TRUSTREID: LEARNING WHICH IDENTITY EVIDENCE TO TRUST FOR UNIVERSAL PERSON RE-IDENTIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Person re-identification (ReID) requires matching identities across changes in viewpoint, clothing, sensing modality, and visibility. In unlabeled videos, strict association criteria fragment identities across views, whereas permissive criteria introduce incorrect identity merges, compromising supervision for pre-training. During matching, visual cues that distinguish identities in one setting may mislead retrieval in another, making fixed evidence fusion inadequate for a shared model. We introduce TrustReID, a two-stage framework that addresses these challenges through view-conditioned identity pre-training and counterfactual evidence routing. In Stage I, View-Conditioned Progressive Identity Discovery (VPID) preserves view-specific prototypes and progressively links observations within and across videos using relation-specific criteria. This process assigns high-purity pseudo-identity labels to over six million person images, supporting transferable identity learning with identity-supervised losses. Under matched data and computational budgets, this identity-supervised objective outperforms contrastive pre-training. In Stage II, Counterfactual Evidence Routing (CER) complements an always-on pretrained identity branch with four experts encoding color, texture, body structure, and visible parts. A shared, task-agnostic router weights these experts for each query–gallery pair. In addition to the final retrieval loss, we supervise expert weights using the change in the positive–negative ranking margin when each expert is masked, guiding the model to select evidence that improves identity ranking. Without downstream fine-tuning, the frozen pretrained model achieves Rank-1/mAP of 77.28%/51.18% on MSMT17, 70.58%/63.84% on AG-ReID v1 under ground-to-aerial retrieval, and 97.70%/96.25% on Occluded-ReID. After joint training, TRUSTREID outperforms several existing unified ReID models on multiple benchmarks, while supporting six ReID settings with a single checkpoint. We will publicly release the pre-training dataset and model weights to facilitate reproducibility and future research.

1 INTRODUCTION

Person re-identification (ReID) encompasses diverse tasks, including standard, clothes-changing, visible–infrared, and occluded retrieval that share the goal of matching person identities. Recent studies have explored unifying these tasks within a single model. Instruct-ReID (He et al., 2024) uses image or language instructions to specify retrieval requirements, AIO (Li et al., 2024) maps heterogeneous modalities into a shared embedding space, and ReID5o (Zuo et al., 2026) activates modality-specific experts according to the input modality. However, a shared retrieval objective does not imply shared requirements for discriminative evidence. Clothing color and texture are useful for standard ReID, yet reliance on them can hinder matching across clothing changes. Similarly, color differences that distinguish identities in visible images do not provide an equivalent basis for visible–infrared matching. These conflicting requirements make evidence selection a central challenge for unified ReID: the same model must exploit a cue where it is informative while limiting its influence where it is misleading.

These conflicting requirements raise a central question: how should a shared model represent complementary identity evidence and determine which evidence to trust during matching? Clothing

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.

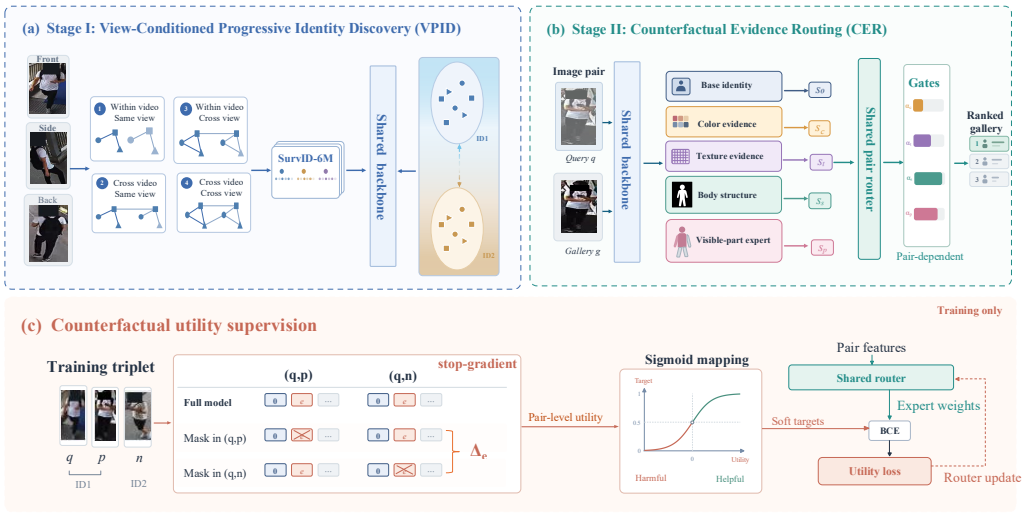


Figure 1: Overview of TRUSTREID. (a) View-conditioned progressive identity discovery constructs pseudo-identities for identity-supervised pre-training. (b) An always-on identity branch and four specialized experts provide complementary matching evidence, fused through pair-dependent routing. (c) During training, expert masking measures utility through ranking-margin changes, providing detached supervision targets for the router.

appearance provides discriminative cues when outfits remain consistent, while clothing changes and modality differences increase the importance of structural cues. Occlusion further requires matching corresponding body regions that are visible in both images. These complementary roles motivate modeling distinct sources of evidence separately, allowing their contributions to be adjusted during matching. However, task instructions and modality information alone cannot determine the appropriate combination for each query–gallery pair: even within the same task, clothing similarity may support a correct match or favor a similarly dressed distractor. We therefore learn to combine evidence based on its contribution to retrieval ranking, emphasizing cues that help distinguish the correct identity from competing candidates. Large vision-language models also support interactive person retrieval, but placing a large gallery within a single context encounters token-budget constraints; ChatReID, for example, first filters candidates using a separate retrieval model (Niu et al., 2025). Our framework implements evidence selection through reusable visual representations and a dedicated matching module, without requiring autoregressive generation for candidate comparison.

Learning complementary evidence for identity matching benefits from transferable person representations. PersonViT (Hu et al., 2025) explores large-scale self-supervised pre-training, while SapiensID (Kim et al., 2025) develops a unified model for face and body recognition. Among methods that discover identity correspondences in unlabeled videos, ISR (Dou et al., 2023) constructs inter-frame positive pairs for reliability-guided contrastive learning, and CION (Zuo et al., 2024) discovers cross-video identities to guide self-distillation. These approaches use recovered identity relations to organize contrastive or distillation-based learning. We investigate whether reliable pseudo-identities can serve directly as supervision for more effective identity-oriented pre-training. Constructing such supervision is complicated by viewpoint variation: appearance differences can leave observations of the same person disconnected, whereas relaxing association criteria risks merging different identities. We address this ambiguity by retaining view-specific prototypes and applying association criteria tailored to video scope and viewpoint relation. The resulting pseudo-identities provide fixed targets for identity-supervised pre-training. Under matched data and computational budgets, this objective outperforms contrastive pre-training, demonstrating the effectiveness of directly using the recovered identities to learn transferable representations.

To build on this identity representation while accommodating conflicting cue requirements across tasks, we introduce TrustReID, a two-stage framework that combines view-conditioned identity pre-training with counterfactual evidence routing. In Stage I, View-Conditioned Progressive Iden-

tity Discovery (VPID) constructs a corpus of over six million person images with pseudo-identity labels for identity-supervised pre-training. In Stage II, Counterfactual Evidence Routing (CER) augments an always-on pretrained identity branch with four complementary experts. Each expert is trained with dedicated constraints that encourage specialization in its intended evidence source. Color and texture experts capture distinct aspects of clothing appearance, a body-structure expert provides structural evidence, and a visible-part expert supports comparison between corresponding visible regions. A shared router weights their outputs for each query–gallery pair without receiving task labels. During training, each expert is masked in turn, and the resulting change in the positive–negative ranking margin provides an explicit utility target for routing. Together, expert-specific constraints and counterfactual supervision guide what evidence each expert encodes and when that evidence should be trusted. At inference, the router predicts the weights directly, without repeated masking.

Our main contributions are summarized as follows:

- We propose evidence-specialized experts with counterfactual routing to address conflicting cue requirements across ReID tasks. Expert-specific constraints guide evidence extraction, while ranking-margin changes under expert masking directly supervise pair-dependent evidence fusion.
- We develop view-conditioned progressive identity discovery with view-specific prototypes and relation-specific association criteria. Under matched data and computation, pre-training with the resulting pseudo-identity supervision outperforms contrastive pre-training.
- We evaluate TRUSTREID across six ReID settings, covering frozen-model direct transfer and single-checkpoint joint-training performance. We will release the pre-training dataset of over six million images and model weights.

2 RELATED WORK

2.1 PERSON RE-IDENTIFICATION PRE-TRAINING

Large-scale pre-training has become an important direction for learning transferable person representations. PersonViT (Hu et al., 2025) combines masked image modeling with discriminative contrastive learning to capture global and local features from unlabeled person images. SapiensID (Kim et al., 2025) explores unified face and body recognition through scale-adaptive tokenization and semantic feature aggregation. These studies demonstrate different approaches to learning person representations from large-scale data. Another line of work explicitly exploits identity correspondences in unlabeled videos. ISR (Dou et al., 2023) constructs inter-frame positive pairs and uses reliability-guided contrastive learning to mitigate noisy correspondences. CION (Zuo et al., 2024) discovers cross-video identities through progressive denoising and incorporates them into identity-guided self-distillation. Our work focuses on constructing reliable pseudo-identities through view-specific prototypes and relation-specific association criteria. We further examine the effectiveness of using these identities directly as fixed supervision, with controlled comparisons against contrastive pre-training.

2.2 UNIFIED PERSON RE-IDENTIFICATION

Recent studies have explored unified person re-identification across different retrieval requirements and input modalities. Instruct-ReID (He et al., 2024) formulates multiple ReID tasks through image or language instructions and introduces OmniReID to support joint training. AIO (Li et al., 2024) combines modality-specific tokenizers with a shared frozen encoder and cross-modal heads to accommodate heterogeneous inputs. ReID5o (Zuo et al., 2026) supports multiple modality combinations through unified encoding and modality-specific experts activated by input modality. ChatReID (Niu et al., 2025) adapts a vision-language model through hierarchical progressive tuning to enable interactive person retrieval. These approaches organize unified retrieval around instructions, modalities, or interactive queries, whereas the usefulness of an identity cue can vary even among image pairs within the same setting. We explicitly specialize complementary evidence through dedicated training constraints and adapt its fusion according to its utility for each match.

3 METHOD

3.1 OVERVIEW

Given a query image q and a gallery $\mathcal{G} = \{g_j\}_{j=1}^{N}$, we aim to rank gallery images by identity similarity using a shared model that supports diverse visual ReID settings without task labels for routing.

As illustrated in Figure 1, TRUSTREID comprises two stages. Stage I progressively associates observations within and across videos under different viewpoint relations, constructing pseudo-identities for identity-supervised pre-training. Stage II augments the pretrained identity representation with four experts specialized in color, texture, body structure, and visible parts through dedicated supervision. A shared router learns pair-dependent expert weights from counterfactual utility, measured by ranking-margin changes under expert masking. At inference, it predicts these weights directly without counterfactual masking.

The final matching score combines the always-on identity branch with adaptively weighted expert evidence:

$$S(q, g) = s_0(q, g) + \sum_{e \in \mathcal{E}(q,g)} \alpha_e(q, g)\, s_e(q, g), \tag{1}$$

where $s_0(q, g)$ denotes the base identity score and $s_e(q, g)$ the calibrated score of expert e. The expert set $\mathcal{E} = \{c, t, s, p\}$ corresponds to color, texture, body structure, and visible parts, respectively, and $\mathcal{E}(q, g) \subseteq \mathcal{E}$ contains the experts available for the image pair. The router predicts independent weights $\alpha_e(q, g) \in (0, 1)$ without task labels. Gallery images are ranked in descending order of $S(q, g)$.

3.2 VIEW-CONDITIONED IDENTITY PRE-TRAINING

We develop VPID to construct reliable pseudo-identities for identity-supervised pre-training. Its central design is to retain view-specific identity representations and progressively associate observations using criteria tailored to their recording scope and viewpoint relation.

Starting from cleaned, view-annotated tracklets, we maintain a separate prototype for each available viewpoint $r \in \{\text{front}, \text{side}, \text{back}\}$ within an identity cluster C:

$$\mu_C^r = \frac{\sum_{i \in C: r_i = r} q_i \phi_i}{\left\| \sum_{i \in C: r_i = r} q_i \phi_i \right\|_2}, \tag{2}$$

where ϕ_i is a normalized bootstrap descriptor and q_i is its quality weight. This representation preserves distinct appearance statistics across viewpoints throughout association.

We progressively establish identity links through four regimes: within-video same-view, cross-video same-view, within-video cross-view, and cross-video cross-view association. Each regime uses relation-specific similarity thresholds, supported by reciprocal-neighbor agreement, camera-time compatibility, and cluster compactness checks. After each merge, view-specific prototypes are updated for subsequent association. A constraint-checked transitive closure then converts the accepted links into fixed pseudo-identity labels.

The resulting SurvID-6M contains over six million person images and supports pre-training a Swin-B backbone (Liu et al., 2021) with identity supervision. Data processing, association rules, and the pre-training objective are detailed in Appendix A. Data processing, association rules, the pre-training objective, and a pseudo-identity quality audit Table 10 are detailed in Appendix B.

3.3 EVIDENCE-SPECIALIZED EXPERTS

We organize identity evidence into clothing appearance, body structure, and local body-region correspondence. Clothing appearance is decomposed into color and texture, allowing chromatic information and local patterns to be weighted separately. Body structure provides geometric cues, while visible parts support regional correspondence under partial visibility. This motivates four complementary experts, trained with dedicated constraints because identity supervision alone does not determine which cues each branch encodes. Their representations need not be mutually exclusive.

Given an image x, each expert maps shared backbone features $F_\theta(x)$ to an evidence-specific representation used to compute $s_e(q, g)$. The router adjusts these scores' contributions for each query–gallery pair, while the pretrained identity branch remains active, as defined in Eq. (1). We next describe the expert-specific training constraints.Examples of the expert-specific training transformations and supervision targets are provided in Appendix E.

Color and texture experts. We specialize these experts through complementary clothing-region transformations. A pretrained SCHP parser (Li et al., 2020) provides a clothing mask $M(x)$, from which we construct

$$\begin{aligned} x^{\text{gray}} &= (1 - M(x)) \odot x + M(x) \odot \mathcal{G}(x), \\ x^{\text{low}} &= (1 - M(x)) \odot x + M(x) \odot \mathcal{B}(x), \end{aligned} \tag{3}$$

where $\mathcal{G}$ is grayscale conversion with three-channel replication, $\mathcal{B}$ is low-pass filtering, and $\odot$ denotes element-wise multiplication. Grayscale conversion removes chromatic information while retaining luminance patterns; low-pass filtering suppresses fine texture while approximately preserving coarse color.

The color expert uses low-pass variants as positives and grayscale variants as negatives, while the texture expert reverses these assignments. For expert $e \in \{c, t\}$, let $z_e(x)$ denote its representation and $x^+_{i,e}$ and $x^-_{i,e}$ the corresponding positive and negative variants of x_i. We pull positive representations closer and push negative representations apart:

$$\begin{aligned} \mathcal{L}^+_e &= \frac{1}{B} \sum_{i=1}^{B} \left[1 - \text{sim}\big(z_e(x_i), z_e(x^+_{i,e})\big)\right], \\ \mathcal{L}^-_e &= \frac{1}{|\Omega|} \sum_{i \in \Omega} \left[m_e + \text{sim}\big(z_e(x_i), z_e(x^-_{i,e})\big)\right]_+, \end{aligned} \tag{4}$$

where sim is cosine similarity, B is the batch size, m_e is an expert-specific margin, $[u]_+ = \max(0, u)$, and Ω contains samples meeting the clothing-mask coverage threshold. These assignments concern evidence similarity and leave identity labels unchanged.

An auxiliary head additionally predicts regional Lab chromatic histograms from $z_c(x)$, supervised by cross-entropy $\mathcal{L}_{\text{hist}}$. The expert objectives are

$$\begin{aligned} \mathcal{L}_c &= \lambda^+_c \mathcal{L}^+_c + \lambda^-_c \mathcal{L}^-_c + \lambda^h_c \mathcal{L}_{\text{hist}}, \\ \mathcal{L}_t &= \lambda^+_t \mathcal{L}^+_t + \lambda^-_t \mathcal{L}^-_t. \end{aligned} \tag{5}$$

Histogram construction and supervision are detailed in Appendix B.

Body-structure expert. Given an input image x, the body-structure expert E_s maps the shared backbone features to a structure representation:

$$z_s(x) = E_s(F_\theta(x)). \tag{6}$$

To encourage this representation to retain body shape information, we construct a binary person foreground mask $M(x)$ from SCHP parsing results (Li et al., 2020) and apply the same geometric transformations to the image and mask to preserve spatial alignment. During training, an auxiliary decoder D_s predicts the silhouette solely from $z_s(x)$:

$$\widehat{M}(x) = \sigma(\text{reshape}(D_s(z_s(x)))), \tag{7}$$

where σ denotes the sigmoid function and the decoder output is reshaped to the spatial dimensions of $M(x)$. The decoder receives no additional backbone features, ensuring that silhouette reconstruction directly constrains the structure representation. We train this expert using binary cross-entropy and Dice losses, without an additional expert-specific identity loss:

$$\mathcal{L}_s = \mathbb{E}_x \Big[\text{BCE}\big(\widehat{M}(x), M(x)\big) + \lambda_{\text{Dice}} \, \text{DiceLoss}\big(\widehat{M}(x), M(x)\big)\Big], \tag{8}$$

where λ_{Dice} controls the relative weight of the Dice loss.

Visible-part expert. To localize body regions for local matching, we generate part pseudo-labels $Y_p(x)$ using SCHP (Li et al., 2020) pretrained on Pascal-Person-Part. The labels comprise five body parts head, torso, arms, legs and background. The image and labels undergo the same geometric transformations to preserve spatial alignment.

Given an image x, the visible-part expert E_p predicts part probabilities from the shared backbone's spatial features:

$$\widehat{A}(x) = \mathrm{Softmax}(E_p(F_\theta(x))), \tag{9}$$

where softmax is applied along the class dimension, and predictions and labels are aligned to the same spatial resolution. We train the expert with cross-entropy over valid labeled locations:

$$\mathcal{L}_p = -\frac{1}{|\mathcal{V}|} \sum_{(x,u)\in\mathcal{V}} \log \widehat{A}_{u,Y_p(x)_u}(x), \tag{10}$$

where $\mathcal{V}$ denotes the set of valid labeled locations in the current mini-batch and u indexes spatial positions. This objective trains the expert to distinguish visible body regions, providing spatial weights for subsequent aggregation of the corresponding backbone features, without an additional expert-specific identity loss.

3.4 Counterfactual Evidence Routing

Expert-specific constraints guide evidence extraction; CER learns its contribution to each query–gallery match through explicit ranking-utility supervision. During initial router training, the backbone and experts are frozen, and only router parameters ϕ are updated.

Let $s_0(q,g)$ denote the identity score and $s_e(q,g)$ the calibrated similarity of expert $e \in \mathcal{E} = \{c,t,s,p\}$. To construct targets independently of current router predictions, we use fixed reference weights β_e:

$$S_{\mathrm{ref}}(q,g) = s_0(q,g) + \sum_{e\in\mathcal{E}(q,g)} \beta_e s_e(q,g), \tag{11}$$

where $\mathcal{E}(q,g)$ contains available experts, excluding the visible-part expert when no corresponding region is visible in both images.

For a query q, a positive candidate g^+, and a negative candidate g^-, let $m_{\mathrm{ref}} = S_{\mathrm{ref}}(q,g^+) - S_{\mathrm{ref}}(q,g^-)$. Removing the same expert e from both scores, with all other contributions fixed, yields $m_{\mathrm{ref},-e}$ and the utility

$$\begin{aligned} \Delta_e(q,g^+,g^-) &= m_{\mathrm{ref}} - m_{\mathrm{ref},-e} \\ &= \beta_e\big[s_e(q,g^+) - s_e(q,g^-)\big]. \end{aligned} \tag{12}$$

Positive utility indicates an increased reference ranking margin. This additive intervention requires no additional network forward passes.

To supervise pair-dependent routing, we aggregate triplet utilities before constructing soft targets. Let $\mathcal{T}_e(q,g)$ contain triplets in which g is either the positive or negative candidate and expert e is available for both comparisons. Candidates are sampled within the same dataset and retrieval condition, excluding self-matches. We define

$$\begin{aligned} u_e(q,g) &= \frac{1}{|\mathcal{T}_e(q,g)|} \sum_{(q,g^+,g^-)\in\mathcal{T}_e(q,g)} \Delta_e(q,g^+,g^-), \\ t_e(q,g) &= \mathrm{stopgrad}\left[\sigma\left(\frac{u_e(q,g)}{\tau}\right)\right], \end{aligned} \tag{13}$$

where $\tau > 0$ controls target smoothness. Pair–expert combinations without valid triplets are excluded from utility supervision.

The router receives a descriptor $\psi(q,g)$ formed by concatenating image representations and expert similarities, without task labels. It predicts independent weights $\alpha_e(q,g) = \sigma(R_\phi(\psi(q,g))_e)$, which

determine the matching score in Eq. (1). We supervise both expert utility and the resulting retrieval ranking:

$$\begin{aligned}\mathcal{L}_{\text{utility}} &= \frac{1}{|\mathcal{I}|}\sum_{(q,g,e)\in\mathcal{I}} \text{BCE}\big(\alpha_e(q,g), t_e(q,g)\big),\\ \mathcal{L}_{\text{rank}} &= \frac{1}{|\mathcal{T}|}\sum_{(q,g^+,g^-)\in\mathcal{T}} \text{softplus}\big(\mu - S(q,g^+) + S(q,g^-)\big),\end{aligned} \tag{14}$$

where $\mathcal{I}$ contains valid pair–expert targets, $\mathcal{T}$ contains training triplets, and μ is the ranking margin. The overall objective is

$$\mathcal{L}_{\text{router}} = \mathcal{L}_{\text{rank}} + \lambda_{\text{utility}}\mathcal{L}_{\text{utility}}. \tag{15}$$

Identity labels construct training comparisons but are not router inputs. At inference, weights are predicted directly without counterfactual target construction or expert masking. Calibration and hyperparameter settings are provided in Appendix B.

Joint fine-tuning. After router training, we keep the expert parameters frozen and jointly fine-tune the shared backbone, identity branch, and router with a reduced learning rate. We retain the identity-supervised objective for the identity branch and the routing objective $\mathcal{L}_{\text{router}}$. Counterfactual supervision targets remain detached from gradient computation.

4 Experiments

4.1 Experimental Setup

Datasets and evaluation protocols. We evaluate TrustReID across six ReID settings: standard, clothes-changing, occluded, visible–infrared, sketch-based, and aerial–ground retrieval. Stage I is evaluated directly with frozen pretrained representations, while Stage II uses a single jointly trained checkpoint. ENTIReID is held out from all training, hyperparameter tuning, and checkpoint selection. We report Rank-1 accuracy and mAP without re-ranking or test-time augmentation. Detailed datasets and evaluation protocols are provided in Appendix C Table 6.

Implementation details. We pre-train Swin-B on SurvID-6M using 224×224 inputs and select the Stage I checkpoint by validation mAP on a held-out pre-training split. Stage II follows the training procedure in Section 3.4 with balanced multi-dataset sampling. Full optimization and hardware configurations are provided in Appendix B.

4.2 Direct Transfer and Unified Adaptation

Figure 2 shows strong Stage I direct transfer, reaching 95.5% and 97.7% Rank-1 on Market1501 and Occluded-ReID. With one checkpoint shared across six settings, Stage II improves Rank-1 from 33.0% to 80.2% on PRCC and from 11.8% to 93.3% on RegDB, while increasing performance on the fully held-out ENTIReID from 66.7% to 70.0%. Notably, our universal model surpasses the selected reported SOTA references on Occluded-ReID (98.9% vs. 94.9%) and eight of the eleven aerial–ground protocols. These results demonstrate that a single shared checkpoint can achieve broad task coverage while exceeding strong reported results on individual benchmarks.

4.3 Comparison with Unified ReID Models

Table 1 compares TrustReID with existing unified ReID models. FlexiReID is included as a unified-architecture reference, as its cited results do not establish a shared checkpoint across all listed datasets. Using a single jointly trained checkpoint, TrustReID achieves the highest listed Rank-1 on five of the seven benchmarks and the highest mAP on five. Its largest advantages appear in clothes-changing and occluded retrieval: it exceeds the strongest listed baseline by 19.45/6.81 percentage points in Rank-1/mAP on PRCC and by 8.47/14.68 points on Occluded-Duke. It also leads both metrics on MSMT17 and SYSU-MM01, demonstrating competitive performance across standard and cross-modal retrieval. Together with the expert and routing ablations in Table 3, these results support combining transferable identity pre-training with utility-supervised, pair-conditioned

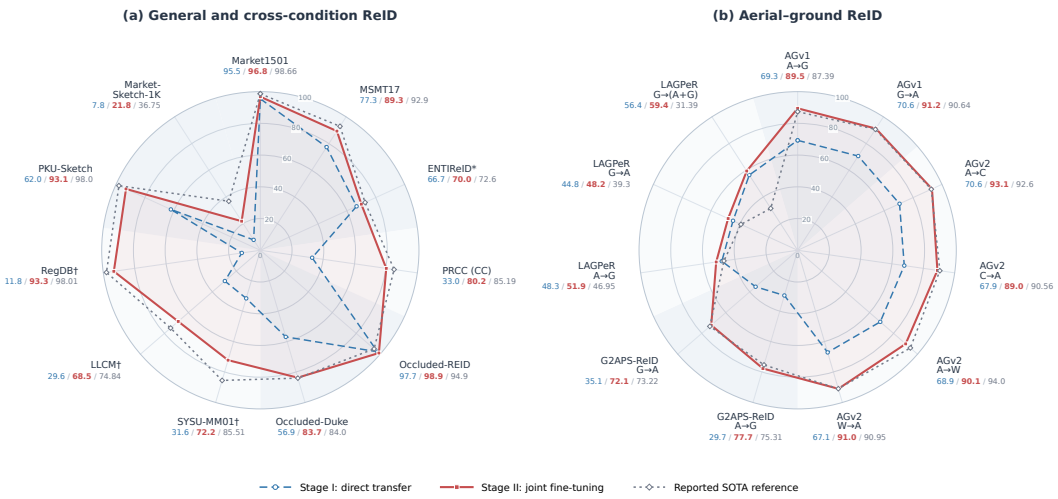


Figure 2: **Rank-1 (%) across six ReID settings and 22 protocols.** (a) Standard and cross-condition; (b) aerial–ground. References may use different protocols and re-ranking. *ENTIReID: held out for ours; supervised reference. †Reference protocols: SYSU-MM01, all-search single-shot IR→VIS; LLCM/RegDB, VIS→IR.

Table 1: Comparison with unified ReID models and a dataset-specific fine-tuning baseline across five retrieval settings. Results are reported as Rank-1 / mAP (%). "–" indicates an unreported result in the cited evaluation. Bold indicates the best reported value for each metric. Dataset protocols are summarized in Table 6.

Dataset / Protocol	Instruct-ReID (MTL) (He et al., 2024)	VersReID (Zheng et al., 2024)	FlexiReID (Sun et al., 2025)	UniGeo (Su et al., 2026)	TRUSTREID
Standard ReID					
Market1501	96.50 / 93.50	96.80 / 93.20	96.00 / 92.10	96.60 / 92.90	**96.79 / 92.28**
MSMT17	86.90 / 72.40	88.80 / 74.20	83.70 / 67.50	87.40 / 71.60	**89.32 / 74.92**
Clothes-changing ReID					
PRCC (CC)	54.20 / 52.30	60.70 / 71.40	–	59.00 / 70.50	**80.15 / 78.21**
Occluded ReID					
Occluded-Duke	–	75.20 / 66.10	–	74.80 / 65.80	**83.67 / 80.78**
Visible–infrared ReID					
SYSU-MM01	–	69.30 / 66.90	67.90 / 62.50	64.30 / 65.40	**72.16 / 72.89**
LLCM	66.70 / **67.50**	–	–	–	**68.45** / 62.91
Sketch-based ReID					
PKU-Sketch	–	–	**93.50** / 91.20	–	93.14 / **91.40**

evidence fusion in a shared model. Additional aerial–ground results are provided in Appendix D, Table 7, extending the evaluation to a sixth setting.

Without downstream training or adaptation, our frozen Stage I model achieves 88.63%, 51.18%, and 96.25% mAP on Market1501, MSMT17, and Occluded-ReID, respectively; complete direct-transfer results across six ReID settings are provided in Appendix D.2.

4.4 Ablation Studies

Scaling Analysis. Figure 3 in the appendix shows that increasing pre-training data from 0.5M to 6.5M images improves direct-transfer Rank-1 and mAP on MSMT17 and ENTIReID, with diminishing gains beyond 4M and marginal improvements from 6M to 6.5M.

Table 2: Effect of pre-training objectives and the VPID strategy on direct transfer. Results are reported as Rank-1 / mAP (%). Best results are highlighted in bold.

Pre-training objective	MSMT17	Market1501	ENTIReID
Contrastive loss	54.01 / 28.94	91.42 / 81.31	49.91 / 49.63
Identity supervision(w/o VPID)	70.15 / 45.52	93.65 / 85.12	61.34 / 60.15
Identity supervision(w/ VPID)	**77.28 / 51.18**	**95.49 / 88.63**	**66.67 / 67.78**

Table 3: Ablation of expert contributions and router objectives. All configurations share the same frozen identity branch and use the same trained experts where applicable. Results are Rank-1 / mAP (%). ✓ indicates an enabled router loss. Arrows indicate changes relative to the identity-only baseline.

Experts	Fusion	$\mathcal{L}_{\text{rank}}$	$\mathcal{L}_{\text{utility}}$	MSMT17	PRCC CC	Occluded-ReID	SYSU-MM01 IR→VIS	PKU-Sketch Sketch→Photo	AG-ReID v1 G→A
None	–	–	–	77.28 / 51.18	33.05 / 35.78	97.70 / 96.25	31.62 / 32.81	62.00 / 65.79	70.58 / 63.84
Color	Fixed	–	–	79.56 / 53.19↑	30.70 / 31.84↓	98.00 / 97.13↑	29.29 / 29.71↓	57.04 / 62.21↓	71.55↑ / 60.12↓
Texture	Fixed	–	–	80.67 / 54.09↑	31.23 / 34.98↓	98.33 / 97.56↑	32.10 / 34.18↑	63.61 / 67.30↑	71.39 / 64.33↑
Body structure	Fixed	–	–	78.42 / 52.08↑	49.70 / 47.29↑	97.13 / 95.05↓	43.19 / 44.08↑	72.10 / 75.28↑	70.00 / 63.69↓
Visible parts	Fixed	–	–	78.06 / 52.77↑	40.30 / 41.62↑	99.07 / 98.41↑	36.09 / 39.46↑	66.03 / 69.43↑	74.12 / 67.55↑
All	Fixed	–	–	83.52 / 60.50	44.91 / 44.26	98.04 / 97.19	39.48 / 40.72	68.10 / 71.39	76.81 / 69.37
All	Adaptive	✓	–	84.38 / 62.71	49.65 / 48.10	98.79 / 97.28	44.92 / 45.62	73.98 / 76.33	79.01 / 71.76
All	Adaptive	✓	✓	**85.73 / 64.21**	**53.01 / 51.39**	**99.07 / 98.45**	**48.45 / 49.00**	**77.18 / 79.46**	**80.42 / 73.33**

Impact of Pre-training Objectives and VPID. Table 2 compares pre-training objectives and evaluates the contribution of VPID. Our contrastive baseline adapts MoCo by constructing positive pairs from distinct images sharing the same pseudo-identity and negative pairs from different pseudo-identities. Positive pairs may come from either the same or different cameras. We compare this baseline with identity supervision using cross-entropy and triplet losses. For the identity-supervised *w/o VPID* variant, we use a unified clustering threshold without viewpoint conditioning and restrict positive-pair sampling to images from the same camera. The full VPID configuration incorporates view-conditioned progressive association to establish identity correspondences across cameras and viewpoints. Both identity-supervised configurations outperform the evaluated contrastive baseline, with VPID achieving the best direct-transfer performance. Compared with the *w/o VPID* variant, VPID improves mAP by 5.66, 3.51, and 7.63 percentage points on MSMT17, Market1501, and ENTIReID, respectively. On the fully held-out ENTIReID benchmark, Rank-1 also increases from 61.34% to 66.67%. These results support the effectiveness of the full VPID configuration for learning transferable identity representations.

Expert Contributions and Routing. Table 3 shows that expert benefits vary across settings: color and texture help MSMT17 but hurt PRCC, while structure and visible parts benefit complementary retrieval conditions. With the identity branch and expert representations fixed, the full CER configuration achieves the highest mAP and best or tied-best Rank-1 among all ablations across all six benchmarks. Utility supervision further improves mAP over rank-only routing by 3.29, 3.38, and 3.13 percentage points on PRCC, SYSU-MM01, and PKU-Sketch, respectively.

5 Conclusion

We presented TrustReID, a two-stage framework that combines transferable identity learning with adaptive evidence selection for universal ReID. VPID constructs pseudo-identities through view-conditioned progressive association, enabling identity-supervised pre-training for direct transfer. Building on these representations, CER uses leave-one-expert-out ranking utility to supervise pair-conditioned evidence fusion, adapting expert contributions to each query–gallery pair. Experiments across six ReID settings demonstrate competitive performance with a single jointly trained checkpoint, while evaluation on the fully held-out ENTIReID benchmark provides evidence of generalization to an unseen domain.

AI USE STATEMENT

Generative AI tools were used to assist with literature searches, language editing, and the organization and formatting of tables. The authors take full responsibility for the accuracy and integrity of the final manuscript, including all references, reported results, and conclusions.

REPRODUCIBILITY STATEMENT

Section 3 describes the model architecture and training objectives. Appendix A details the pseudo-identity construction pipeline, Appendix B provides training and optimization settings, and Appendix C summarizes the evaluation datasets and protocols. Additional experimental results are reported in Appendix D.

REFERENCES

Zhaopeng Dou, Zhongdao Wang, Yali Li, and Shengjin Wang. Identity-seeking self-supervised representation learning for generalizable person re-identification. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15801–15812. IEEE, 2023.

Weizhen He, Yiheng Deng, Shixiang Tang, Qihao Chen, Qingsong Xie, Yizhou Wang, Lei Bai, Feng Zhu, Rui Zhao, Wanli Ouyang, et al. Instruct-reid: A multi-purpose person re-identification task with instructions. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17521–17531. IEEE, 2024.

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.

Minchul Kim, Dingqiang Ye, Yiyang Su, Feng Liu, and Xiaoming Liu. Sapiensid: Foundation for human recognition. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13937–13947. IEEE, 2025.

He Li, Mang Ye, Ming Zhang, and Bo Du. All in one framework for multimodal re-identification in the wild. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17459–17469. IEEE, 2024.

Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self-correction for human parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):3260–3271, 2020.

Kejun Lin, Zhixiang Wang, Zheng Wang, Yinqiang Zheng, and Shin'ichi Satoh. Beyond domain gap: Exploiting subjectivity in sketch-based person retrieval. In *Proceedings of the 31st ACM international conference on multimedia*, pp. 2078–2089, 2023.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9992–10002. Ieee, 2021.

Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. Pose-guided feature alignment for occluded person re-identification. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 542–551. IEEE, 2019.

Dat Tien Nguyen, Hyung Gil Hong, Ki Wan Kim, and Kang Ryoung Park. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors*, 17(3):605, 2017.

Huy Nguyen, Kien Nguyen, Sridha Sridharan, and Clinton Fookes. Aerial-ground person re-id. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 2585–2590. IEEE, 2023.

Huy Nguyen, Kien Nguyen, Sridha Sridharan, and Clinton Fookes. Ag-reid. v2: Bridging aerial and ground views for person re-identification. *IEEE Transactions on Information Forensics and Security*, 19:2896–2908, 2024.

Ke Niu, Haiyang Yu, Mengyang Zhao, Teng Fu, Siyang Yi, Wei Lu, Bin Li, Xuelin Qian, and Xiangyang Xue. Chatreid: Open-ended interactive person retrieval via hierarchical progressive tuning for vision language models. pp. 24245–24254, 2025.

Lu Pang, Yaowei Wang, Yi-Zhe Song, Tiejun Huang, and Yonghong Tian. Cross-domain adversarial feature learning for sketch re-identification. In *Proceedings of the 26th ACM international conference on Multimedia*, pp. 609–617, 2018.

Bohan Su, Jiashuo Wang, Fangyi Liu, and Mang Ye. Reliability-aware 3d geometric injection for universal person re-identification. *arXiv preprint arXiv:2607.18863*, 2026.

Zhen Sun, Lei Tan, Yunhang Shen, Chengmao Cai, Xing Sun, Pingyang Dai, Liujuan Cao, and Rongrong Ji. Flexireid: Adaptive mixture of expert for multi-modal person re-identification. 2025. URL https://arxiv.org/abs/2510.15595.

Shining Wang, Yunlong Wang, Ruiqi Wu, Bingliang Jiao, Wenxuan Wang, and Peng Wang. Secap: Self-calibrating and adaptive prompts for cross-view person re-identification in aerial-ground networks. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22119–22128. IEEE, 2025.

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. IEEE, 2018.

Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *2017 IEEE international conference on computer vision (ICCV)*, pp. 5390–5399. IEEE, 2017.

Qize Yang, Ancong Wu, and Wei-Shi Zheng. Person re-identification by contour sketch under moderate clothing change. *IEEE transactions on pattern analysis and machine intelligence*, 43 (6):2029–2046, 2019.

Serdar Yıldız and Ahmet Nezih Kasım. Entire-id: An extensive and diverse dataset for person re-identification. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 1–5. IEEE, 2024.

Yukang Zhang and Hanzi Wang. Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2153–2162. IEEE, 2023.

Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pp. 1116–1124, 2015.

Wei-Shi Zheng, Junkai Yan, and Yi-Xing Peng. A versatile framework for multi-scene person re-identification. 2024. URL https://arxiv.org/abs/2403.11121.

Jiaxuan Zhuo, Zeyu Chen, Jianhuang Lai, and Guangcong Wang. Occluded person re-identification. In *2018 IEEE international conference on multimedia and expo (ICME)*, pp. 1–6. IEEE, 2018.

Jialong Zuo, Ying Nie, Hanyu Zhou, Huaxin Zhang, Haoyu Wang, Tianyu Guo, Nong Sang, and Changxin Gao. Cross-video identity correlating for person re-identification pre-training. volume 37, pp. 25228–25250, 2024.

Jialong Zuo, Yongtai Deng, Mengdan Tan, Rui Jin, Dongyue Wu, Nong Sang, Liang Pan, and Changxin Gao. Reid5o: Achieving omni multi-modal person re-identification in a single model. volume 38, pp. 53401–53424, 2026.

A Implementation Details of VPID

Each video is associated with a video index v, camera index c, and timestamp. We first construct cleaned, view-annotated tracklets and then progressively associate them into pseudo-identities.

Table 4: View-dependent quality grading of candidate crops. Grade A takes precedence over grade B.

Viewpoint	Grade	Criteria
Front	A	$N_i(\mathcal{K}, 0.8) \geq 15$ and $h_i/w_i \geq 2$.
	B	$\big(N_i(\mathcal{H}, 0.9) \geq 1 \vee N_i(\mathcal{H}, 0.6) \geq 3\big)$ and $N_i(\mathcal{U}, 0.6) \geq 4$.
Back	A	$N_i(\mathcal{H}, 0.5) \geq 2$, $N_i(\mathcal{K}, 0.8) \geq 15$, and $h_i/w_i \geq 2$.
	B	$N_i(\mathcal{H}, 0.5) \geq 1$ and $\max\{N_i(\mathcal{U}, 0.9), N_i(\mathcal{L}, 0.9)\} \geq 4$.
Side	A	$N_i(\mathcal{H}, 0.8) \geq 1$ and $N_i(\mathcal{B}, 0.8) \geq 8$.

Pose-aware detection, tracking, and capture. We process every fifth frame using RTMO for person detection and keypoint estimation, and link detections using ByteTrack. We retain candidate crops x_i with height $h_i \geq 128$ pixels and width $w_i \geq 64$ pixels. Each candidate is assigned a coarse viewpoint $r_i \in \{\text{front}, \text{side}, \text{back}\}$ and a view-dependent quality grade using its 17 predicted keypoints K_i.

Let $\mathbf{p}_{i,k} = (u_{i,k}, v_{i,k})$ denote the position of keypoint k, with indices 0, 5, and 6 denoting the nose, left shoulder, and right shoulder, respectively. The cosine of the angle subtended by the shoulders at the nose is

$$\rho_i = \frac{(\mathbf{p}_{i,5} - \mathbf{p}_{i,0})^\top (\mathbf{p}_{i,6} - \mathbf{p}_{i,0})}{\|\mathbf{p}_{i,5} - \mathbf{p}_{i,0}\|_2 \|\mathbf{p}_{i,6} - \mathbf{p}_{i,0}\|_2}. \tag{16}$$

We assign the side viewpoint when $0.6 \leq \rho_i \leq 1$, the nose lies near or beyond an endpoint of the shoulder segment, and $|u_{i,5} - u_{i,6}| < w_i/4$. Otherwise, we assign the front viewpoint when

$$u_{i,a} > u_{i,b}, \qquad \forall (a, b) \in \{(5, 6), (7, 8), (9, 10), (11, 12)\}, \tag{17}$$

where the pairs correspond to the left and right shoulders, elbows, wrists, and hips. All remaining candidates are assigned the back viewpoint.

Candidate quality is assessed using keypoint confidence and crop aspect ratio. Let $\gamma_{i,k}$ denote the confidence of keypoint k, and define the head, upper-body, and lower-body sets as $\mathcal{H} = \{0, \ldots, 4\}$, $\mathcal{U} = \{5, \ldots, 10\}$, and $\mathcal{L} = \{11, \ldots, 16\}$. We use $\mathcal{B} = \mathcal{U} \cup \mathcal{L}$ for body keypoints and $\mathcal{K} = \mathcal{H} \cup \mathcal{B}$ for all keypoints. The number of keypoints in a set $\mathcal{S}$ with confidence above t is

$$N_i(\mathcal{S}, t) = \sum_{k \in \mathcal{S}} \mathbf{1}[\gamma_{i,k} > t]. \tag{18}$$

Table 4 summarizes the view-dependent grading rules. For front and back views, grade B is assigned only when the grade A criteria are not satisfied.

Candidates are buffered until a track has been absent for 60 frames or its length reaches 1,500 frames, and then undergo the tracklet-cleaning procedure described below. After cleaning, we retain the highest-quality surviving candidate for each available viewpoint, prioritizing grade A over grade B. The viewpoint labels are retained for view-specific prototype construction, while candidate quality informs the reliability weights q_i used in cleaning and prototype estimation.

Bootstrap features and tracklet cleaning. A frozen visual encoder Φ extracts normalized person descriptors:

$$\phi_i = \frac{\Phi(x_i)}{\|\Phi(x_i)\|_2}. \tag{19}$$

Our implementation uses the pre-projector visual tokens of Qwen2.5-VL, mean-pooled within the person crop. The encoder is used only for data construction and is excluded from the pretrained and downstream models.

To reduce tracking-induced identity switches, we iteratively compute a leave-one-out prototype and each candidate's deviation:

$$\bar{\phi}_{-i} = \frac{\sum_{j \neq i} q_j \phi_j}{\left\| \sum_{j \neq i} q_j \phi_j \right\|_2}, \qquad d_i^{\text{trk}} = 1 - \phi_i^\top \bar{\phi}_{-i}, \tag{20}$$

Table 5: Four association regimes in VPID. Thresholds are implementation-specific operating points.

Step	Recording scope	View relation	Supporting evidence	τ_l
1	Within video	Same view	Same camera, short time	0.85
2	Cross video	Same view	Appearance and time prior	0.80
3	Within video	Cross view	Track/time continuity	0.75
4	Cross video	Cross view	Appearance and time prior	0.70

where q_i combines detection, pose, and image-quality confidence. We discard the candidate with the largest deviation while $\max_i d_i^{\mathrm{trk}} > \sigma_{\mathrm{trk}}$. Surviving view fragments from the same cleaned tracklet are must-linked, while their view-specific statistics remain separate.

View-conditioned similarity. We maintain view-specific cluster prototypes using Eq. 2. For clusters C_i and C_j, the appearance similarity under view pair (r, r') is

$$s_{ij}^{r,r'} = (\mu_{C_i}^{r})^{\top} \mu_{C_j}^{r'}. \tag{21}$$

Camera-time compatibility. When a reliable camera-pair transition prior is available, we denote its normal travel-time interval by $[T_{c_i,c_j}^{\min}, T_{c_i,c_j}^{\max}]$. Widths δ^- and δ^+ control the early and late penalty regions:

$$g(\Delta t) = \begin{cases} 0, & \Delta t < T^{\min} - \delta^- \text{ or } \Delta t > T^{\max} + \delta^+, \\ 1 - \left(\dfrac{T^{\min} - \Delta t}{\delta^-} \right)^2, & T^{\min} - \delta^- \le \Delta t < T^{\min}, \\ 1, & T^{\min} \le \Delta t \le T^{\max}, \\ 1 - \left(\dfrac{\Delta t - T^{\max}}{\delta^+} \right)^2, & T^{\max} < \Delta t \le T^{\max} + \delta^+. \end{cases} \tag{22}$$

Camera-pair subscripts are omitted for clarity. The adjusted association score is $\tilde{s}_{ij}^{r,r'} = s_{ij}^{r,r'} g(\Delta t)$. For example, a normal interval of $[4, 5]$ minutes with $\delta^- = \delta^+ = 3$ minutes gives a feasible interval of $[1, 8]$ minutes. If no reliable transition prior exists, we set $g = 1$. Simultaneously visible tracklets from the same camera provide a cannot-link constraint.

Progressive association and label assignment. Table 5 summarizes the four association regimes. Thresholds are selected on a held-out audited subset and fixed before downstream evaluation.

At step l, a candidate edge is accepted only when: (i) its adjusted similarity exceeds τ_l; (ii) the clusters are reciprocal top-k neighbors within the same relation type; (iii) no camera-time cannot-link constraint is violated; and (iv) all affected view-specific prototypes satisfy a compactness bound after the tentative merge. Accepted clusters are merged and their prototypes are recomputed before subsequent association.

After the final step, a constraint-checked transitive closure produces the connected components. Each component defines a fixed pseudo-identity $\tilde{y}$, while every image retains its camera label c and viewpoint r. The resulting SurvID-6M contains over six million images collected from 49 cameras and organized into 163,992 pseudo-identities.

Pre-training objective. We optimize the Swin-B backbone B_θ and base identity projection E_0 using pseudo-identity cross-entropy and batch-hard triplet losses:

$$\mathcal{L}_{\mathrm{pre}} = \mathcal{L}_{\mathrm{CE}}(\tilde{y}) + \lambda_{\mathrm{tri}} \mathcal{L}_{\mathrm{tri}}(\tilde{y}). \tag{23}$$

Identity-balanced, confidence-weighted sampling limits the influence of large or low-confidence pseudo-identities. Pre-training uses neither a teacher–student loss nor downstream annotations.

Scaling Analysis. We examine how pre-training data scale affects frozen-model direct transfer, with detailed results provided in Figure 3

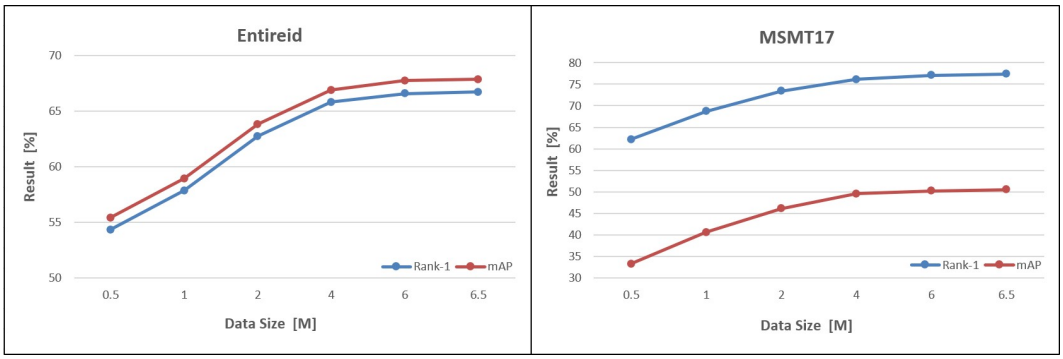


Figure 3: Effect of pre-training data scale on direct-transfer mAP on ENTIReID and MSMT17.

Table 6: Datasets and evaluation protocols used in our experiments.

Setting	**Dataset**	**Evaluation protocol**
Standard	Market1501 (Zheng et al., 2015)	Standard within-dataset retrieval
	MSMT17 (Wei et al., 2018)	Standard within-dataset retrieval
	ENTIRe-ID (Yıldız & Kasım, 2024)	Blurred-image variant
Clothes-changing	PRCC (Yang et al., 2019)	Clothes-changing protocol
Occluded	Occluded-ReID (Zhuo et al., 2018)	Occluded-query retrieval
	Occluded-Duke (Miao et al., 2019)	Occluded-query retrieval
Visible–infrared	SYSU-MM01 (Wu et al., 2017)	single-shot IR→VIS;
	LLCM (Zhang & Wang, 2023)	VIS→IR
	RegDB (Nguyen et al., 2017)	VIS→IR
Sketch-based	PKU-Sketch (Pang et al., 2018)	Sketch-to-photo retrieval
	Market-Sketch-1K (Lin et al., 2023)	Sketch-to-photo retrieval; fixed test split; single query
Aerial–ground	AG-ReID v1 (Nguyen et al., 2023)	A→G; G→A
	AG-ReID v2 (Nguyen et al., 2024)	A→C; C→A; A→W; W→A
	G2APS-ReID (Wang et al., 2025)	A→G; G→A
	LAGPeR (Wang et al., 2025)	A→G; G→A; G→(A+G)

Directions are written as query→gallery. A, G, C, and W denote aerial, ground, CCTV, and wearable views, respectively. VIS, IR, and NIR denote visible, infrared, and near-infrared images.

B Experiments Details

Pseudo-identity Quality Audit. Table 10 reports an audit of 200 sampled associations from each of the four video–view regimes and 500 within-cluster pairs from the final pseudo-identities. Precision is computed as the proportion of correct associations among determinate cases, excluding undetermined samples. Association precision ranges from 96.4After multi-stage association, the final pseudo-identities achieve a sampled-pair precision of 88.1These results quantify the consistency of the resulting identity supervision while revealing residual pairwise label noise.

Stage I pre-training. We train the Swin-B backbone on SurvID-6M, which contains over six million person images with automatically discovered pseudo-identity labels. Input images are directly resized to 224×224 pixels. We use identity-based $P \times K$ sampling, where P denotes the number of identities and K the number of images per identity. Each mini-batch contains $P = 16$ identities and $K = 8$ images per identity, yielding a global batch size of 128. All backbone parameters are optimized for 60 epochs using SGD, with an initial learning rate of 1×10^{-4} and a cosine annealing schedule. We select the checkpoint with the highest mAP on a held-out validation split of the pre-training corpus and use the same frozen checkpoint for all downstream direct-transfer evaluations.

Table 7: Additional results on aerial–ground ReID benchmarks. Each entry reports Rank-1 / mAP (%). "–" indicates an unreported result. Bold highlights the better value for metrics reported by both methods.

Dataset	Protocol	UniGeo (Su et al., 2026)	**TRUSTREID**
AG-ReID v1	A→G	–	89.50 / 71.33
	G→A	–	91.21 / 80.65
AG-ReID v2	A→C	92.60 / **88.80**	**93.11** / 84.09
	C→A	88.90 / 83.90	**89.05 / 83.92**
	A→W	**93.70 / 90.10**	90.10 / 87.34
	W→A	90.50 / 85.60	**91.04 / 86.98**
G2APS-ReID	A→G	–	77.70 / 61.12
	G→A	–	72.07 / 59.11
LAGPeR	A→G	–	51.89 / 43.00
	G→A	–	48.16 / 43.28
	G→(A+G)	–	59.37 / 46.34

Expert training. Expert training uses the same input resolution, optimizer, initial learning rate, learning-rate schedule, and training duration as Stage I. We balance dataset participation across mini-batches. Each mini-batch contains two identities from each of two selected datasets, with eight images per identity, yielding $P = 4$, $K = 8$, and a global batch size of 32. Identity labels are kept distinct across datasets. The pretrained backbone remains frozen throughout expert training; only the expert-specific modules are optimized using their respective specialization objectives.

Regional color histogram supervision. To explicitly supervise chromatic content, an auxiliary head predicts regional Lab color histograms from the color representation $z_c(x)$. We divide the clothing mask into $R = 2$ spatial regions, corresponding to its upper and lower portions. For each region, we compute normalized soft histograms with $K = 16$ bins for each chromatic channel, a and b, excluding pixels outside the clothing mask. The auxiliary objective is

$$\mathcal{L}_{\text{hist}} = -\frac{1}{|\Omega_h|} \sum_{i \in \Omega_h} \frac{1}{2R} \sum_{r=1}^{R} \sum_{u \in \{a,b\}} \sum_{k=1}^{K} h_{i,r,u,k} \log \hat{h}_{i,r,u,k}, \tag{24}$$

where h denotes the target histogram and $\hat{h}$ the predicted distribution, obtained by applying softmax over histogram bins. The set Ω_h contains samples with valid regional mask coverage. This loss encourages the color representation to encode regional chromatic distributions and contributes to the color expert objective in Eq. (5).

Router training. After expert training, we freeze the backbone and experts and optimize only the router using Eq. 15. We train the router for 20 epochs with AdamW, an initial learning rate of 3×10^{-4}, and a weight decay of 1×10^{-4}. The learning rate increases linearly during the first epoch and then follows a cosine annealing schedule, decreasing to 1×10^{-6}.

We balance dataset participation across mini-batches. Each mini-batch contains eight identities from each of two selected datasets, with eight images per identity, yielding $P = 16$, $K = 8$, and a global batch size of 128. Training triplets are constructed within the same dataset and retrieval condition, excluding self-matches. Sampling ensures that valid positive and negative candidates are available under the corresponding retrieval protocol.

Before router training, we standardize each expert's similarity scores using statistics estimated from dataset-balanced training pairs. These statistics remain fixed during router training and are shared by the reference and router-weighted fusion. We set the reference coefficients to $\beta_e = 1$ for all experts, the target temperature to $\tau = 1$, the ranking margin to $\mu = 0.3$, and the utility loss weight to $\lambda_{\text{utility}} = 1$. Counterfactual utility targets are detached from gradient computation.

Table 8: Direct transfer of pretrained models across five ReID settings without target-dataset fine-tuning. Results are reported as Rank-1 / mAP (%). "–" denotes an unavailable result in the cited evaluation or our current measurements. Bold indicates the best reported value for each metric.

Dataset / Protocol	ISR (ResNet-50) (Dou et al., 2023)	ISR (Swin) (Dou et al., 2023)	CION (Swin-S) (Zuo et al., 2024)	SapiensID (ViT-B) (Kim et al., 2025)	AIO (Li et al., 2024)	TRUSTREID (Stage I)
Standard ReID						
Market1501	85.10 / 65.10	87.00 / 70.50	66.51 / 38.75	88.18 / 68.26	79.60 / 59.90	**95.49 / 88.63**
MSMT17	45.70 / 21.20	56.40 / 30.30	24.51 / 9.22	67.25 / 31.02	–	**77.28 / 51.18**
ENTIRe-ID (blurred)	–	–	13.89 / 13.75	18.36 / 17.39	–	**66.67 / 67.78**
Clothes-changing ReID						
PRCC (CC)	–	–	–	**78.75 / 72.60**	–	33.05 / 35.78
Occluded ReID						
Occluded-ReID	–	–	–	87.30 / 75.57	–	**97.70 / 96.25**
Occluded-Duke	–	–	–	–	–	**56.92 / 46.49**
Visible–infrared ReID						
SYSU-MM01 (IR→VIS)	–	–	–	–	**57.60 / 51.90**	31.62 / 32.81
LLCM (VIS→IR)	–	–	–	–	–	**29.60 / 33.90**
RegDB (VIS→IR)	–	–	–	–	–	**11.78 / 13.50**
Sketch-based ReID						
PKU-Sketch	–	–	–	–	**70.20 / 73.50**	62.00 / 65.79
Market-Sketch-1K	–	–	–	–	–	**7.81 / 10.79**



Figure 4: Expert training examples: original images, clothing-region low-pass and grayscale transformations, silhouette targets, human-parsing labels, and label overlays.

We evaluate the router after each epoch on held-out validation splits drawn from the participating training sets. A single checkpoint is selected by the unweighted mean validation mAP across datasets.

Joint fine-tuning. After router training, we keep the expert parameters frozen and jointly fine-tune the shared backbone, identity branch, and router with a reduced learning rate. We retain the identity-supervised objective for the identity branch and the routing objective $\mathcal{L}_{\text{router}}$. Expert-specific specialization losses are not used during this phase. Counterfactual supervision targets remain detached from gradient computation. The final model uses a single checkpoint for evaluation across the corresponding downstream test sets.

Hardware. Training is conducted on four NVIDIA RTX 4090 GPUs. All batch sizes reported above are global batch sizes across the four GPUs.

Table 9: Direct transfer of the frozen Stage I model on aerial–ground ReID benchmarks. All results are reported in percentages. A, G, C, and W denote aerial, ground, CCTV, and wearable views, respectively.

Dataset	Protocol	Rank-1	mAP
AG-ReID v1	A→G	69.28	61.48
	G→A	70.58	63.84
AG-ReID v2	A→C	70.63	60.51
	C→A	67.92	58.90
	A→W	68.90	57.36
	W→A	67.09	57.98
G2APS-ReID	A→G	29.70	20.60
	G→A	35.07	22.81
LAGPeR	A→G	48.33	37.59
	G→A	44.81	41.28
	G→(A+G)	56.37	40.34

Table 10: VPID Multi-stage Clustering Association Audit Report.

Association Type	Sample Size	Undetermined	Correct	Incorrect	Precision	Threshold
Intra-video, Same-view	200	3	190	7	96.4%	0.88
Cross-video, Same-view	200	5	178	17	91.3%	0.82
Intra-video, Cross-view	200	8	164	28	85.4%	0.75
Cross-video, Cross-view	200	12	146	42	77.7%	0.68
Sampled-pairs in Final Pseudo-identities	500	20	423	57	88.1%	–

C Datasets and Evaluation Protocols

Table 6 summarizes the datasets and evaluation protocols across six ReID settings: standard, clothes-changing, occluded, visible–infrared, sketch-based, and aerial–ground. We use the blurred-image variant of ENTIRe-ID and the fixed test split with single-query evaluation for Market-Sketch-1K. All cross-modal and cross-view retrieval directions are specified in query-to-gallery order.

D Additional Evaluation

D.1 Joint-Training Performance on Aerial–Ground Benchmarks

Table 7 reports the aerial–ground retrieval performance of the same Stage II checkpoint on AG-ReID v1, AG-ReID v2, G2APS-ReID, and LAGPeR. We include the available UniGeo results on AG-ReID v2 for comparison. Results on the remaining benchmarks extend the evaluation coverage of TrustReID. A, G, C, and W denote aerial, ground, CCTV, and wearable views, respectively.

D.2 Direct Transfer of the Pretrained Model

Table 8 shows that our frozen Stage I model achieves 88.63% and 51.18% mAP on Market1501 and MSMT17, exceeding the published SapiensID results by 20.37 and 20.16 percentage points, respectively. It also reaches 96.25% mAP on Occluded-ReID. These results demonstrate strong transfer performance on standard and occluded ReID benchmarks using the representations learned through view-conditioned pseudo-identity discovery and identity-supervised pre-training. AIO reports stronger cross-modal results, although its source training includes infrared and sketch images, whereas our Stage I uses visible images only. Additional direct-transfer results on four aerial–ground benchmarks are provided in Table 8.

Table 9 reports direct-transfer results on AG-ReID v1, AG-ReID v2, G2APS-ReID, and LAGPeR. The same frozen Stage I checkpoint is evaluated across eleven retrieval protocols without down-

stream training or adaptation. These results extend the evaluation of the pretrained representation to aerial–ground matching.

E EXPERT TRAINING SAMPLES

Figure 4 illustrates the data used for expert specialization. Clothing-region low-pass filtering produces positive examples for the color expert and negative examples for the texture expert, while grayscale conversion reverses these roles. Silhouette targets and human-parsing labels supervise the body-structure and visible-part experts, respectively. Label overlays are included for visualization only.