HEAD SIMILARITY: MODELING STRUCTURED WHOLE-HEAD APPEARANCE BEYOND FACE RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Many vision applications require consistency in both identity and appearance beyond biometric verification, particularly across viewpoints or when facial cues are limited. Conventional face recognition models prioritize identity discrimination and are not explicitly trained to distinguish appearance states within the same identity. We introduce *Head Similarity*, a whole-head matching task that determines whether two images depict the same person with a consistent head appearance. We formulate this as exact identity–appearance matching, where same-identity, same-state pairs are matches and all other pairs are non-matches. To retain identity discrimination beyond this binary objective, our method additionally encourages same-identity, different-state pairs to score above different-identity pairs. We construct a benchmark from long-form videos with human-verified identity and appearance-state annotations for both training and testing, covering diverse viewpoints, occlusions, and temporal appearance changes. We develop a Dual-CLS framework combining identity distillation, hierarchical similarity learning, global score supervision, and training-only patch-pair auxiliary learning. Experiments show improved identity–appearance matching over the evaluated baselines while retaining strong identity verification performance.

1 INTRODUCTION

Many vision applications require *identity consistency* rather than strict biometric recognition. For example, video generation, reenactment, and editing systems need a person to remain visually consistent across frames despite changes in viewpoint, expression, or hairstyle Wang et al. (2018c); Song et al. (2022); Siarohin et al. (2019). The goal is not to verify legal identity, but to preserve the perception that the sequence depicts the same person. Conventional face recognition is designed for a different objective. Modern systems learn identity-invariant embeddings with margin-based metric learning losses Deng et al. (2019); Kim et al. (2022), deliberately suppressing intra-identity variations such as hairstyle, makeup, or aging. While effective for biometric verification, such embeddings often collapse different appearance states of the same identity into a single cluster. As shown in Fig. 1, AdaFace scores can be inconsistent with whole-head appearance similarity on synthetic examples generated by Qwen Xu et al. (2025): a visually closer same-identity sample may receive a lower similarity score. This suggests that identity-centric embeddings are insufficient for appearance-sensitive whole-head similarity. This limitation becomes more pronounced in real-world scenarios with partial, non-frontal, or rear-view observations, where facial cues may be weak or invisible. In such cases, whole-head cues such as hair, contour, and accessories remain important for perceived identity consistency. Recent hairstyle modeling and transfer works Chang et al. (2023) further highlight the importance of head appearance beyond facial identity. These observations reveal a gap between face similarity and holistic head-level similarity.

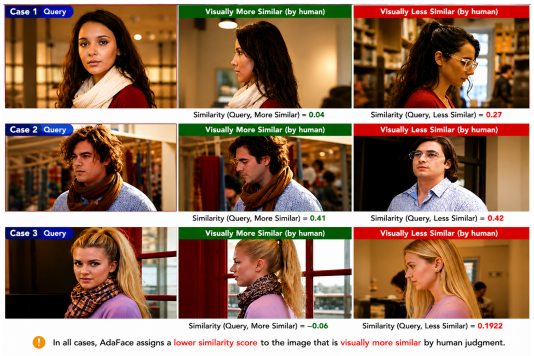


Figure 1: Failure cases of AdaFace on whole-head similarity.

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.

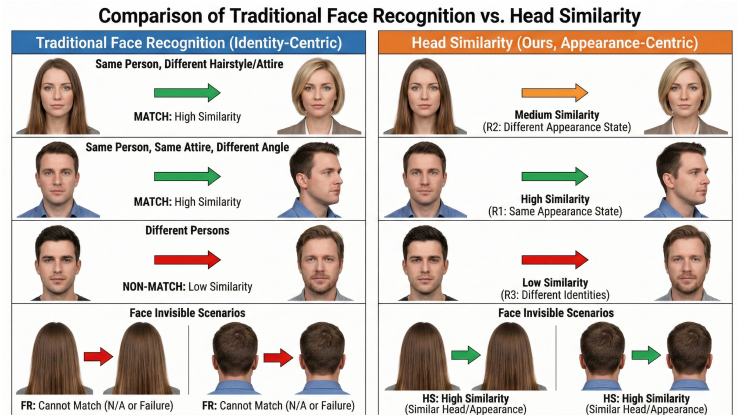


Figure 2: Conceptual illustration of our Head Similarity modeling. The primary task distinguishes R_1 from $R_2 \cup R_3$, while our method additionally encourages $R_1 > R_2 > R_3$.

To address this gap, we introduce *Head Similarity*, a whole-head task for exact identity–appearance matching. Unlike face recognition, it distinguishes different appearance states within the same identity. Our method additionally preserves identity-aware structure by encouraging same-identity, different-state pairs to score above different-identity pairs (Fig. 2). We construct a long-form video benchmark with human-verified identity and appearance-state annotations and propose a Dual-CLS framework with identity distillation, hierarchical similarity learning, global score supervision, and patch-pair auxiliary learning.

In summary, our contributions are:

- We introduce *Head Similarity*, a whole-head matching task for identifying exact identity–appearance matches beyond identity-invariant face recognition.
- We construct a benchmark through automatic temporal grouping and identity clustering followed by human verification of the resulting identity and appearance-state assignments for both training and testing.
- We propose a Dual-CLS framework that augments identity–appearance matching with identity-preserving hierarchical similarity learning, together with identity distillation, global score supervision, and patch-pair auxiliary learning.
- We show improved whole-head identity–appearance matching over the evaluated baselines while retaining strong identity verification performance.

2 Related Works

2.1 Face Recognition and Person Re-Identification

Modern face recognition learns identity-invariant embeddings using margin-based losses such as ArcFace Deng et al. (2019), CosFace Wang et al. (2018b), and AdaFace Kim et al. (2022), typically with ResNet He et al. (2016) or ViT Dosovitskiy et al. (2020) backbones. These methods collapse intra-identity variations into compact clusters, which benefits verification but suppresses appearance changes. Person Re-ID instead matches identities across cameras using global appearance cues Sun et al. (2019); Wang et al. (2018a); He et al. (2021), often assuming consistent clothing, although recent methods consider clothing changes Gu et al. (2022); Qian et al. (2020). In contrast to both tasks, we operate on whole-head inputs and explicitly model structured appearance states within the same identity.

2.2 Robust Recognition and Structured Similarity Learning

Pose-invariant and occlusion-robust methods address extreme views or missing regions through frontalization or feature learning Tran et al. (2017); Huang et al. (2017); Wan & Chen (2017); An et al. (2021), while Head360 He et al. (2024) extends recognition to full-head geometry. Metric-learning methods structure embedding spaces using contrastive or proxy-based objectives Schroff et al. (2015); Hadsell et al. (2006); Movshovitz-Attias et al. (2017); Khosla et al. (2020), and hierarchical approaches model relations across classes Kim et al. (2023); Gonzalez-Jimenez et al. (2024). Unlike these methods, our primary task focuses on whole-head identity–appearance matching, while our model additionally preserves identity-aware structure across appearance states.

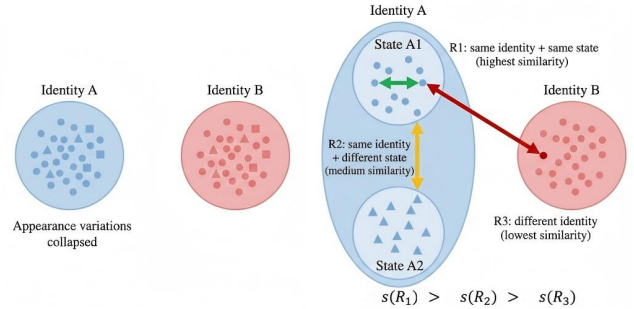
3 PROBLEM FORMULATION

Identity-Centric Face Recognition. Conventional face recognition learns identity-invariant embeddings from aligned facial crops. Given $\mathcal{D}_{FR} = \{(x_i, u_i)\}_{i=1}^{N}$, where u_i denotes identity, the objective is to learn an embedding function f_θ such that

$$s_\theta(x_i, x_j) > s_\theta(x_i, x_k), \quad \text{whenever } u_i = u_j,\ u_i \neq u_k,$$

where s_θ denotes cosine similarity. This encourages compact identity clusters and reduces sensitivity to intra-identity appearance variations such as hairstyle or grooming changes. While effective for biometric verification, it is less suitable for appearance-sensitive whole-head matching, where different appearance states of the same person may need to be distinguished.

Head Image Factors. We consider a head image x to depend on identity u, appearance state a, and nuisance factors e. The appearance state describes relatively stable visual attributes, including hairstyle, visible head contour, makeup, and accessories, while illumination, background, and minor pose are treated as nuisance factors.



(a) Identity-Centric Face Recognition (b) Hierarchical Head Similarity (Ours)

Figure 3: Illustration of our identity-preserving Head Similarity modeling. Head Similarity treats R_1 as exact identity–appearance matches, while our model additionally encourages $R_1 > R_2 > R_3$ to preserve identity-aware structure.

Two samples share the same appearance state when these attributes remain visually consistent despite nuisance changes. Conventional face recognition primarily preserves identity while suppressing intra-identity appearance variation. In contrast, Head Similarity requires sensitivity to both identity and appearance state for exact matching while remaining robust to nuisance factors.

We define

$$\mathcal{D} = \{(x_i, u_i, a_i)\}_{i=1}^{N}.$$

For training, temporal grouping is used to generate candidate appearance states, which are subsequently manually verified to ensure appearance consistency within each state.

Head Similarity Task. Based on identity and appearance state, we define

$$\begin{aligned} R_1 &= \{(i,j) \mid u_i = u_j \wedge a_i = a_j\}, \\ R_2 &= \{(i,j) \mid u_i = u_j \wedge a_i \neq a_j\}, \\ R_3 &= \{(i,j) \mid u_i \neq u_j\}. \end{aligned} \tag{1}$$

Here, R_1 denotes the same identity and appearance state, R_2 the same identity but different appearance states, and R_3 different identities.

Head Similarity treats R_1 as exact identity–appearance matches and $R_2 \cup R_3$ as non-matches. Thus, the primary task is to distinguish exact identity–appearance matches from all other pairs.

Identity-Preserving Similarity Structure. Beyond the binary Head Similarity objective, our model additionally preserves identity-aware structure. Let $f_\theta : \mathcal{X} \to \mathbb{R}^d$ and define

$$s_\theta(x_i, x_j) = \cos\big(f_\theta(x_i), f_\theta(x_j)\big).$$

We encourage

$$s_\theta(x_i, x_j) > s_\theta(x_i, x_k) > s_\theta(x_i, x_\ell),$$

for $(i,j) \in R_1$, $(i,k) \in R_2$, and $(i,\ell) \in R_3$.

As illustrated in Fig. 3, this structure encourages R_1 pairs to receive the highest similarity, while R_2 pairs remain more similar than R_3 pairs. This ordering is an additional modeling objective rather than a requirement of the Head Similarity task itself.

4 METHOD

As illustrated in Fig. 4, our Dual-CLS backbone produces an identity embedding z^{id} and an auxiliary head-similarity embedding z^{head}. Both embeddings receive hierarchical similarity supervision, while z^{id} additionally receives identity distillation and global score supervision. A training-only patch-pair classifier supervises the shared backbone through patch-token interactions. Inference uses only z^{id} for cosine similarity.

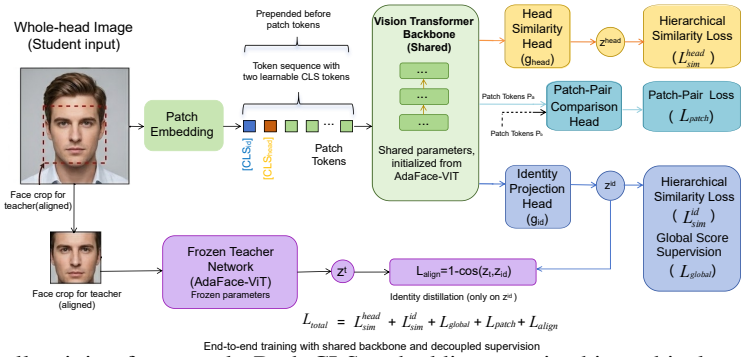


Figure 4: Overall training framework. Both CLS embeddings receive hierarchical supervision, while the identity embedding additionally receives identity distillation and global score supervision. P_a and P_b denote patch-token sequences from the anchor x_a and paired image x_b. A training-only Patch-Pair Comparison Head provides auxiliary supervision to the backbone. Only the identity embedding is used for inference.

4.1 DUAL-CLS VISION TRANSFORMER

Building on Sec. 3, we aim to model whole-head appearance while preserving identity discrimination and reducing sensitivity to nuisance factors. Directly adapting a face-aligned backbone to whole-head inputs induces distribution shift and tends to collapse intra-identity appearance differences. We therefore adopt a dual-CLS Vision Transformer that explicitly separates identity preservation from appearance-sensitive head similarity within a shared backbone. We instantiate the backbone with a ViT model pre-trained using AdaFace. Unlike conventional face-recognition systems that operate on aligned facial crops, our network takes whole-head images as input and can exploit cues such as hairstyle and head contour. Following prior work on multiple learnable tokens in transformers Ryoo et al. (2021), we adopt a simple dual-CLS design as a baseline architecture for the proposed framework.

Two CLS tokens. Given an input head image x, we divide it into non-overlapping patches and project them into patch tokens. We prepend two learnable tokens, CLS^{id} and CLS^{head}, forming $[\text{CLS}^{id}, \text{CLS}^{head}, x_{\text{patches}}] \in \mathbb{R}^{(2+N)\times d}$, where N is the number of patches and d is the embedding dimension. Both tokens participate in self-attention and aggregate global head-level information, but are optimized under different objectives.

Identity and Head-Similarity Embeddings. We extract two CLS tokens from the transformer and project them via g_{id} and g_{head}, followed by L2 normalization: z^{id}, $z^{head} \in \mathbb{R}^d$. The identity embedding z^{id} is used at inference and trained with identity distillation, hierarchical similarity supervision, and global score supervision. The auxiliary embedding z^{head} is only used during training to provide appearance-sensitive signals and is discarded at inference. Both embeddings are supervised by the hierarchical similarity objective, while z^{id} is additionally aligned with a frozen teacher.

4.2 DISTILLATION-BASED IDENTITY PRESERVATION

To preserve compatibility with conventional face-recognition embeddings, we align z^{id} with a frozen teacher pre-trained on aligned face images. Given a face crop x^{face} and a corresponding head image x^{head}, the teacher produces z^t and the student produces z^{id}. Distillation is applied only when valid face crops are available. We minimize: $L_{\text{align}} = 1 - \cos(z^t, z^{id})$. The teacher remains frozen. This preserves identity consistency under distribution shift from face crops to whole-head inputs, while allowing appearance variation to be modeled via the similarity objective.

4.3 SIMILARITY LEARNING

For each anchor x_{anc}, we sample a counterpart x_k for each relation R_k, $k \in \{1, 2, 3\}$ (Sec. 3). Their cosine similarity is $s^{(k)} = z_{\text{anc}}^{\top} z_k$, where z denotes the L2-normalized embedding from either branch. Expectations below are over sampled training examples.

Hierarchical supervision. Beyond the binary matching objective, we preserve identity-aware structure through hierarchical supervision. We encourage $s^{(1)} > s^{(2)} > s^{(3)}$ through:

$$L_{\text{sim}} = \mathbb{E}\Big[\,\text{Softplus}(m_1 + s^{(2)} - s^{(1)}) + \text{Softplus}(m_2 + s^{(3)} - s^{(1)}) + \text{Softplus}(m_3 + s^{(3)} - s^{(2)})\Big], \tag{2}$$

where $m_1, m_2, m_3 > 0$ are margins for the (R_1, R_2), (R_1, R_3), and (R_2, R_3) comparisons, respectively. The loss is applied separately to z^{id} and z^{head}.

Global score supervision. To complement within-anchor ordering, we compare positive and negative cosine scores from z^{id} that need not share an anchor. Let s^+ be an R_1 score and s^- an R_2 or R_3 score, with the two negative relations sampled equally. We optimize:

$$L_{\text{global}} = \tau\, \mathbb{E}_{(s^+, s^-)} \left[\text{Softplus}\left(\frac{m_g + s^- - s^+}{\tau}\right)\right], \tag{3}$$

where $m_g > 0$ is the separation margin and $\tau > 0$ is the temperature.

Patch-pair auxiliary supervision. For an image pair (x_i, x_j) with backbone patch tokens P_i, P_j, a shared learnable projection Φ gives $U = \Phi(P_i)$ and $V = \Phi(P_j)$. Multi-head cross-attention computes $M = \text{MHA}(U, V, V)$. A shared MLP processes $[U, M, |U - M|, U \odot M]$, where $[\cdot]$ denotes feature concatenation and $\odot$ element-wise multiplication. Its outputs are mean- and max-pooled over tokens, concatenated, and mapped by a shared classifier to a directional logit $q(U, V)$. We symmetrize it as $\ell(x_i, x_j) = [q(U, V) + q(V, U)]/2$.

With $\ell^{(k)} = \ell(x_{\text{anc}}, x_k)$, we treat R_1 as positive and R_2, R_3 as negative:

$$L_{\text{patch}} = \frac{1}{4}\mathbb{E}\left[2b(\ell^{(1)}, 1) + b(\ell^{(2)}, 0) + b(\ell^{(3)}, 0)\right], \tag{4}$$

where $b(u, y) = \text{Softplus}(u) - yu$ is binary cross-entropy with logits. The weights balance the positive and combined negative terms. This training-only loss updates the shared backbone through patch tokens. Inference uses only cosine similarity between z^{id} embeddings.

4.4 DATASET CONSTRUCTION

The dataset construction pipeline is illustrated in Fig. 5. **Shot segmentation** is first performed using an HSV histogram-based method Truong & Venkatesh (2007) to obtain temporally coherent segments. Within each shot, we conduct **head detection and tracking** using RetinaFace Deng et al. (2020) and short-term IoU-based tracking to produce local track segments. We then apply **segment filtering and representative crop selection**, retaining segments with at least 5 frames, $\geq$20% face-visible frames, and $\geq$10% face-invisible frames. For each segment, a *best-face* frame (face area >50% of the head box) is selected for identity clustering, while full head crops are preserved for training. Next, **cross-video identity clustering** is performed using face embeddings, providing pair relations for training, where samples within a segment are treated as R_1, samples across segments within the same identity cluster as R_2, and samples from different clusters as R_3. All retained training samples and appearance-state assignments are subsequently manually verified.

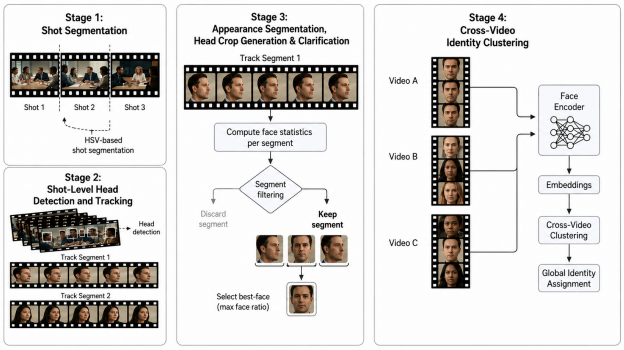


Figure 5: Four-stage construction pipeline of the HeadSim-Head dataset.

5 EXPERIMENTS

5.1 DATASETS

We use VGGFace2 Cao et al. (2018) and VoxCeleb2 Chung et al. (2018) for identity distillation, and HeadSim-Head for evaluation. VGGFace2 is used only for distillation, where aligned face crops are processed by a frozen teacher and corresponding head images are fed to the student. VoxCeleb2 is used for both distillation and evaluation. We follow the standard verification protocol on the official test split with strict train/test separation. HeadSim-Head is our benchmark dataset (Sec. 4.4), with human-verified identity and appearance-state annotations for both the training and test splits. As summarized in Table 1, the dataset contains 731/187 identities, 6,126/2,489 videos, and 39,890/8,870 images for train/test splits, with disjoint identities. Each identity has multiple appearance states (avg. 8), covering diverse poses (38% frontal, 53% profile, 9% back-view). HeadSim-Head identities do not overlap with VoxCeleb2 test identities. For evaluation, we replace the original background of each HeadSim-Head test image with a background randomly sampled from the COCO dataset Lin et al. (2014), while preserving the foreground head region. Training strategies, implementation details, and evaluation metrics are provided in Appendix A.

Table 1: HeadSim statistics

Metric	Train	Test
# Identities	731	187
# Videos	6,126	2,489
# Images	39,890	8,870
Avg. states	8	13

5.2 FROM FACE RECOGNITION TO WHOLE-HEAD INPUT

We analyze the effect of adapting face-recognition models from aligned-face inputs to unaligned whole-head images. As shown in Table 2, directly applying face-trained models substantially degrades verification performance, particularly at low FAR. At FAR=10^{-3}, TAR decreases from 0.912 to 0.081 for AdaFace-ViT and from 0.869 to 0.025 for ArcFace-ResNet50 Deng et al. (2019), highlighting the input-domain mismatch. TransReID-ViT He et al. (2021) also performs poorly, achieving TAR@FAR=10^{-3} of 0.025 and AUC of 0.570. These results suggest that strong person re-identification performance does not necessarily transfer to whole-head identity verification without task-specific adaptation.

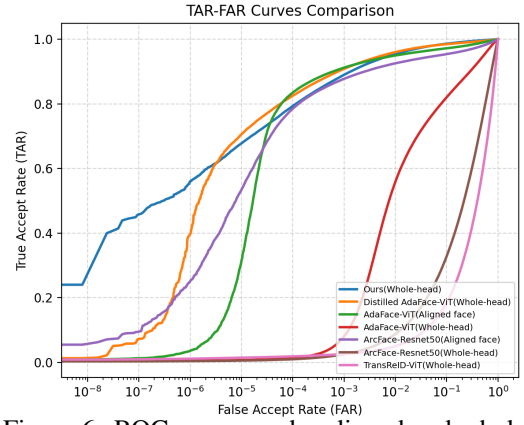


Figure 6: ROC curves under aligned and whole-head inputs.

Table 2: Face verification performance under aligned-face and whole-head inputs.

Model	Input	TAR@FAR=10^{-2}	TAR@FAR=10^{-3}	TAR@FAR=10^{-4}	AUC
AdaFace-ViT	Aligned face	0.950	0.912	0.834	0.987
AdaFace-ViT	Whole-head	0.561	0.081	0.014	0.920
Dist. AdaFace	Whole-head	0.960	0.908	0.825	0.993
ArcFace-Resnet50	Aligned face	0.924	0.869	0.782	0.977
ArcFace-Resnet50	Whole-head	0.083	0.025	0.010	0.700
TransReID-ViT	Whole-head	0.035	0.025	0.018	0.570
Ours	Whole-head	0.954	0.890	0.793	0.994

Distillation-based adaptation substantially improves whole-head verification performance. Among whole-head methods, Distilled AdaFace achieves the highest TAR at all reported operating points: 0.960, 0.908, and 0.825 at FAR=10^{-2}, 10^{-3}, and 10^{-4}, respectively, closely approaching the aligned-face AdaFace-ViT baseline at low FAR. Our method achieves corresponding TAR values of 0.954, 0.890, and 0.793, together with the highest reported AUC of 0.994. It substantially outperforms the directly applied whole-head baselines, retaining strong identity discrimination while modeling appearance-sensitive similarity. Compared with distillation-only adaptation, our method obtains a slightly higher AUC (0.994 versus 0.993) but lower TAR at all reported operating points, indicating that identity-only distillation remains more effective at low FAR. As further shown in Appendix A.4, adding similarity-learning objectives slightly reduces low-FAR TAR while preserving strong identity verification performance. Figure 6 presents the corresponding ROC comparison.

5.3 HEAD SIMILARITY

As shown in Table 3, directly applying identity-oriented models to whole-head similarity yields limited performance. AdaFace-ViT achieves TAR@FAR=10^{-3} of 0.222 and AUC of 0.782. Identity distillation improves these to 0.273 and 0.848, respectively, but remains well below our method under the identity–appearance matching protocol. These results suggest that identity-focused pretraining or adaptation alone is insufficient for strong performance on this benchmark. We further evaluate the framework with ArcFace-ResNet50 and TransReID-ViT. For ArcFace, adaptation increases TAR@FAR=10^{-3} from 0.223 to 0.475 and AUC from 0.769 to 0.977. For TransReID, the corresponding metrics improve from 0.199 to 0.212 and from 0.675 to 0.748, indicating more modest gains, particularly at low FAR. The complete AdaFace-ViT model achieves the highest performance across all reported metrics, with TAR values of 0.975, 0.919, and 0.828 at FAR=10^{-2}, 10^{-3}, and 10^{-4}, respectively, and AUC of 0.998. These results demonstrate its effectiveness in distinguishing identity–appearance matches from different-state and different-identity pairs. Figure 7 shows the corresponding ROC curves.

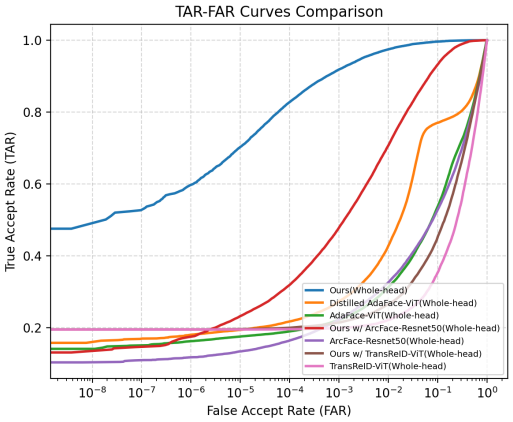


Figure 7: ROC curves comparison on the HeadSim-Head dataset.

Table 3: Head similarity performance on the HeadSim-Head dataset.

Model	Input	TAR@FAR=10^{-2}	TAR@FAR=10^{-3}	TAR@FAR=10^{-4}	AUC
Ours	Whole-head	0.975	0.919	0.828	0.998
Dist. AdaFace-ViT	Whole-head	0.427	0.273	0.217	0.848
AdaFace-ViT	Whole-head	0.313	0.222	0.190	0.782
Ours w/ ArcFace-ResNet50	Whole-head	0.706	0.475	0.319	0.977
ArcFace-ResNet50	Whole-head	0.326	0.223	0.165	0.769
Ours w/ TransReID-ViT	Whole-head	0.262	0.212	0.199	0.748
TransReID-ViT	Whole-head	0.221	0.199	0.196	0.675

Figure 8 presents top-3 retrieval results and similarity scores for AdaFace and our method on the HeadSim-Head test set under two settings: same background and background changed. Background replacement preserves the foreground head, allowing us to examine the effect of contextual changes on retrieval.

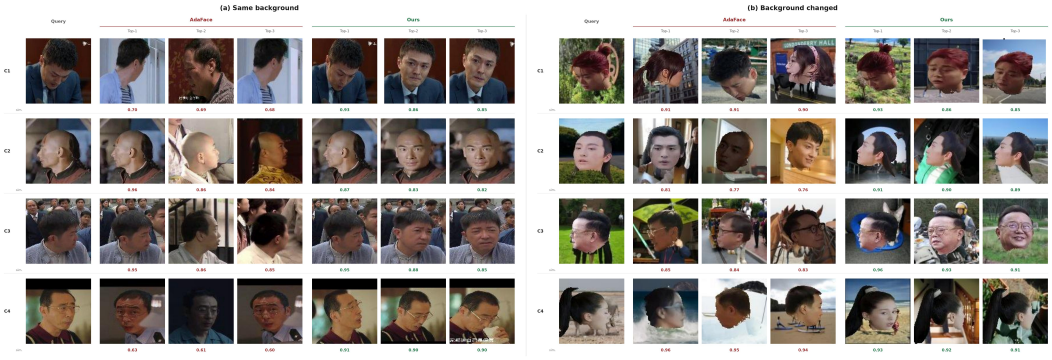


Figure 8: Top-3 retrieval results on the HeadSim-Head test set for AdaFace and our method under two settings: same background (left) and background changed (right). Each row shows a query and its top-3 retrieved samples with corresponding similarity scores.

In these examples, our method retrieves samples that better match the query's identity and overall head appearance than AdaFace under both background settings. These qualitative results complement the verification results in Table 3, although dataset-wide background robustness requires separate quantitative evaluation.

5.4 Ablation Studies on Architecture and Learning Objectives

We first compare architectural designs and examine the incremental effects of auxiliary supervision. We first evaluate the primary binary Head Similarity objective, and then examine whether the proposed identity-preserving supervision retains the additional ordering $R_1 > R_2 > R_3$.

Architectural design and auxiliary supervision. Table 4 compares four architectural configurations and two additional supervision variants based on Dual-CLS. *Shared Embedding* applies L_{align} and L_{sim} to one representation. *Dual-Head (Split)* assigns these objectives to separate branches, while *Dual-Head (Both)* additionally applies L_{sim} to the identity branch. *Dual-CLS* uses two CLS tokens to support the two objectives. Building on Dual-CLS, we first introduce L_{global} to separate positive and negative similarity scores across sampled anchors. We then add L_{patch} as a training-only patch-pair classification objective that provides additional supervision to the shared backbone.

Table 4: Ablation of architectural designs and auxiliary supervision. The loss column lists the ID-embedding objectives and, where applicable, the patch-pair auxiliary objective, which supervises the shared backbone through patch tokens.

Model	#CLS	Loss configuration	TAR@FAR=10^{-2}	TAR@FAR=10^{-3}	TAR@FAR=10^{-4}	AUC
Shared Embedding	1	$L_{\text{align}} + L_{\text{sim}}$	0.853	0.734	0.439	0.968
Dual-Head (Split)	1	L_{align}	0.360	0.217	0.185	0.843
Dual-Head (Both)	1	$L_{\text{align}} + L_{\text{sim}}$	0.879	0.773	0.498	0.971
Dual-CLS (Ours)	2	$L_{\text{align}} + L_{\text{sim}}$	0.959	0.901	0.760	0.996
Dual-CLS (Ours)	2	$L_{\text{align}} + L_{\text{sim}} + L_{\text{global}}$	0.965	0.906	0.811	0.996
Dual-CLS (Ours)	2	$L_{\text{align}} + L_{\text{sim}} + L_{\text{global}} + L_{\text{patch}}$	0.975	0.919	0.828	0.998

Among the four architectural designs without the additional global and patch-pair objectives, Dual-CLS achieves the highest reported performance, with TAR values of 0.959, 0.901, and 0.760 at FAR=10^{-2}, 10^{-3}, and 10^{-4}, respectively, and an AUC of 0.996. These results motivate our choice of Dual-CLS as the base architecture. Adding L_{global} improves the corresponding TAR values to 0.965, 0.906, and 0.811, with the largest gain of 5.1 percentage points at FAR=10^{-4}. Further introducing L_{patch} increases TAR to 0.975, 0.919, and 0.828, yielding additional gains of 1.0, 1.3, and 1.7 percentage points, respectively, while increasing AUC from 0.996 to 0.998. These results show improvements from both auxiliary objectives, with the largest TAR gains observed at the lowest reported FAR.

Comparison of learning objectives. Table 5 compares different learning objectives on the primary binary Head Similarity task. BCE directly separates R_1 from $\{R_2, R_3\}$, while triplet learning additionally imposes $R_1 > R_2$ and $R_1 > R_3$. Our hierarchical objective further introduces the identity-preserving relation $R_2 > R_3$. The complete method combines this structured objective with the auxiliary supervision described above.

Table 5: Verification performance of different learning configurations using Dual-CLS under a common evaluation protocol. All configurations retain L_{global} and L_{patch}; the main similarity objective and the use of distillation vary as indicated.

Architecture	Objective	Similarity supervision	Distillation	TAR@FAR=10^{-2}	TAR@FAR=10^{-3}	TAR@FAR=10^{-4}
Dual-CLS	BCE	R_1 vs. $\{R_2, R_3\}$	No	0.320	0.228	0.187
Dual-CLS	BCE	R_1 vs. $\{R_2, R_3\}$	Yes	0.880	0.823	0.722
Dual-CLS	Triplet	$R_1 > R_2,\ R_1 > R_3$	No	0.641	0.401	0.285
Dual-CLS	Triplet	$R_1 > R_2,\ R_1 > R_3$	Yes	0.950	0.824	0.687
Dual-CLS	Ours w/o $R_2 > R_3$	$R_1 > R_2,\ R_1 > R_3$	No	0.740	0.436	0.251
Dual-CLS	Ours w/o $R_2 > R_3$	$R_1 > R_2,\ R_1 > R_3$	Yes	0.962	0.834	0.703
Dual-CLS	Ours (Full)	$R_1 > R_2 > R_3$	Yes	**0.975**	**0.919**	**0.828**

For each of the three objective families evaluated both with and without distillation, the distilled configuration achieves higher TAR at all reported operating points. The relative ranking of the remaining configurations depends on the operating point. For example, distilled BCE achieves TAR@FAR=10^{-4} of 0.722, compared with 0.687 for distilled triplet learning and 0.703 for the distilled hierarchical variant without explicit $R_2 > R_3$ supervision. The complete method achieves the highest reported TAR values of 0.975, 0.919, and 0.828. These exceed the corresponding values of 0.962, 0.834, and 0.703 for the distilled variant without explicit $R_2 > R_3$ supervision by 1.3, 8.5, and 12.5 percentage points, respectively.

Evaluation of identity-preserving structure. We further evaluate whether the proposed hierarchical supervision preserves identity-aware ordering beyond the binary Head Similarity task. We evaluate all methods on the same independently human-verified test subset of HeadSim-Head, with verified R_1, R_2, R_3 relations. For each anchor, we sample one counterpart from each relation and compute cosine similarities using z^{id}. All methods share the same test samples, visibility-selection criteria, and sampling seed. We report the three pairwise ordering accuracies and complete-order accuracy $P(R_1 > R_2 > R_3)$.

Table 6: Evaluation of the identity-preserving similarity ordering induced by hierarchical supervision.

Architecture	Objective	Distillation	Explicit $R_2 > R_3$	$P(R_1 > R_2)$	$P(R_1 > R_3)$	$P(R_2 > R_3)$	$P(R_1 > R_2 > R_3)$
Dual-CLS	BCE	No	No	0.799	0.805	0.507	0.360
Dual-CLS	BCE	Yes	No	0.971	0.978	0.510	0.485
Dual-CLS	Triplet	No	No	0.961	0.969	0.641	0.579
Dual-CLS	Triplet	Yes	No	0.985	0.991	0.643	0.590
Dual-CLS	Ours w/o $R_2 > R_3$	No	No	0.985	0.996	0.804	0.759
Dual-CLS	Ours w/o $R_2 > R_3$	Yes	No	0.990	0.997	0.815	0.788
Dual-CLS	Ours (Full)	Yes	Yes	**0.992**	**0.998**	**0.864**	**0.837**

The full method achieves $P(R_1 > R_2) = 0.992$, $P(R_1 > R_3) = 0.998$, and $P(R_2 > R_3) = 0.864$, with a complete-order accuracy of 0.837. Notably, $P(R_2 > R_3)$ increases from near chance under BCE (0.507–0.510) to 0.641–0.643 under triplet supervision, and further to 0.804–0.815 under our formulation without explicit $R_2 > R_3$ supervision. Adding the intermediate-order term further improves $P(R_2 > R_3)$ from 0.815 to 0.864 and complete-order accuracy from 0.788 to 0.837. These results show that the proposed formulation preserves identity-aware similarity structure beyond the binary exact-match objective, while explicit $R_2 > R_3$ supervision further strengthens the intermediate ordering.

5.5 Human Evaluation

To evaluate alignment with human perception, we conduct a user study with 14 participants and 30 ranking questions, covering identity-aware hierarchy validation and retrieval preference. Options are randomly shuffled to reduce positional bias. The complete questionnaire is provided in Appendix B.

Identity-aware hierarchy validation. The first 15 questions examine whether the identity-aware ordering used by our model is consistent with human judgments. For each query, participants rank one candidate from each of R_1, R_2, and R_3 by overall head similarity. As shown in Table 7, participants predominantly rank R_1 first and R_3 last, supporting the perceptual plausibility of the identity-aware ordering.

Table 7: Human ranking results for hierarchy validation.

Relation	1st	2nd	3rd
R_1	84.8%	15.2%	0.0%
R_2	15.2%	81.4%	3.3%
R_3	0.0%	3.3%	96.7%

Retrieval preference. The remaining 15 questions compare retrieval results from AdaFace, TransReID, and our method. Most queries come from the HeadSim-Head test set, with two generated by Qwen and Gemini to probe generalization beyond real data. Participants rank the top-3 results from each method by similarity to the query. As shown in Table 8, our method ranks first in 84.3% of responses, ahead of AdaFace and TransReID.

Table 8: Human ranking results for retrieval preference.

Method	1st	2nd	3rd
AdaFace	15.2%	68.6%	16.2%
Ours	84.3%	14.8%	1.0%
TransReID	0.5%	16.7%	82.8%

6 Conclusion and Limitations

We introduced *Head Similarity*, a whole-head task for exact identity–appearance matching beyond identity-invariant face recognition. We constructed a benchmark and developed a Dual-CLS framework that augments this binary objective with identity-preserving hierarchical learning, identity distillation, global score supervision, and training-only patch-pair supervision. The resulting model improves identity–appearance matching while retaining strong identity verification performance using a single embedding at inference. Our work has several limitations: the dataset remains limited in scale and diversity, and the additional identity-aware ordering $R_1 > R_2 > R_3$ may not fully capture subjective human perception.

AI USE STATEMENT

We used generative AI tools to assist with language editing, LaTeX formatting, and code development. All AI-assisted code was manually reviewed and tested by the authors. We also used generative AI tools to generate the synthetic examples shown in Fig. 1; these examples were used only for illustration and were not included in training or quantitative evaluation. We take full responsibility for the final text, code, results, claims, and artifacts presented in this work.

REFERENCES

Xiang An, Xuhan Zhu, Yuan Gao, Yang Xiao, Yongle Zhao, Ziyong Feng, Lan Wu, Bin Qin, Ming Zhang, Debing Zhang, et al. Partial fc: Training 10 million identities on a single machine. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1445–1449, 2021.

Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pp. 67–74. IEEE, 2018.

Seunggyu Chang, Gihoon Kim, and Hayeon Kim. Hairnerf: Geometry-aware image synthesis for hairstyle transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2448–2458, 2023.

Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4690–4699, 2019.

Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5203–5212, 2020.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Alvaro Gonzalez-Jimenez, Simone Lionetti, Dena Bazazian, Philippe Gottfrois, Fabian Gröger, Alexander Navarini, and Marc Pouly. Hyperbolic metric learning for visual outlier detection. In *European Conference on Computer Vision*, pp. 327–344. Springer, 2024.

Xinqian Gu, Hong Chang, Bingpeng Ma, Shutao Bai, Shiguang Shan, and Xilin Chen. Clothes-changing person re-identification with rgb modality only. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1060–1069, 2022.

Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 2, pp. 1735–1742. IEEE, 2006.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

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*, pp. 15013–15022, 2021.

Yuxiao He, Yiyu Zhuang, Yanwen Wang, Yao Yao, Siyu Zhu, Xiaoyu Li, Qi Zhang, Xun Cao, and Hao Zhu. Head360: Learning a parametric 3d full-head for free-view synthesis in 360°. In *European Conference on Computer Vision*, pp. 254–272. Springer, 2024.

Rui Huang, Shu Zhang, Tianyu Li, and Ran He. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. In *Proceedings of the IEEE international conference on computer vision*, pp. 2439–2448, 2017.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

Minchul Kim, Anil K Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18750–18759, 2022.

Sungyeon Kim, Boseung Jeong, and Suha Kwak. Hier: Metric learning beyond class labels via hierarchical regularization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19903–19912, 2023.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Yair Movshovitz-Attias, Alexander Toshev, Thomas K Leung, Sergey Ioffe, and Saurabh Singh. No fuss distance metric learning using proxies. In *Proceedings of the IEEE international conference on computer vision*, pp. 360–368, 2017.

Xuelin Qian, Wenxuan Wang, Li Zhang, Fangrui Zhu, Yanwei Fu, Tao Xiang, Yu-Gang Jiang, and Xiangyang Xue. Long-term cloth-changing person re-identification. In *Proceedings of the Asian conference on computer vision*, 2020.

Michael S Ryoo, AJ Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. Token-learner: What can 8 learned tokens do for images and videos? *arXiv preprint arXiv:2106.11297*, 2021.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in neural information processing systems*, 32, 2019.

Linsen Song, Wayne Wu, Chen Qian, Ran He, and Chen Change Loy. Everybody's talkin': Let me talk as you want. *IEEE Transactions on Information Forensics and Security*, 17:585–598, 2022.

Yifan Sun, Liang Zheng, Yali Li, Yi Yang, Qi Tian, and Shengjin Wang. Learning part-based convolutional features for person re-identification. *IEEE transactions on pattern analysis and machine intelligence*, 43(3):902–917, 2019.

Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1415–1424, 2017.

Ba Tu Truong and Svetha Venkatesh. Video abstraction: A systematic review and classification. *ACM transactions on multimedia computing, communications, and applications (TOMM)*, 3(1):3–es, 2007.

Weitao Wan and Jiansheng Chen. Occlusion robust face recognition based on mask learning. In *2017 IEEE international conference on image processing (ICIP)*, pp. 3795–3799. IEEE, 2017.

Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of the 26th ACM international conference on Multimedia*, pp. 274–282, 2018a.

Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5265–5274, 2018b.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. *arXiv preprint arXiv:1808.06601*, 2018c.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.

APPENDIX

A TRAINING AND EVALUATION DETAILS

A.1 TRAINING STRATEGY

For training, we combine HeadSim-Head with large-scale face-recognition data. Face-recognition samples from VoxCeleb2 Chung et al. (2018) and VGGFace2 Cao et al. (2018) are used for distillation, where aligned face crops are processed by a frozen teacher and corresponding head images are fed to the student. We adopt a mixed-batch strategy, where head samples provide (R_1, R_2, R_3) relations with online hard negative mining, while face samples contribute only to the alignment loss. We apply background randomization by replacing backgrounds with random samples from the COCO dataset Lin et al. (2014) to reduce spurious correlations and improve robustness.

A.2 IMPLEMENTATION DETAILS

We use a ViT backbone initialized from an AdaFace-pretrained checkpoint. Whole-head images are resized and normalized following the teacher. We train with AdamW (lr=$1e{-4}$, weight decay=$5e{-2}$, batch size=256) for 30 epochs. Margins are set to $m_1 = 0.1$, $m_2 = 0.3$, and $m_3 = 0.2$. Experiments are conducted on two RTX 5090 GPUs.

A.3 EVALUATION METRICS

We report true accept rate (TAR) at fixed false accept rates (FAR), including FAR=10^{-2}, 10^{-3}, and 10^{-4}, together with the area under the ROC curve (AUC).

On VoxCeleb2, positive and negative pairs are defined according to whether the two samples share the same identity. On HeadSim-Head, R_1 pairs (same identity and same appearance state) are treated as positive, whereas both R_2 pairs (same identity but different appearance states) and R_3 pairs (different identities) are treated as negative when computing TAR@FAR and AUC. Thus, these binary verification metrics evaluate whether a model can retrieve samples from the same identity-appearance state while rejecting both different-state and different-identity samples.

Because this binary protocol does not directly measure the intermediate ordering between R_2 and R_3, we additionally report $P(R_1 > R_2)$, $P(R_1 > R_3)$, $P(R_2 > R_3)$, and $P(R_1 > R_2 > R_3)$ in Table 6. Unless otherwise specified, all results use the identity embedding z^{id} at inference; z^{head} is used only during training.

A.4 IDENTITY-PRESERVATION ABLATION

We further examine how the similarity-learning objectives affect whole-head identity verification. Starting from identity distillation only, we progressively add L_{sim}, L_{global}, and L_{patch}. All variants use the same backbone and evaluation protocol.

Table 9: Ablation of similarity-learning objectives on whole-head identity verification.

Configuration	TAR@FAR=10^{-2}	TAR@FAR=10^{-3}	TAR@FAR=10^{-4}	AUC
Distillation only	0.960	0.908	0.825	0.993
$+L_{\text{sim}}$	0.955	0.891	0.794	0.998
$+L_{\text{global}}$	0.956	0.894	0.800	0.998
$+L_{\text{patch}}$	0.954	0.890	0.793	0.994

B EXAMPLE OF USER STUDY QUESTIONNAIRE

We include the example questionnaire used in the human evaluation. The questionnaire contains 30 ranking questions. The first 15 questions are used for hierarchy validation, where the three options correspond to R_1, R_2, and R_3. The remaining 15 questions are used for retrieval preference evaluation, where the three options correspond to AdaFace, our method, and TransReID. Options are randomly shuffled during annotation to avoid positional bias.

Head Similarity Survey

Please evaluate the images based on **overall head similarity**, i.e., determine whether the head in the candidate image looks like the same person as in the reference image, and whether the overall appearance is consistent.

Please consider the following factors comprehensively: identity, head shape and contour, hairstyle and hair volume, and head-related attributes (such as glasses, hats, etc.). Even if the face is not visible, please judge based on the overall head appearance.

Please do not consider background, clothing, or image quality. Rank the candidate images from most similar to least similar compared to the reference image.

1.  Please rank the options from most similar to least similar to the reference image. [Ranking question, please fill in numbers in the brackets in order] *


[ ]Option1


[ ]Option2


[ ]Option3

19.  Please rank the three images in the options based on their similarity to the reference image, from most similar to least similar. [Ranking question, please fill in numbers in the brackets in order] *




[ ]Option1




[ ]Option2




[ ]Option3