and reverse KL divergence ($\mathcal{L}_{\mathrm{RKL}}$) (Kullback & Leibler, 1951):

$$\mathcal{L}_{\mathrm{CE}}^\ell = -\sum_i \hat{a}_{T,i}^\ell \log \hat{a}_{S,i}^\ell, \qquad \mathcal{L}_{\mathrm{RKL}}^\ell = \sum_i \hat{a}_{S,i}^\ell \log \frac{\hat{a}_{S,i}^\ell}{\hat{a}_{T,i}^\ell}. \tag{3}$$

Since cross-entropy is weighted by the teacher distributions, it mainly emphasizes positions with high teacher probability and is relatively less sensitive to mismatches in low-probability regions. In Re-ID, however, such low-attention regions are also important, as they often correspond to background or other non-discriminative context that should be suppressed. By incorporating $\mathcal{L}_{\mathrm{RKL}}$, we additionally penalize cases where the student assigns high attention to regions receiving low teacher attention.

The overall training objective $\mathcal{L}$ integrates these patterns across the set of selected transformer blocks $\mathcal{B}$ alongside standard Re-ID supervision:

$$\mathcal{L} = \mathcal{L}_{\mathrm{cls}} + \mathcal{L}_{\mathrm{tri}} + \underbrace{\frac{1}{|\mathcal{B}|}\sum_{\ell\in\mathcal{B}} \left(\mathcal{L}_{\mathrm{CE}}^\ell + \mathcal{L}_{\mathrm{RKL}}^\ell\right)}_{\mathcal{L}_{\mathrm{attn}}}, \tag{4}$$

where $\mathcal{L}_{\mathrm{cls}}$ and $\mathcal{L}_{\mathrm{tri}}$ denote the standard classification and triplet losses used in Re-ID (Luo et al., 2019; He et al., 2021), $\mathcal{L}_{\mathrm{attn}}$ denotes the spatial aggregation transfer loss, and $\mathcal{B}$ denotes the set of selected transformer blocks. This formulation is mathematically equivalent to a symmetric KL-based objective up to a constant independent of the student (Jeffreys, 1946; Kullback & Leibler, 1951), but it admits a more interpretable decomposition. Cross-entropy guides the student toward teacher-attended regions, while reverse KL suppresses attention on teacher-inconsistent regions. We use a fixed unit weight for $\mathcal{L}_{\mathrm{attn}}$ throughout all experiments.

4 Experiments

4.1 Experimental Settings

We evaluate spatial aggregation transfer on three ViT-based Re-ID frameworks: TransReID, TransReID-SSL, and PersonViT (He et al., 2021; Luo et al., 2021; Hu et al., 2025), which span different backbone pretraining sources and strategies, but share the same supervised Re-ID training objectives and evaluation protocol, while exhibiting different baseline performance levels (Krizhevsky et al., 2012; Fu et al., 2021; He et al., 2022; Caron et al., 2021; Chen et al., 2021). This tests whether late-block [CLS] attention provides a consistent spatial aggregation signal across standard and stronger ViT-based Re-ID models. For TransReID, we use the plain ViT backbone configuration without additional architecture-specific modules, following recent practice for controlled

Table 1: **Evaluation of spatial aggregation transfer for Re-ID.** Results are reported for ViT-based Re-ID models under ViT-B → B and ViT-B → S transfer settings on five standard benchmarks.

Framework	Setting	Market-1501		DukeMTMC		MSMT17		CUHK03		NightReID	
		mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1
TransReID	ViT-B (teacher)	87.7	94.8	81.2	90.6	66.3	85.2	74.7	77.3	40.8	57.7
	ViT-S	83.6	92.5	76.6	87.4	59.5	80.4	64.0	66.4	33.9	48.5
	ViT-B → B	**88.2**	**95.0**	**81.7**	90.6	**67.6**	**86.0**	**76.7**	**80.1**	**42.4**	**59.0**
	ViT-B → S	**84.3**	**93.1**	**78.0**	**88.5**	**59.8**	**80.6**	**68.7**	**71.7**	**37.0**	**51.4**
TransReID-SSL	ViT-B (teacher)	92.7	96.3	83.4	91.7	73.2	88.7	85.1	86.6	51.9	66.5
	ViT-S	91.2	96.1	82.0	90.4	68.5	86.2	81.8	84.1	49.1	64.6
	ViT-B → B	**92.8**	**96.4**	**83.7**	91.7	**74.0**	**89.1**	**86.0**	**87.7**	**53.3**	**68.1**
	ViT-B → S	**92.0**	**96.2**	**82.7**	**90.9**	**71.6**	**88.1**	**83.5**	**85.6**	**52.8**	**68.4**
PersonViT	ViT-B (teacher)	94.9	97.5	88.0	93.8	80.8	92.0	89.4	90.1	63.1	75.0
	ViT-S	93.0	96.9	84.7	92.0	74.4	88.9	87.0	88.4	58.0	71.3
	ViT-B → B	94.9	**97.6**	88.0	93.8	**81.0**	**92.2**	**89.8**	**90.7**	**64.0**	**75.4**
	ViT-B → S	**93.8**	**97.2**	**86.0**	**92.9**	**76.9**	**90.5**	**87.6**	**89.0**	**61.0**	**74.2**

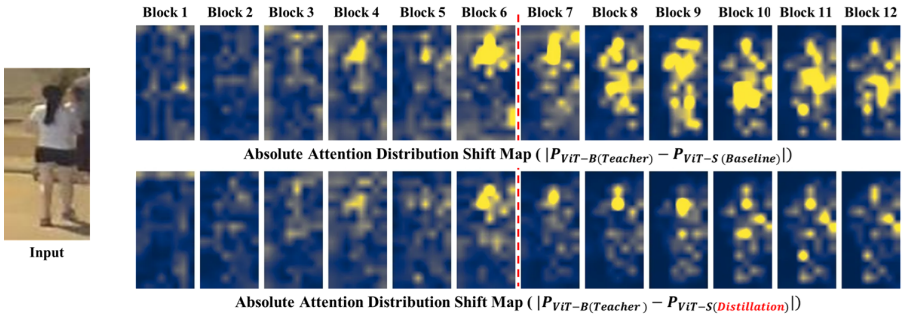

Figure 3: **Visualization of attention maps and absolute attention distribution shift** on NightReID. Brighter colors indicate larger attention-distribution differences. The upper row shows the absolute difference between ViT-B (teacher) and ViT-S (baseline). The lower row shows the absolute difference between ViT-B (teacher) and ViT-S trained with spatial aggregation transfer.

ViT-based Re-ID evaluation (He et al., 2021; Ye et al., 2025). This isolates spatial aggregation supervision from Re-ID-specific engineering components. For each framework, we consider both same-capacity refinement and lightweight student distillation. The teacher is a ViT-B model trained under the corresponding framework, while the student is another ViT-B or ViT-S. We evaluate on five standard RGB person Re-ID benchmarks: Market-1501, DukeMTMC-reID, MSMT17, CUHK03, and NightReID (Zheng et al., 2015; Gou et al., 2017; Wei et al., 2018; Li et al., 2014; Zhao et al., 2025). For MSMT17, we follow the setting without Mosaic augmentation. Full implementation details are provided in Appendix A.3.

4.2 Main Results

Tab. 1 shows that spatial aggregation transfer consistently improves the respective baselines across all three ViT-based Re-ID frameworks. In the cross-architecture setting, ViT-B → S improves the ViT-S student on every benchmark, supporting late `[CLS]` attention as an effective signal for transferring spatial aggregation behavior to compact models. Importantly, ViT-B → B also improves over the ViT-B baseline, suggesting that spatial aggregation supervision provides a beneficial optimization bias even without capacity transfer. Teacher attention explicitly guides the student toward teacher-relevant spatial regions, while allowing it to learn its own Re-ID embedding under the standard objective. The consistent gains on challenging datasets such as MSMT17 and NightReID further indicate that spatial aggregation guidance is useful under large appearance variation and difficult visual conditions. Additional random-seed stability results are provided in Appendix A.5.

Table 2: **Comparison of distillation methods** on TransReID-SSL. Performance is reported as ViT-B / ViT-S; all distillation experiments use a ViT-B teacher. "–" denotes a setting not directly supported by the original formulation.

Method	MSMT17		NightReID	
	mAP	R-1	mAP	R-1
Baseline - trained alone (ViT-B/ViT-S)	73.2 / 68.5	88.7 / 86.2	51.9 / 49.1	66.5 / 64.6
Logits	73.7 / 70.6	88.9 / 87.6	52.8 / 51.4	66.4 / 66.6
Final features	73.3 / 68.0	88.7 / 85.6	51.8 / 49.0	67.1 / 64.3
SPKD (Tung & Mori, 2019)	73.8 / 71.0	88.7 / 87.2	53.2 / 51.3	67.4 / 65.7
FMD (Hao et al., 2022)	73.6 / 69.3	88.8 / 86.9	53.0 / 50.8	68.0 / 65.8
ViT Attention Transfer (Li et al., 2024)	73.6 / –	88.9 / –	52.8 / –	67.3 / –
Spatial Aggregation (Ours)	**74.0 / 71.6**	**89.1 / 88.1**	**53.3 / 52.8**	**68.1 / 68.4**

As shown in Fig. 3, the distilled student produces attention maps that more closely align with the ViT-B teacher, with a substantially smaller discrepancy than the baseline. This verifies that the supervision affects the intended spatial aggregation behavior, guiding the student toward teacher-relevant regions. Interestingly, although distillation is applied only to the later blocks (Blocks 7–12), partial alignment is also observed in earlier blocks, suggesting that its effect propagates beyond the directly supervised layers. More qualitative results on Re-ID and attention map changes are provided in Appendix A.7.

5 Ablations and Analysis

In the following, we conduct ablation studies to examine which components support spatial aggregation transfer. For computational efficiency and consistent comparison, all ablations are performed on the TransReID-SSL framework, which represents a mid-performance setting among the evaluated models (Luo et al., 2021). We focus on the challenging benchmarks MSMT17 and NightReID, where the effect of spatial aggregation supervision is more pronounced under large appearance variation and difficult visual conditions (Wei et al., 2018; Zhao et al., 2025).

5.1 Comparison of Teacher Signals and Distillation Methods

This experiment examines whether spatial aggregation provides effective supervision for open-set Re-ID compared with both conventional teacher signals and stronger structure-preserving distillation methods. In addition to classification logits and final feature embeddings, we compare with SPKD (Tung & Mori, 2019), which preserves pairwise feature similarities, FMD (Hao et al., 2022), which transfers token-level manifold structure, and the original ViT attention-transfer method of Li et al. (Li et al., 2024). The original attention-transfer method is evaluated only for ViT-B $\rightarrow$ B because it requires direct correspondence between layers and attention heads.

As shown in Tab. 2, spatial aggregation transfer outperforms the compared alternatives in each available setting. This is consistent with Fig. 1, supporting late `[CLS]` attention as an effective teacher signal while allowing the student to learn its own retrieval embedding.

5.2 Effect of Distilled Block Position

This experiment examines where spatial aggregation supervision is most effective across transformer depth. We vary the distilled block range and compare early, middle, and late-block distillation. As shown in Fig. 4, distilling early blocks provides little benefit and can even degrade performance, while distilling late blocks consistently improves performance. This supports our observation that early `[CLS]` attention is less semantically aligned with Re-ID, whereas late-block attention better captures task-relevant spatial aggregation. Among the tested settings, Blocks 7–12 give the best overall performance and are adopted as our default. The full numerical results, including both mAP and Rank-1 accuracy, are provided in Appendix A.4.

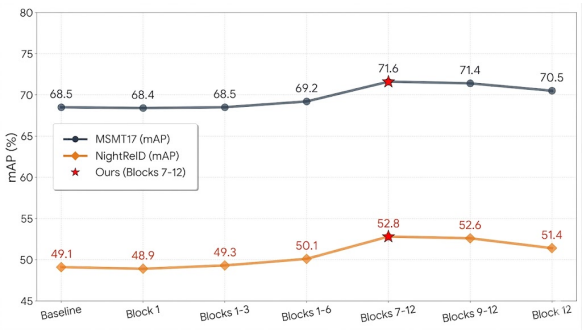

Figure 4: **Effect of distilled block position** on TransReID-SSL for ViT-B → ViT-S. The results show mAP on MSMT17 and NightReID.

Table 3: **Attention-region ablation** for spatial aggregation transfer on TransReID-SSL. Performance is reported as ViT-B / ViT-S.

Distillation Target	MSMT17		NightReID	
	mAP	R-1	mAP	R-1
Baseline (ViT-B/ViT-S)	73.2 / 68.5	88.7 / 86.2	51.9 / 49.1	66.5 / 64.6
Full attention matrix	74.0 / 71.3	88.8 / 87.8	53.0 / 52.2	68.0 / 67.5
`[CLS]` query only (Ours)	74.0 / **71.6**	**89.1 / 88.1**	**53.3 / 52.8**	**68.1 / 68.4**

5.3 `[CLS]` Attention vs. Full Attention Distillation

This experiment examines whether the benefit comes from transferring global-token spatial aggregation, rather than from generic attention matching. We compare our `[CLS]`-query target with full attention matrix distillation. While the `[CLS]`-query target focuses on how the global representation aggregates spatial evidence, full attention distillation additionally constrains all token-to-token interactions. As shown in Tab. 3, full attention distillation does not improve over the `[CLS]`-query target. Instead, the `[CLS]`-query formulation achieves slightly better overall performance, especially for the ViT-S student. This suggests that the key supervision lies in guiding global-token spatial aggregation, while matching the entire attention matrix may impose unnecessary constraints.

5.4 Multi-Dataset Training and Zero-Shot Generalization

This experiment examines whether spatial aggregation guidance transfers across domains, rather than simply fitting dataset-specific cues through additional distillation training. We jointly train models on DukeMTMC, MSMT17, CUHK03, and NightReID, and evaluate them on both the seen datasets and the unseen Market-1501 benchmark. Multi-dataset training provides a stronger testbed than single-dataset training, as teacher guidance must remain useful across diverse camera systems, illumination conditions, and annotation distributions. The seen-dataset results test whether spatial aggregation transfer remains effective under large-scale heterogeneous training, while the zero-shot Market-1501 result evaluates transfer beyond the training domains. As shown in Tab. 4, spatial aggregation transfer remains effective and improves zero-shot generalization to Market-1501, suggesting that the learned guidance is not merely dataset-specific. Notably, ViT-B → S substantially reduces the ViT-S–ViT-B gap across all datasets, showing that teacher-guided spatial aggregation is especially effective for compact models in deployment-oriented training. This is further supported by the qualitative results on unseen Market-1501 in Appendix A.7, where the distilled ViT-S exhibits smaller spatial aggregation discrepancies from the ViT-B teacher despite Market-1501 being excluded from training.

5.5 Effect of Loss Function

This experiment examines how the loss function affects spatial aggregation supervision. As shown in Tab. 5, mean squared error (MSE) performs poorly, suggesting that `[CLS]` attention maps should be

Table 4: **Multi-dataset training and zero-shot evaluation** on TransReID-SSL. Models are jointly trained on DukeMTMC, MSMT17, CUHK03, and NightReID, and evaluated on the seen datasets and the unseen Market-1501 benchmark.

Method	Seen Datasets								Unseen	
	DukeMTMC		MSMT17		CUHK03		NightReID		Market-1501	
	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1
Baseline (ViT-B)	82.4	91.1	71.2	87.7	81.8	83.4	50.3	64.3	73.2	90.4
Baseline (ViT-S)	79.5	88.8	65.9	84.5	78.6	80.6	46.4	61.0	65.5	85.3
ViT-B $\rightarrow$ B	**82.6**	90.7	**71.5**	**87.9**	**82.5**	**84.2**	**50.9**	**65.6**	**73.9**	90.4
ViT-B $\rightarrow$ S	**81.4**	**90.0**	**69.5**	**86.9**	**80.3**	**81.9**	**50.1**	**64.8**	**70.4**	**88.2**

Table 5: **Spatial aggregation loss ablation** for ViT-B $\rightarrow$ S on TransReID-SSL.

Loss	MSMT17		NightReID	
	mAP	R-1	mAP	R-1
Baseline (ViT-S)	68.5	86.2	49.1	64.6
MSE	68.2	85.8	49.0	64.0
CE	71.3	88.0	52.4	67.6
Reverse KL	71.3	87.6	52.3	**68.4**
Ours	**71.6**	**88.1**	**52.8**	**68.4**

Table 6: **`[CLS]` attention target construction ablation** for ViT-B $\rightarrow$ B on TransReID-SSL.

Method	Remove [CLS] Key	Mean Heads	MSMT17		NightReID	
			mAP	R-1	mAP	R-1
Baseline (ViT-B)	–	–	73.2	88.7	51.9	66.5
Distill.	✗	✗	74.0	88.9	53.2	67.7
	✓	✗	73.9	**89.1**	53.2	67.4
	✗	✓	**74.0**	**89.1**	**53.3**	**68.1**
	✓	✓	**74.0**	**89.1**	**53.3**	**68.1**

treated as probability distributions rather than ordinary continuous features. Cross-entropy, which is commonly used for attention transfer (Li et al., 2024), performs much better by emphasizing teacher-attended regions. Further combining reverse KL divergence gives the best overall performance, supporting the role of both emphasizing teacher-attended regions and suppressing teacher-inconsistent regions in spatial aggregation transfer.

5.6 Robustness to Attention Target Construction

This experiment examines whether the performance gain depends on minor target-construction choices. We conduct this ablation in the ViT-B $\rightarrow$ ViT-B setting, where head-wise and head-averaged variants can both be compared. We vary two implementation details: removing the `[CLS]` key entry and averaging attention heads. Head averaging is used for simplicity and cross-architecture compatibility, since ViT-B and ViT-S have different numbers of attention heads. The `[CLS]` key entry is removed so that the target focuses on how spatial patch tokens contribute to the global representation, rather than on self-attention to the `[CLS]` token. As shown in Tab. 6, all variants achieve comparable performance. This indicates that the improvement does not come from fragile implementation details, but from the core idea of transferring late-block `[CLS]` spatial aggregation.

6 Conclusion

In this work, we revisit knowledge distillation for ViT-based Re-ID from the perspective of open-set retrieval. Our analysis shows that large changes in individual descriptor values do not necessarily imply comparable degradation of retrieval structure, motivating the transfer of how descriptors are formed rather than their exact values. We instantiate this perspective using late-block `[CLS]`-to-patch attention as an observable proxy for spatial aggregation. Across multiple ViT-based Re-ID frameworks, spatial aggregation transfer provides effective supervision for both same-capacity refinement and lightweight student distillation without additional inference-time components. Preliminary ViT-L $\rightarrow$ ViT-B results in Appendix A.6 provide evidence that spatial aggregation can also be transferred without direct layer correspondence, while developing a robust general alignment strategy for heterogeneous-depth architectures remains future work.

AI USE STATEMENT

In this work, we used generative AI tools to assist with grammar correction, language polishing, and improving the clarity and presentation of figure-related text and visual references. The tools were not used to generate experimental results, conduct analyses, or determine the scientific conclusions of the paper. All AI-assisted content was reviewed and revised by the authors, who take full responsibility for the final manuscript.

ETHICS STATEMENT

Pedestrian Re-ID can support applications such as person search and camera-network analysis, but it also raises concerns regarding privacy, surveillance, consent, and fairness. More accurate and efficient Re-ID models may increase the scalability of persistent tracking, particularly when deployed without appropriate authorization or oversight. In addition, existing Re-ID benchmarks do not provide sufficient demographic annotations to establish fairness across population groups, and models may inherit biases related to people, cameras, or environments represented in the training data. Practical deployment should therefore consider data provenance and consent, appropriate access control and privacy protection, demographic and cross-camera auditing where possible, and human oversight in consequential applications.

REPRODUCIBILITY STATEMENT

We provide the model formulation, training objective, attention-target construction, and distillation procedure in the main paper. Experimental settings, implementation details, dataset protocols, and additional numerical results are provided in the appendix. All experiments are based on standard ViT architectures and established public Re-ID benchmarks, and we report the relevant training and evaluation settings to facilitate independent reproduction of the results.

REFERENCES

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9630–9640, 2021. doi: 10.1109/ICCV48922.2021.00951.

Xinlei Chen, Saining Xie, and Kaiming He. An Empirical Study of Training Self-Supervised Vision Transformers . In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9620–9629, Los Alamitos, CA, USA, October 2021. IEEE Computer Society. doi: 10.1109/ICCV48922.2021.00950. URL https://doi.ieeecomputersociety.org/10.1109/ICCV48922.2021.00950.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

Dengpan Fu, Dongdong Chen, Jianmin Bao, Hao Yang, Lu Yuan, Lei Zhang, Houqiang Li, and Dong Chen. Unsupervised pre-training for person re-identification. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2021.

Jianping Gou, Baosheng Yu, Stephen John Maybank, and Dacheng Tao. Knowledge distillation: A survey. *CoRR*, abs/2006.05525, 2020. URL https://arxiv.org/abs/2006.05525.

Mengran Gou, Srikrishna Karanam, Wenqian Liu, Octavia Camps, and Richard J. Radke. Dukemtmc4reid: A large-scale multi-camera person re-identification dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1425–1434, 2017. doi: 10.1109/CVPRW.2017.185.

Zhiwei Hao, Jianyuan Guo, Ding Jia, Kai Han, Yehui Tang, Chao Zhang, Han Hu, and Yunhe Wang. Learning efficient vision transformers via fine-grained manifold distillation. In *Advances in Neural Information Processing Systems*, 2022.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16000–16009, June 2022.

Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15013–15022, October 2021.

Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In *International Conference on Computer Vision (ICCV)*, 2019.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

Bin Hu, Xinggang Wang, and Wenyu Liu. Personvit: Large-scale self-supervised vision transformer for person re-identification. *Machine Vision and Applications*, 36(2):32, 2025. doi: 10.1007/s00138-025-01659-y. URL https://doi.org/10.1007/s00138-025-01659-y.

Harold Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 186(1007):453–461, 1946. ISSN 00804630. doi: 10.2307/97883. URL http://dx.doi.org/10.2307/97883.

Samy Jelassi, Michael Sander, and Yuanzhi Li. Vision transformers provably learn spatial structure. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 37822–37836. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/f69707de866eb0805683d3521756b73f-Paper-Conference.pdf.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMlR, 2019.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, March 1951. doi: 10.1214/aoms/1177729694. URL https://doi.org/10.1214%2Faoms%2F1177729694.

Alexander Cong Li, Yuandong Tian, Beidi Chen, Deepak Pathak, and Xinlei Chen. On the surprising effectiveness of attention transfer for vision transformers. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 152–159, 2014. doi: 10.1109/CVPR.2014.27.

J. Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991. doi: 10.1109/18.61115.

Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

Hao Luo, Pichao Wang, Yi Xu, Feng Ding, Yanxin Zhou, Fan Wang, Hao Li, and Rong Jin. Self-supervised pre-training for transformer-based person re-identification, 2021. URL https://arxiv.org/abs/2111.12084.

Hao Ni, Yuke Li, Lianli Gao, Heng Tao Shen, and Jingkuan Song. Part-aware transformer for generalizable person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11280–11289, 2023.

Namuk Park and Songkuk Kim. How do vision transformers work? In *International Conference on Learning Representations*, 2022.

Namuk Park, Wonjae Kim, Byeongho Heo, Taekyung Kim, and Sangdoo Yun. What do self-supervised vision transformers learn? In *International Conference on Learning Representations*, 2023.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3962–3971, 2019. doi: 10.1109/CVPR.2019.00409.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 12116–12128. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/652cf38361a209088302ba2b8b7f51e0-Paper.pdf.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6550.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *International Conference on Learning Representations*, 2020.

Fred Tung and Greg Mori. Similarity-preserving knowledge distillation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1365–1374, 2019. doi: 10.1109/ICCV.2019.00145.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 79–88, 2018. doi: 10.1109/CVPR.2018.00016.

Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi. Deep Learning for Person Re-Identification: A Survey and Outlook . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 44(06):2872–2893, June 2022. ISSN 1939-3539. doi: 10.1109/TPAMI.2021.3054775. URL https://doi.ieeecomputersociety.org/10.1109/TPAMI.2021.3054775.

Mang Ye, Shijian Chen, Chao Li, et al. Transformer for object re-identification: A survey. *International Journal of Computer Vision*, 133:2410–2440, 2025. doi: 10.1007/s11263-024-02284-4. URL https://doi.org/10.1007/s11263-024-02284-4.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Sks9_ajex.

Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11943–11952, 2022. doi: 10.1109/CVPR52688.2022.01165.

Yuxuan Zhao, Weijian Ruan, He Li, and Mang Ye. Nightreid: A large-scale nighttime person re-identification benchmark. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 10519–10527, 2025. doi: 10.1609/aaai.v39i10.33142. URL https://ojs.aaai.org/index.php/AAAI/article/view/33142.

Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1116–1124, 2015. doi: 10.1109/ICCV.2015.133.

Liang Zheng, Yi Yang, and Alexander Hauptmann. Person re-identification: Past, present and future. *ArXiv*, abs/1610.02984, 2016. URL https://api.semanticscholar.org/CorpusID:13599403.

A Appendix

A.1 [CLS] Token Intervention at Inference

Table 7: **Effect of [CLS] token interventions across transformer blocks** on Market-1501. Feature and attention shifts are measured between the original and intervened outputs for the same image, using cosine distance for descriptors and Jensen–Shannon divergence for [CLS]-to-patch attention maps. Arrows indicate whether lower (↓) or higher (↑) values are preferable.

Injection Vector	Block	Performance		Feature Geometry		Shift from Original	
		R-1 ↑	mAP ↑	Intra-ID ↓	Inter-ID ↑	Feature ↓	Attn. JS ↓
No intervention	–	96.3	92.7	0.206 ± 0.122	0.993 ± 0.140	–	–
Initial [CLS] token	2	32.7	10.1	0.196 ± 0.201	0.335 ± 0.264	0.909 ± 0.070	0.0409 ± 0.0147
Initial [CLS] token	6	64.5	29.3	0.174 ± 0.076	0.338 ± 0.097	0.753 ± 0.082	0.0372 ± 0.0142
Initial [CLS] token	12	**90.6**	**73.2**	0.005 ± 0.004	0.023 ± 0.009	0.931 ± 0.045	$\mathbf{0.0082 \pm 0.0030}$
Zero vector	2	27.1	8.1	0.284 ± 0.278	0.435 ± 0.302	0.915 ± 0.073	0.0382 ± 0.0151
Zero vector	6	53.4	21.1	0.175 ± 0.074	0.308 ± 0.090	0.805 ± 0.078	0.0403 ± 0.0143
Zero vector	12	**88.6**	**66.0**	0.046 ± 0.033	0.188 ± 0.073	0.831 ± 0.064	$\mathbf{0.0019 \pm 0.0011}$
Random vector	2	22.8	6.5	0.311 ± 0.321	0.460 ± 0.340	0.927 ± 0.070	0.0387 ± 0.0143
Random vector	6	48.8	18.3	0.166 ± 0.073	0.292 ± 0.092	0.812 ± 0.076	0.0411 ± 0.0134
Random vector	12	**89.0**	**66.9**	0.039 ± 0.028	0.149 ± 0.049	0.834 ± 0.060	$\mathbf{0.0033 \pm 0.0015}$

Table 7 provides additional quantitative results for the [CLS] token intervention study in Section 1. The model is a TransReID ViT-B trained and evaluated on Market-1501, and all interventions are applied only at inference time while keeping the model weights fixed. We compare three image-independent replacements of the [CLS] token: the learned initial [CLS] token, a zero vector, and a fixed Gaussian random vector.

All interventions substantially change the final descriptor relative to the original model, with feature shifts comparable to the original inter-ID distance. However, their retrieval impact varies substantially and more closely follows the change in the [CLS] attention pattern. Across the three intervention types, interventions at early and intermediate blocks induce larger attention shifts and cause severe retrieval degradation, whereas interventions at the final block preserve the attention distribution relatively well and retain much higher retrieval accuracy. Notably, this trend holds even when feature shifts remain similarly large, showing that displacement of individual descriptors alone is insufficient to characterize retrieval degradation. We next examine the corresponding inter-image retrieval geometry more directly.

Table 8: **Retrieval-geometry preservation under [CLS] token interventions** on Market-1501. Q–G Dist. Corr. denotes the Pearson correlation between the original and intervened query–gallery distance matrices; Rank Corr. denotes the mean per-query Spearman correlation between gallery rankings; and NN Overlap@10 measures the overlap between the original and intervened top-10 retrieval neighborhoods. Higher values indicate better preservation of the original retrieval structure.

Injection Vector	Block	mAP ↑	Feature Dist. ↓	Retrieval-Geometry Preservation		
				Q–G Dist. Corr. ↑	Rank Corr. ↑	NN Overlap@10 ↑
Initial [CLS] token	2	10.1	0.909	0.0447	0.1352	0.1350
Initial [CLS] token	12	**73.2**	0.931	**0.4210**	**0.5002**	**0.6392**
Zero vector	2	8.1	0.915	0.0426	0.1172	0.1075
Zero vector	12	**66.0**	0.831	**0.3438**	**0.4398**	**0.5929**

Table 8 further shows that the same trend holds for retrieval-relevant inter-image structure: late interventions preserve substantially more query–gallery distance structure, ranking correlation, and nearest-neighbor overlap despite similarly large changes in individual descriptor values. This does not imply that embedding geometry is secondary to attention; retrieval ultimately depends on the relative geometry of the learned descriptors. Rather, our results suggest that preserving the teacher's

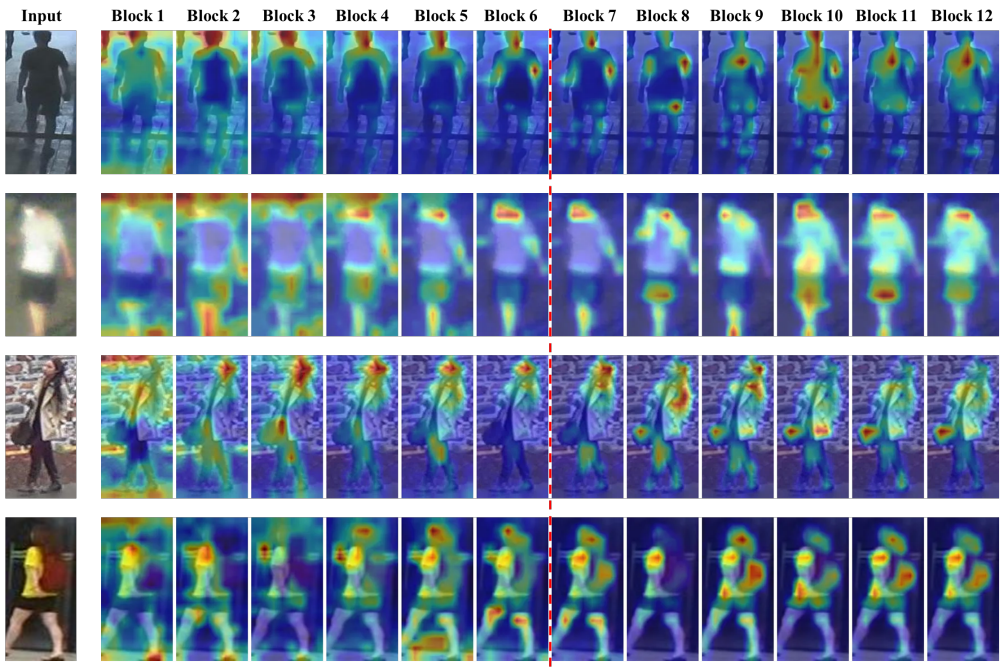

Figure 5: **Layer-wise `[CLS]`-to-patch attention maps** from a TransReID-SSL ViT-B model trained jointly on five Re-ID datasets. Each row shows one input image, followed by attention overlays from Blocks 1–12. The red dashed line separates the early and late transformer blocks used in our default distillation setting.

exact descriptor values is not necessary for retaining this structure, while `[CLS]` attention provides an observable signal associated with how the retrieval-relevant representation is formed. The same trend holds for nearest-neighbor overlap with $k \in 1, 5, 20$.

A.2 `[CLS]` Token Attention Map Visualization

Figure 5 provides qualitative examples of how `[CLS]` attention evolves across transformer depth. The examples include images from different Re-ID benchmarks, indicating that this trend is not specific to a single dataset or visual condition. Early blocks often produce diffuse or background-influenced attention, while later blocks show more coherent attention over person-centric regions such as the torso, clothing, head, legs, and carried items. This supports our choice of applying attention distillation to the later transformer blocks, where the `[CLS]` token captures more task-relevant spatial aggregation patterns for person re-identification.

A.3 Implementation Details and Compute

Tab. 9 lists the hyperparameters used for training and evaluation. Baseline and distilled models are trained with the same data pipeline, optimizer, batch size, and number of epochs. During distillation, the teacher model is kept frozen and only the student model is optimized. The attention distillation loss is added to the standard Re-ID objective with unit weight.

Tab. 10 reports the parameter count and memory footprint of the ViT-B teacher and ViT-S student measured from the inference model. Tab. 11 reports representative observed throughput from training logs rather than a dedicated hardware benchmark. Distillation reduces training throughput because it requires an additional teacher forward pass and attention-loss computation. This overhead is only incurred during training; at inference time, the student architecture is unchanged and no teacher model or additional module is used.

Table 9: **Hyperparameter settings** used in our Re-ID experiments.

Component	Setting
Backbones	ViT-B / ViT-S with patch size 16×16
Input resolution	256×128
Optimizer	SGD
Training epochs	120
Batch size	64
Learning rate	8×10^{-4} for TransReID; 4×10^{-4} for TransReID-SSL and PersonViT
Warmup	20 epochs with cosine warmup
Weight decay	1×10^{-4}
Instances per identity	4
Triplet loss margin	0.3
Label smoothing	Off
Data augmentation	Random horizontal flip, random erasing, and padding
Augmentation probabilities	Flip probability 0.5, random erasing probability 0.5
Padding size	10
Image normalization	Mean $(0.5, 0.5, 0.5)$, std $(0.5, 0.5, 0.5)$
Random seed	1234 for default runs
Attention loss weight	1.0
Re-ranking	Not used

Table 10: **Model complexity** of the ViT teacher and student architectures used in our experiments. Parameter counts are measured from the inference model.

Model	Depth	Embedding Dim.	Heads	Parameters	Param. Memory
ViT-B	12	768	12	90.3M	344.62 MB
ViT-S	12	384	6	23.7M	90.35 MB

Table 11: **Representative observed training throughput** on a single NVIDIA RTX A5000. Throughput is rounded and reported in samples per second to account for different dataset sizes.

Model	Setting	Observed Throughput
ViT-B	Baseline training	$\sim$330 samples/s
ViT-B	Distillation training	$\sim$250 samples/s
ViT-S	Baseline training	$\sim$530 samples/s
ViT-S	Distillation training	$\sim$340 samples/s

A.4 Effect of Distilled Block Position

Table 12: **Full numerical results for the distilled block-position ablation** on TransReID-SSL. All results are reported for ViT-B → ViT-S.

Distilled Blocks	MSMT17		NightReID	
	mAP	R-1	mAP	R-1
Baseline	68.5	86.2	49.1	64.6
Block 1	68.4	85.8	48.9	63.8
Blocks 1–3	68.5	86.2	49.3	64.3
Blocks 1–6	69.2	86.9	50.1	64.8
Blocks 7–12 (Ours)	**71.6**	**88.1**	**52.8**	68.4
Blocks 9–12	71.4	87.7	52.6	**69.3**
Block 12	70.5	87.2	51.4	66.2

Tab. 12 provides the full numerical results for the block-position ablation shown in the main paper. Distilling early blocks provides little or no improvement over the baseline, whereas supervising later blocks consistently improves performance on both MSMT17 and NightReID. This empirical result is consistent with the attention-map visualization in Fig. 5, where later blocks show more coherent person-centric attention than earlier blocks. The strongest overall results are obtained when distilling a contiguous range of late blocks, supporting our default choice of Blocks 7–12.

A.5 Random Seed Stability

Table 13: **Random-seed stability** on TransReID-SSL. ViT-B → ViT-S distillation methods are reported as mean ± standard deviation over four seeds. We additionally report ViT-B → B stability for our method. The independently trained ViT-B and ViT-S models are shown as references.

Method	MSMT17		NightReID	
	mAP	R-1	mAP	R-1
ViT-B (baseline)	73.2	88.7	51.9	66.5
ViT-S (baseline)	68.5	86.2	49.1	64.6
Feature Distillation (B→S)	68.10 ± 0.14	85.93 ± 0.28	49.23 ± 0.17	64.35 ± 0.53
Logit Distillation (B→S)	70.58 ± 0.13	87.38 ± 0.22	51.20 ± 0.27	66.90 ± 0.26
Ours (B→S)	$\mathbf{71.53 \pm 0.05}$	$\mathbf{87.93 \pm 0.13}$	$\mathbf{52.80 \pm 0.16}$	$\mathbf{68.35 \pm 0.37}$
Ours (B→B)	$\mathbf{73.95 \pm 0.13}$	$\mathbf{88.98 \pm 0.15}$	$\mathbf{53.45 \pm 0.13}$	$\mathbf{68.10 \pm 0.29}$

We evaluate random-seed stability on TransReID-SSL for both lightweight student distillation and same-capacity refinement. In the ViT-B → ViT-S setting, spatial aggregation transfer achieves the best mean performance among the compared teacher signals with low variance. The ViT-B → B results are also consistent across four seeds, indicating that the same-capacity gains are consistent across runs.

A.6 Preliminary Heterogeneous-Depth Transfer

Table 14: **Preliminary heterogeneous-depth transfer** on NightReID under the standard TransReID framework. ViT-L and ViT-B contain 24 and 12 transformer blocks, respectively.

Setting	mAP	R-1
ViT-L (teacher)	48.7	63.2
ViT-B (baseline)	40.8	57.7
ViT-L → ViT-B	**42.6**	**58.3**

To examine heterogeneous-depth transfer, we evaluate ViT-L (24 blocks) $\rightarrow$ ViT-B (12 blocks) on NightReID under standard TransReID, where a compatible pretrained ViT-L checkpoint is available. Rather than enforcing block-wise correspondence, we obtain a single teacher spatial aggregation map through attention rollout and transfer it to the student. As shown in Tab. 14, this simple cross-depth adaptation improves ViT-B from 40.8/57.7 to 42.6/58.3 mAP/R-1, showing that spatial aggregation transfer remains applicable without matched transformer depths. More adaptive cross-depth aggregation strategies may further improve this extension.

A.7 Qualitative Results

We provide qualitative results using a single model trained jointly on DukeMTMC, MSMT17, CUHK03, and NightReID. We visualize retrieval and attention results on both the seen training datasets and the unseen Market-1501 benchmark. This setting allows us to examine whether the same learned spatial aggregation guidance is effective within the training domains and also transfers to an unseen domain.

A.8 Rank-10 Visualization

As illustrated in Fig. 6, spatial aggregation transfer qualitatively improves retrieval quality for both ViT-B $\rightarrow$ B and ViT-B $\rightarrow$ S. Both distilled models retrieve more correct matches than their corresponding baselines. Notably, this improvement is observed not only on the seen training datasets, but also on the unseen Market-1501 benchmark.

A.9 Attention Map Visualization

Fig. 7 visualizes late-block attention maps for the ViT-B teacher, ViT-S baseline, and distilled ViT-S student, along with the corresponding attention-difference maps relative to the teacher. After spatial aggregation transfer, the ViT-S attention maps become more closely aligned with the teacher, and the difference maps become less pronounced. This effect is observed not only on validation images from seen training domains, but also on images from the unseen Market-1501 benchmark. These results suggest that spatial aggregation transfer guides the student toward teacher-consistent attention behavior, rather than merely fitting dataset-specific visual patterns.



(a) Distilled ViT-B student vs. ViT-B baseline.

(b) Distilled ViT-S student vs. ViT-S baseline.

Figure 6: **Rank-10 retrieval visualization** using the same jointly trained model evaluated on both seen datasets and the zero-shot Market-1501 benchmark. For each dataset, the **top row shows the distilled student** and the **bottom row shows the corresponding baseline**. Green and red boxes denote correct and incorrect retrievals, respectively.



(a) DukeMTMC (b) MSMT17

(c) Market-1501 (d) Market-1501

Figure 7: **Attention map visualizations and attention-difference maps.** In each panel, the top row shows ViT-B teacher attention maps. The middle group shows ViT-S baseline attention maps and their absolute difference maps relative to the teacher. The bottom group shows distilled ViT-S attention maps and their absolute difference maps relative to the teacher. Brighter regions indicate larger attention discrepancies. Panels (a) and (b) are from validation images of seen training domains, while (c) and (d) are from the unseen Market-1501 domain.