Colors See Colors Ignore: Clothes Changing ReID with Color Disentanglement

Priyank Pathak Yogesh S. Rawat

University of Central Florida

{priyank, yogesh}@ucf.edu

https://ucf-crcv.github.io/ReID/CSCI

Abstract

*Clothes-Changing Re-Identification (CC-ReID) aims to recognize individuals across different locations and times, irrespective of clothing. Existing methods often rely on additional models or annotations to learn robust, clothing-invariant features, making them resource-intensive. In contrast, we explore the use of color—specifically foreground and background colors—as a lightweight, annotation-free proxy for mitigating appearance bias in ReID models. We propose **Colors See, Colors Ignore (CSCI)**, an RGB-only method that leverages color information directly from raw images or video frames. CSCI efficiently captures color-related appearance bias (**'Color See'**) while disentangling it from identity-relevant ReID features (**'Color Ignore'**). To achieve this, we introduce **S2A self-attention**, a novel self-attention to prevent information leak between color and identity cues within the feature space. Our analysis shows a strong correspondence between learned color embeddings and clothing attributes, validating color as an effective proxy when explicit clothing labels are unavailable. We demonstrate the effectiveness of CSCI on both image and video ReID with extensive experiments on four CC-ReID datasets. We improve baseline by Top-1 2.9% on LTCC and 5.0% on PRCC for image-based ReID, and 1.0% on CCVID and 2.5% on MeVID for video-based ReID without relying on additional supervision. Our results highlight the potential of color as a cost-effective solution for addressing appearance bias in CC-ReID. Github: **https://github.com/ppriyank/ICCV-CSCI-Person-ReID**.*

1. Introduction

Person Re-Identification (ReID) aims to recognize individuals across different cameras, times, and locations based on visual cues. While traditional ReID [39, 41] assumes consistent clothing, real-world scenarios often involve changes in attire, making Clothes-Changing ReID (CC-ReID) [19] a more challenging yet practical task. In CC-ReID, the model must distinguish identities despite drastic variations in appearance, while also handling challenges such as different individuals wearing similar clothing. Given its importance in real-world surveillance and security applications [12, 20], we focus on addressing CC-ReID in this work.

Existing CC-ReID approaches often use external biometric modalities such as gait [25], face [45], and body shape [9], but their *pre-processing cost* and reliance on additional models limit real-world applicability. RGB-only methods like CAL [19] learn clothing-invariant features but require *external clothing annotations*, adding manual overhead. Recent works [23, 33] leverage *fine-grained attributes* (*e.g.* black pants, has umbrella), but these are *computationally expensive*, *short-term* (per frame), and sensitive to *clothing modifications*, *occlusion*, and *illumination changes*, making them unreliable over time. For practical deployment, it is crucial to reduce reliance on such external attributes while effectively handling clothing variations.

Our work explores a simple, efficient color-based CC-ReID approach, eliminating the need for external annotations. By leveraging color cues, our method offers a computationally lightweight yet effective alternative for reducing appearance bias, particularly when clothing annotations are unavailable. Color information in an image possesses three key properties: **1) Efficient**: colors can be extracted in real-time with minimal overhead **2) Adaptive**: colors dynamically adjust to changes in illumination, occlusion, and environmental conditions **3) Contextual**: colors capture both clothing and background information, enabling differentiation between similar clothing in different contexts. Figure 1 illustrates the adaptive and contextual properties through clustering of images via RGB-derived color representations.

We propose **Color See, Color Ignore (CSCI)**, a transformer-based method that learns color representations (*Color See*) from raw RGB-only input while disentangling them from ReID features (*Color Ignore*). CSCI introduces a dedicated **Color token**, analogous to the class token in ViTs, to predict color embeddings, enabling explicit color representation learning. However, the *parameter-efficient* nature of this token risks color bias leakage into ReID features via self-attention. To address this, we propose a middle-ground solution between two extremes: (1) complete overlap of ReID and Color (appearance) tokens [19,

Figure 1. **Clustering of color representation**: Clusters (2 images/cluster, same outline color) generated via RGB-uv projection [1]. *(a):* Clothing changes in video are clustered differently *(b & c):* Clustering based on similar clothing & background / illumination. *(d):* Different clothing but similar overall color profile clustered together. Images from CCVID (a) and LTCC (b, c, d).

21, 35], and (2) fully separate dual-branch [33, 51] via our **S2A self-attention**, enabling ReID and Color tokens to perform independent self-attention, preventing cross-information transfer. We demonstrate CSCI's effectiveness and generalization across both image and video ReID tasks.

In summary, 1) 'CSCI' learns to ignore color representation using RGB-only inputs *without* any external annotation or modality. 2) Using the Color token and S2A self-attentions, we disentangle color bias from the ReID features. 3) CSCI outperforms existing methods on two image CC-ReID benchmarks: LTCC and PRCC, and two video benchmarks: CCVID and MeVID. 4) The ablation study provides evidence of colors' viability as a cost-efficient alternative for external annotations/modalities.

2. Related Work

CC-ReID External modalities such as gait [6, 17], face [5], body-shape [32], background [34], pose [35] *etc.* are often used to capture biometric signatures. However, these modalities have computationally expensive preprocessing, limiting their application. Conversely, RGB-only models [19, 21, 51] are less resource-intensive but rely on external clothing annotations [48]. Moreover, these attributes are usually 'short-term', *i.e.* clothing can change throughout a video, potentially requiring frame-level annotations. To address these issues, we propose learning the RGB-only color representations, that can serve as cost-free frame-level annotations for appearance bias removal without relying on external annotations. In the past, colors have mostly been used as augmentation [32, 43] rather than learnable attributes. Our approach can potentially offer a more cost-efficient CC-ReID in the absence of clothing attributes.

Transformers in ReID TransReID [22], and follow-ups [8, 10] have used ViT class token for ReID. Other early works have combined traditional ReID with transformers, *e.g.* memory banks [38], external modalities [11], CNNs features [24] *etc.*). Recent methods [7, 23, 31, 33, 47] incorporate text-based fine-grained descriptions into the ReID features. Ye et al. surveyed the growing popularity of transformers in ReID, with state-of-the-art largely dominated by transformers. With the growing complexity of transformers, the added cost of external attributes could become a bottleneck. Thus our design primarily focuses on transformer-centric solutions, via a computationally inexpensive learnable token to learn and disentangle color-related features.

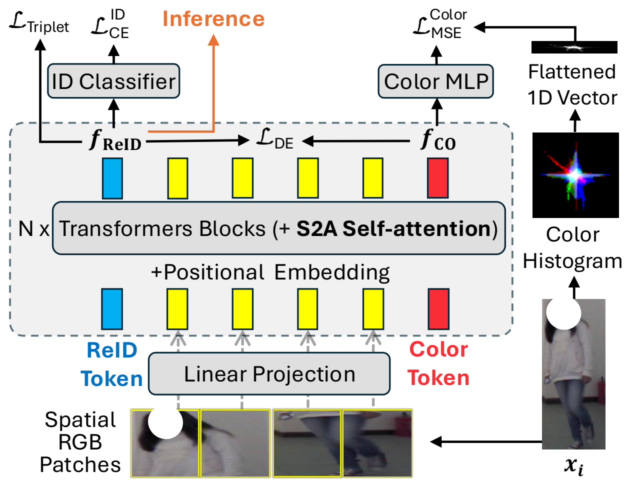
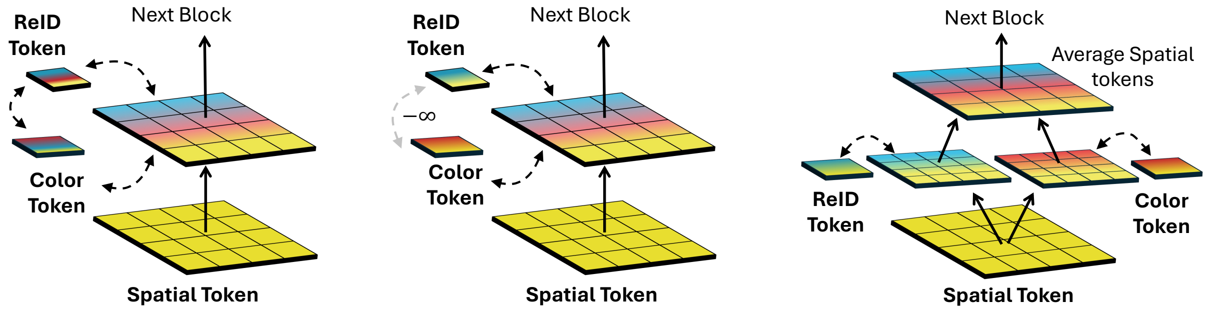


Figure 2. **CSCI Framework**: Image spatial tokens (yellow) are appended with ReID (f_{ReID}), and Color (f_{CO}) tokens. Tokens are passed through transformer blocks (modified self-attention), discarding spatial ones after the last block. f_{ReID} (inference) predicts the identity labels and trains with triplet loss. f_{CO} predicts flattened color histograms and disentangled from f_{ReID}.

3. Methodology

Color See, Color Ignore (CSCI), learns to generate color-agnostic CC-ReID features in a parameter-efficient manner. Given a training set of N data points $(x_i, y_i)_{i=1}^{N}$, where x_i are RGB images and y_i are identity labels, **CSCI does not rely on any external or internal (clothing) annotations**. During both training and testing, only RGB images x_i are provided as input, producing ReID features f_{REID} and color embeddings f_{CO} (Fig. 2), while only f_{REID} is used for inference. Inference involves matching query f_{REID} against images/videos in the gallery. A dedicated **Color token** learns color embeddings, mitigating the color-induced appearance bias. We employ a two-step self-attention mechanism, **S2A Self-Attention**, to prevent information leak between ReID and color features. In this section, we describe the each component of CSCI.



(a) Tradational Self-Attention (b) Masked Self-Attention (c) S2A Self-Attention

Figure 3. **Self-Attention Variants**: Dotted arrows and shades of coloring indicate information exchange via self-attention. For explanation, spatial tokens are grouped as one, skipping their internal self-attention. *(a) Traditional self-attention*, all tokens attend each other; *(b) Masked self-attention* is the same as traditional without *ReID-Color* token attention; *(c) S2A self-attention* has two independent self-attention steps, one for ReID and the other for Color token. Spatial tokens are averaged across the two steps.

3.1. Color See Color Ignore (CSCI)

The idea behind our proposed method *'Color See Color Ignore'* is to learn color representation from RGB-only inputs (*'See Color'*) and using the learned color representation, we disentangle it from the ReID features (*'Color Ignore'*). An overview of the proposed method is shown in Figure 2.

3.1.1. Image Encoder

We propose a transformers-based approach where the input image is first converted into non-overlapping spatial tokens (yellow). Two identical learnable head tokens, ReID token (popularly called 'class token') (blue), and Color token (red) are appended to these spatial tokens. All tokens along with their positional embedding are passed through a series of transformer blocks (w/ modified self-attention). ReID and Color tokens derive their respective embeddings from the context of spatial tokens.

After the last block, spatial tokens are discarded, ReID token serves as ReID features f_{ReID}, while the Color token represents color embeddings f_{CO}. An ID classifier head is added on top of f_{REID} for classifying the identity labels y_i of the individuals in the training set. On the other hand, f_{CO} learns color embedding by predicting/regressing the color representation vector via a fully connected MLP layer. Predicting color vectors removes test time dependency on these representations. The cosine distance between f_{REID} and f_{CO} is maximized (inducing orthogonality [28]) for removing color bias from ReID features.

$$\mathcal{L}_{DE}(f_{CO}, f_{REID}) = \left| \frac{f_{CO}}{\|f_{CO}\|_2} \cdot \frac{f_{REID}}{\|f_{REID}\|_2} \right| \quad (1)$$

This additional head token-based approach [44] is parameter-efficient and can help generate color embeddings in the same feature space as that of the ReID token. We adapt EVA-02 [18] vision transformer for image encoder in our approach. In addition, we also show its generalization to other existing transformer-based approaches for ReID.

3.1.2. S2A Self-Attention

In *Traditional Self-Attention* (100% overlap of color and ReID, Fig. 3a), the ReID (class) token attends (exchanges information) with every other token, and vice versa. It's commonly depicted for one head self-attention without token dimension as, $Att(Q, K, V) = softmax(Q.K^T)V$ where Q, K, and V are Query, Key, and Value representations. The weights assigned to each token are shown as

$$QK^T = \begin{array}{c|ccc} & k_{CO} & k_{ID} & k_{SP} \\ \hline q_{CO} & q_{CO} \cdot k_{CO} & q_{CO} \cdot k_{ID} & q_{CO} \cdot k_{SP} \\ q_{ID} & q_{ID} \cdot k_{CO} & q_{ID} \cdot k_{ID} & q_{ID} \cdot k_{SP} \\ q_{SP} & q_{SP} \cdot k_{CO} & q_{SP} \cdot k_{ID} & q_{SP} \cdot k_{SP} \end{array} \quad (2)$$

Here, the Color token (CO) exchanges information directly with the ReID token (ID) and spatial token(s) (grouped as SP), risking entangling color bias into ReID features. A simple solution is to use *Masked Self-attention* [15], where tokens are masked with $-\infty$ weights (0 after softmax, Fig. 3b). When computing self-attention for the ReID token, Color token is assigned $-\infty$ weight, and similarly Color token assigns $-\infty$ weight to the ReID token.

$$QK^T_{\text{Mask}} = \begin{array}{c|ccc} & k_{CO} & k_{ID} & k_{SP} \\ \hline q_{CO} & q_{CO} \cdot k_{CO} & -\infty & q_{CO} \cdot k_{SP} \\ q_{ID} & -\infty & q_{ID} \cdot k_{ID} & q_{ID} \cdot k_{SP} \\ q_{SP} & q_{SP} \cdot k_{CO} & q_{SP} \cdot k_{ID} & q_{SP} \cdot k_{SP} \end{array} \quad (3)$$

While this reduces the direct information leak between the Color and ReID tokens, spatial tokens however are exposed to both ReID and Color tokens. Both tokens influence each other's weight (Eq. (3) last row) on final spatial (context) embedding. We believe that the ReID token, representing biometrics, is as important as the appearance bias represented by the Color token. Therefore, to give equal weight to both tokens, we separate the self-attention step entirely for ReID and Color tokens. We name this ***S2A Self-Attention*** (Fig. 3c). Mathematically, it can be represented as two self-attention steps, one without the Color token ($\sim CO$), Eq. (4)), and one without the ReID token ($\sim ID$), Eq. (5)), and averaging for Spatial Tokens Eq. (6)).

$$QK^T_{\sim CO} = \begin{array}{c|cc} & k_{ID} & k_{SP} \\ \hline q_{ID} & q_{ID}\cdot k_{ID} & q_{ID}\cdot k_{SP} \\ q_{SP} & q_{SP}\cdot k_{ID} & q_{SP}\cdot k_{SP} \end{array} \quad (4)$$

$$QK^T_{\sim ID} = \begin{array}{c|cc} & k_{CO} & k_{SP} \\ \hline q_{CO} & q_{CO}\cdot k_{CO} & q_{CO}\cdot k_{SP} \\ q_{SP} & q_{SP}\cdot k_{CO} & q_{SP}\cdot k_{SP} \end{array} \quad (5)$$

$$Att(Q^{SP}, K^{SP}, V^{SP}) = \Big(Att(Q^{SP}_{\sim ID}, K^{SP}_{\sim ID}, V^{SP}_{\sim ID}) + Att(Q^{SP}_{\sim CO}, K^{SP}_{\sim CO}, V^{SP}_{\sim CO})\Big)/2 \quad (6)$$

These two self-attention(s) add minimal computational overhead while outperforming its counterparts (Tab. 4a). An alternative would be to use dual-branch (one for color, and the other for ReID) [40, 51] to completely nullify the overlap of these tokens, *computationally impractical*. S2A with some leak between ReID and color token offers a middle ground between complete (Eq. (2)) and no overlap.

3.2. Color Representation

Colors from images/video frames can provide an *adaptive* (current state *e.g.* illumination) and *contextual* (clothes and background colors) profile of the input (Fig. 1). Here, we explore two different color representations: *'Pixel Binning (Pix Bin)'* and *'RGB-uv projection (RGB-uv)'* [1]. For a given bin size h, *Pix Bin* creates a 3D histogram based on pixel intensity/values (0-256) and counts the number of pixels in each bin. This generates a $h \times h \times h$ 3D histogram. On the other hand, RGB-uv generates 2D histogram projections ($h \times h$) for each R,G,B channel, via a set of differential operations (*Supplementary*)[1]. This generates a $3 \times h \times h$ histogram. RGB-uv projections have been shown to effectively style transfer in GANs [2]. Any of these methods can be used to represent colors (Fig. 4), and by flattening these histograms into 1D vectors we can represent these histograms as ground truth color representations. Both, 'Pix Bin' and 'RGB-uv' can be computed efficiently on the fly from RGB images *i.e.* without any external dependency.

3.3. Learning Objective

EVA-02 (f_{REID}) is fine-tuned on target ReID like traditional ReID [22], with standard cross-entropy loss ($\mathcal{L}^{ID}_{CE}$) on ID logits for predicting the identities of individuals in the training set, and Triplet loss ($\mathcal{L}_{Tripelt}$) on f_{ReID}. For predicting the color representation vector, we use MSE regression loss ($\mathcal{L}^{Color}_{MSE}$) on f_{CO}. $\mathcal{L}_{DE}$ (Eq. (1)) disentangles f_{ReID} with f_{CO}. Our overall training objective L_{ReID} is defined as

$$L_{ReID} = \mathcal{L}^{ID}_{CE} + \mathcal{L}^{Color}_{MSE} + \mathcal{L}_{Triplet} + \mathcal{L}_{DE} \quad (7)$$

Details of loss functions are in *Supplementary*.

[1]Code: https://github.com/mahmoudnafifi/HistoGAN

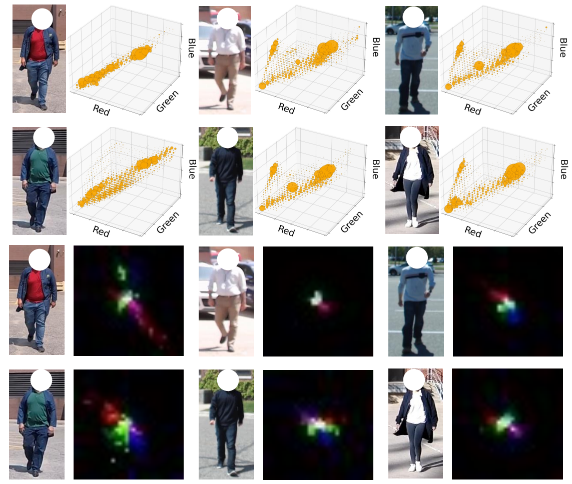


Figure 4. **Color Representation**: *Top Pixel Binning:* 3D Histogram ($20\times20\times20$); *Bottom:RGB-uv 2D Projection/Histogram* ($3\times32\times32$, upscaled for viewability. Images taken from CCVID.

3.4. Generalization to Video ReID

We generalize CSCI to Video ReID by integrating our Color token with a lightweight video adaptation strategy. Specifically, we use EZ-CLIP [3, 4] technique, which efficiently extends image-based transformers to videos using temporal tokens (CSCI-V, detailed architecture in *Supplementary*). Following its design, we first pre-train the image model as a standard Image ReID model (w/o the Color token) using randomly sampled frames from video clips. We then fine-tune the model for video sequences by incorporating temporal tokens while training in a **two-step alternating process**: **1) Vanilla Video ReID**: The model processes video frames using temporal tokens, similar to standard Video ReID, w/o the Color token. **2) CSCI Image ReID**: The temporal tokens are disabled, effectively converting the model back into an image-based transformer. The Color token is then used to disentangle color information, training on randomly sampled video frames. *Note:* Temporal and Color tokens are not trained simultaneously, as temporal tokens have their own self-attention mechanism. Combining this with our two-step S2A self-attention would introduce multiple levels of self-attention in each transformer block, significantly increasing computational overhead.

4. Experiments

4.1. Experiment Settings

Dataset To show the effectiveness of our method we use 2 *Image CC-ReID* datasets: (i) **LTCC** [42]: 77 training identities (IDs) (9,576 images) & 75 test IDs (7,543 images). (ii) **PRCC** [50]: 150 training IDs (17,896 images) & 71 test IDs (10,800 images). For the *video CC-ReID* dataset, we choose (iii) **CCVID** [19] has 75 training IDs (948 clips)

	Method	Venue	Additional Attributes		LTCC				PRCC			
					CC		General		CC		SC	
			Int.	External	R-1↑	mAP↑	R-1↑	mAP↑	R-1↑	mAP↑	R-1↑	mAP↑
ResNet	CAL [19]	CVPR'22	CL	-	40.1	18.0	74.2	40.8	55.2	55.8	**100**	99.8
	CCFA [21]	CVPR'23	CL	-	45.3	22.1	75.8	42.5	61.2	58.4	99.6	98.7
	FIRe^2 [48]	TIFS'24	-	-	44.6	19.1	75.9	39.9	65.0	63.1	**100**	99.5
Multi-ResNets	3DInvar. [32]	ICCV'23	CL	Po+BS	40.9	18.9	-	-	56.5	57.2	-	-
	AIM [51]	CVPR'23	CL	-	40.6	19.1	76.3	41.1	57.9	58.3	**100**	**99.9**
	DCR. [13]	TCSVT'23	CL	BP+Co.	41.1	20.4	76.1	42.3	57.2	57.4	**100**	99.7
	CVSL [35]	WACV'24	CL	Po	44.5	21.3	76.4	41.9	57.5	56.9	97.5	99.1
	CCPG [36]	CVPR'24	CL	Po	46.2	22.9	77.2	42.9	61.8	58.3	**100**	99.6
	RLQ [40]	BMVC'25	CL	Po+Gender+BP	46.4	21.5	76.9	41.8	64.0	**63.2**	**100**	99.8
Transformer	CLIP3D [33]	CVPR'24	-	BS+Text	42.1	21.7	-	-	60.6	59.3	-	-
	IRM [23]†	CVPR'24	CL	Text+BP	46.7	-	66.7	-	-	-	-	-
	EVA-02		-	-	44.9	23.1	80.3	45.9	61.6	59.0	**100**	**99.9**
	CSCI - Pix. Bin *(our)*		-	-	46.7$^{+1.8}$	23.6$^{+0.5}$	80.8$^{+0.5}$	45.9	**66.6**$^{+5.0}$	60.7$^{+1.7}$	**100**	**99.9**
	CSCI - RGB-uv *(our)*		-	-	**47.8**$^{+2.9}$	**24.4**$^{+1.3}$	82.6$^{+2.3}$	48.0$^{+2.1}$	66.2$^{+4.6}$	61.3$^{+2.3}$	**100**	**99.9**

Table 1. **LTCC & PRCC Results (%)**. Shorthand notations are: **Best results** and second best results, 'Int.': internal ('CL' clothing labels), 'Po': Pose/skeleton, 'Co': Contours, 'BS': 3D Body shape, 'BP': Body parse silhouettes, '-': results/attributes not reported/used, ↑: Higher means better performance, †, IRM multi-dataset training results skipped. **The improvement X% over the baseline is indicated by** $^{+X}$.

	Method	Venue	Additional Attributes		CCVID				MeVID			
					CC		General		Overall (General)			
			Int.	External	R-1↑	mAP↑	R-1↑	mAP↑	R-1↑	R-5↑	R-10↑	mAP↑
	CAL [19]	CVPR'22	CL	-	81.7	79.6	82.6	81.3	52.5	66.5	73.7	27.1
M-CNN	3DInvar. [32]	ICCV'23	CL	Po+BS	84.3	81.3	83.9	82.6	-	-	-	-
	SEMI [37]	WACV'24	-	BS	82.5	81.9	83.1	81.8	-	-	-	-
	ShARc [53]	WACV'24	-	Sil+BS+Po	84.7	85.2	89.8	90.2	59.5	70.3	77.2	29.6
Transformer	CLIP3D [33]	CVPR'24		BS+Text	82.4	83.2	84.5	83.9	-	-	-	-
	GBO [27]	ArXiv'24	CL	BP+Po+Diff	86.9	83.5	89.7	87.1	-	-	-	-
	EVA-02 Image		-	-	86.4	87.0	88.8	89.0	73.5	83.7	86.2	47.5
	EVA-02 + EZ-CLIP [3]		-	-	89.8	90.1	91.1	91.1	76.6	84.8	87.0	55.5
	CSCI-V - Pix. Bin *(our)*		-	-	90.5$^{+0.7}$	90.8$^{+0.7}$	91.5$^{+0.4}$	91.6$^{+0.5}$	**79.1**$^{+2.5}$	**87.5**$^{+2.7}$	**89.1**$^{+2.1}$	**56.9**$^{+1.4}$
	CSCI-V - RGB-uv *(our)*		-	-	**90.8**$^{+1.0}$	**91.3**$^{+1.2}$	**91.7**$^{+0.6}$	**92.2**$^{+1.1}$	**79.1**$^{+2.5}$	87.2$^{+2.4}$	88.6$^{+1.6}$	55.6$^{+0.1}$

Table 2. **CCVID & MEVID Results (%)**. Same shorthand notation as Tab. 1. Additional attributes include 'Diff.' for the Diffusion model and 'Sil' for silhouettes. The baseline here is the Video ReID 'EVA-02 + EZ-CLIP'. M-CNN refers to Multi-CNN architecture.

Model	Venue	ViT Backbone	Attributes Int.	Ext.	LTCC (CC) R-1↑	mAP↑
TransReID[22]	ICCV'22	ViT-B (86 M)	-	-	31.1	15.1
+CSCI	*(ours)*		-	-	36.0	15.5
TMGF [26]	WACV'23	ViT-S (22 M)	-	-	32.9	17.5
+CSCI	*(ours)*		-	-	39.5	20.9
PAT [38]	ICCV'23	ViT-B (86 M)	-	-	31.1	13.9
+CSCI	*(ours)*		-	-	32.7	13.8
TCiP [49]	CGI'22	ViT-B (86 M)	-	Pose	35.2	19.0
			CL	Pose	36.5	19.0
+CSCI	*(ours)*		-	Pose	38.8	19.7
Baseline		EVA-02 (303 M)	-		44.9	23.6
			CL		46.3	24.0
+CSCI	*(ours)*		-		47.8	24.4

Table 3. **CSCI on other Transformers**. Only TCiP is CC-ReID

with 151 test IDs (1908 clips) and (iv) **MeVID** [14] has 104 training IDs (6338 clips) with 54 test IDs (1754 clips).

Evaluation We report Cumulative Matching Characteristics (CMC, rank 1(R-1) and rank 5 (R-5), *etc.*) and Mean Average Precision (mAP) as an average of two runs. Evaluation for each dataset: LTCC and CCVID both use the Cloth-Changing (CC) (Query-Gallery have different clothes) and the General Protocol (entire test set). MeVID only uses the General Protocol. Other protocols for MeVID in *Supplementary*. PRCC uses the CC and the Same-Clothes (SC) protocol (Query-Gallery wears the same clothes).

4.2. Implementation

The input resolution of all input is 224×224 with batch size 8 and we use n=4 number of frames for videos. We use ViT-L backbone of EVA-02 with its official pre-trained weights [16, 18], architectural details in *Supplementary*. Generalization to other transformers is in Section 4.3 (Tab. 3). The

Method (Self-Attention)	# Param ↓ (M)	#FLOP↓ (G)	LTCC		PRCC	
			R-1	mAP	R-1	mAP
Baseline	303.47	81.18	44.9	23.1	61.6	59.0
Traditional	311.87	81.52	46.8	23.5	63.5	60.3
Masked	311.87	81.52	46.7	23.7	65.3	61.3
S2A *(our)*	311.87	84.74	**47.8**	**24.4**	**66.2**	**61.3**

(a) **Self-Attention Variations**: 'M' means in million, 'G' is giga FLOPs.

Method	LTCC	
	R-1	mAP
Baseline	44.9	23.1
W/ Clothes	46.3	24.0
Grey Aug.	38.3	18.5
Colors *(Our)*	**47.8**	**24.4**

(b) **Attributes Variation**.

Method	PRCC	
	R-1	mAP
Baseline	61.6	59.0
Feed	63.1	57.6
Pred. (Cos)	65.0	60.9
Pred. (MSE)	**66.2**	**61.3**

(c) **Colors Design Variation**.

Table 4. **Design Choices**: All evaluations done on cloth changing (CC) protocol. # of Params & GFLOPs calculated for PRCC dataset with RGB-uv color $32 \times 32 = 1024$-dim projections (averaged across channels).

image encoder uses absolute position embeddings, with 24 blocks, patch size 14, 16 multi-head self-attention and 1024 dimensional tokens. Color and ReID tokens (and their positional embeddings) are initialized with pre-trained class token weights. The image and video models are trained on two 16GB and 32GB Tesla V100 GPUs, respectively. Pixel Binning is computed using OpenCV 'cv2.calcHist' method, with bin size 20, creating a flattened $20 \times 20 \times 20 = 8000$-Dim color vector. For RGB-uv projections, hyperparameters were determined empirically (*Supplementary* for all color-related hyperparameters). RGB-uv projections are averaged across 3 channels, except for LTCC where it is concatenated. Images and videos use bin sizes of 32 and 16.

4.3. Comparison with SOTA Methods

Additional attributes used by previous approaches are either *'internal'* like clothing labels (via dataset) or *'external'*, generated from off-the-shelf models. We have skipped the results for model agnostic technique-based methods like ReFace [5], CCVReID [46], and DCGN [29] which improve the accuracy of existing methods with facial features. *Supplementary* contains all results.

Image ReID (LTCC & PRCC) Table 1 shows CSCI with color disengagement outperforms existing methods on both LTCC and PRCC despite not using any additional attributes. FIRe^2 [48] is RGB-only approach (unsupervised clustering for pseudo attributes) with comparable performance on PRCC. CNN models (w/ additional attributes) mostly fall short of EVA-02 baseline, reinforcing transformers superiority. Other transformer-based approaches (CLIP3D [33], IRM [23]) using large-scale Image-Text models, similar to our EVA-02. Color disentanglement improvement over image baseline (Top-1) using RGB-uv projection (*2.9%-2.3% on LTCC* and 4.6% on PRCC) outperforms pixel binning (1.8%-0.5% on LTCC and *5% on PRCC*).

Video ReID (CCVID & MeVID) Table 2 shows video adaptation of image baseline using EZ-CLIP significantly improves the performance of the vanilla Image ReID model. Video adaptation with color disentanglement (CSCI-V) outperforms existing methods for CCVID and MeVID. For videos, RGB-uv projection (Top-1 gain with *1.0%-0.6%* for CCVID and 2.5% on MeVID) in general has the same improvement over baseline (in all categories) as that of pixel binning (Top-1 gain with *0.7%-0.4%* for CCVID and *2.5%* for MeVID. The EVA-02 baseline significantly outperforms prior MeVID results, likely due to strong alignment between real-world MeVID data and the model's pre-trained weights. Other foundation model-based GBO [27] uses diffusion to generate data augmentations, in addition to other attributes like pose, clothes, and body parsing; it has comparable results to ours. *Efficient video adaptation using EZ-CLIP [3, 4] incorporates image ReID achievements into video ReID, offering a viable future research direction.*

Method	R-1↑	mAP↑
Baseline	44.9	23.1
(Directly) Colors Vectors (1024-dim)	45.7	21.4
(Directly) Project ReID to Colors (3072-dim)	44.9	23.8
Using Color tokens (3072-dim.) (Our)	47.8	24.4

Table 5. **Learning vs directly using colors**. LTCC results.

Generalization of CSCI To demonstrate the generalization and fairness of our method, CSCI (Color token and S2A self-attention), we applied colors to other transformer backbones, specifically ViT-small (TMGF [26], 22M parameters) and ViT-base (TransReID [22], TCiP [49], and PAT [38], 86M parameters), as shown in Table 3. These backbones are significantly lighter than our backbone, which primarily uses a ViT-Large. TransReID, PAT, and TMGF are the same clothes ReID models; hence, CSCI mainly reflects performance enhancement, particularly for TMGF, with a 7% Top-1 improvement. For the CC-ReID model TCiP, color outperforms clothing labels.

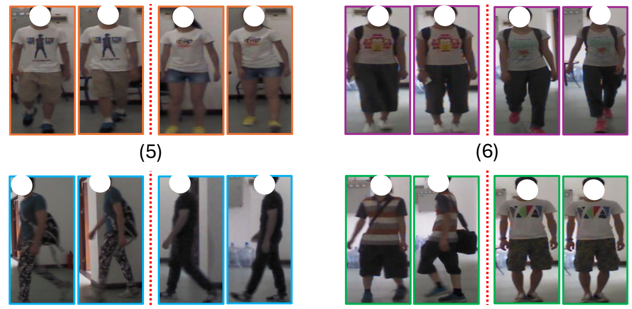
5. Ablation Studies & Analysis

All the analyses below pertain specifically to PRCC and LTCC under the CC protocol. Experimental details and hyperparameters of the following are in *Supplementary*.

5.1. Design Choices & Cost Analysis

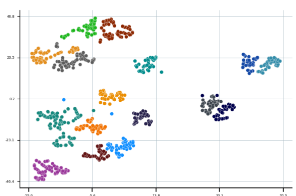
Table 4a demonstrates that our proposed S2A Self-Attention (Fig. 3c) outperforms the other two variants: Traditional (Fig. 3a) and Masked Self-Attention (Fig. 3b). The relatively lower performance of traditional self-attention on

(a) **Same** cloth label **Different** k-means cluster of Color token

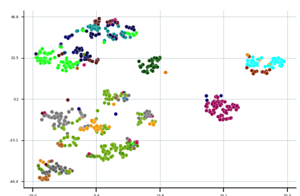


(b) **Different** cloth label **Same** k-means cluster of Color token

Figure 5. **Clustering of Color token**: Images from PRCC gallery set. (a) Two images per cluster (same outline colors), indicating color embedding clusters based on background & illumination in case of the same clothing. (b) Cluster include similar-looking different clothes (5,6) (ideal case) and similar-looking individuals (7,8) possibly indicating *i.e.* ReID information leak in color embeddings (error case).



(a) Clothes Labels coloring



(b) f_{CO} K-means (k=24) coloring

Figure 6. **Color token embeddings (**f_{CO}**) t-SNE**: t-SNE clusters of f_{CO} can be separated into (a) clothes labels *(left)* indicating a correlation between learned f_{CO} and clothes labels. (b) K-means clusters of f_{CO} differ in clothing *i.e.* different clothing may share similar colors. For example, 'green' cluster on *(b) right* have different clothing, blue, orange, brown on *(a) left* (PRCC test subset)

the PRCC dataset suggests an information leak between Color and ReID tokens. CSCI parameter overhead is *2.77%* due to the Color token, its positional embeddings, and the MLP head, validating the parameter-efficient nature of using the Color token as an additional head token. FLOP was calculated using the fvcore library. The increase in FLOP count for S2A Self-Attention is *4.39%* for 24 blocks, with a *0.18%* increment per block, due to the two-step self-attention process. Table 4b shows the additional head token-based learning for clothing labels ('w/ Clothes'), instead of color, is sub-optimal in reducing appearance bias. Grey-scaled augmentation [30] ('Grey Aug.') also performs worse, likely due to the distribution shift from RGB pretrained weights to greyscale training. Using random initialization to avoid RGB bias would force training transformers on small ReID datasets without large-scale pre-training (random initialized EVA achieves a 2.6% Top-1 accuracy on LTCC (CC)). No pertaining for transformer is not practical given model's size and complexity. Table 4c indicates that input color representations ('Feed'), instead of predicting them, result in inferior performance while also introducing a test-time dependency on colors. Additionally, MSE regression loss ($\mathcal{L}_{MSE}^{Color}$) outperforms cosine similarity for predicting color histogram vectors.

Learning color representations instead of directly using raw color vectors offers two key benefits: 1) Learned color representations are in the same feature space as ReID embeddings (shared transformer) which eases disentanglement. 2) The ReID token is 1024 dimensional, while the optimal color representation of the color vector (e.g. LTCC is 3072 dim) need not fit this 1024 dimensionality. As shown in Tab. 5 using sub-par color vectors (1024-dim.) or projecting ReID features to match the optimal color vectors (3072-dim.) performs inferior to learning color embeddings.

5.2. Color Token Embeddings

Use of colors is motivated by its ability to distinguish between similar clothing and backgrounds. This is shown in Fig. 1, where RGB-uv projections create separable clusters (averaged across 3 channels, L2 normalized, t-SNE dimensionality reduction $\in \mathbb{R}^{32\times32} \rightarrow \mathbb{R}^2$, and DBSCAN). These clusters show (a) color representation can change within the same video (adaptive), (b & c) separability based on clothing, background & illumination, and (d) matching overall color profiles. *Supplementary* has GIFs of all clusters.

Color token embedding (f_{CO}) t-SNE plot using the PRCC test subset (Fig. 6) reveals separable clusters. These clusters can easily be segregated based on clothing labels (Fig. 6a) *i.e.* there exists a **correlation between clothing labels and learned color representation**. Coloring these clusters via f_{CO} k-means clusters (Fig. 6b), reveals these clusters (clothes-segregated) share color profiles across them. Simply put, different clothing labels (clusters) share learned color representation across them.

For the PRCC gallery with 71 clothing labels, k-means clustering (k=71+10 clusters) of f_{CO} reveals two key findings: (a) **Same clothing labels but different k-means clusters:** (Fig. 5a), Similar clothes are clustered separately, likely due to variations in background and illumination. (b) **Different clothing labels but same k-means clusters:** (Fig. 5b), Despite different clothing annotations, visually

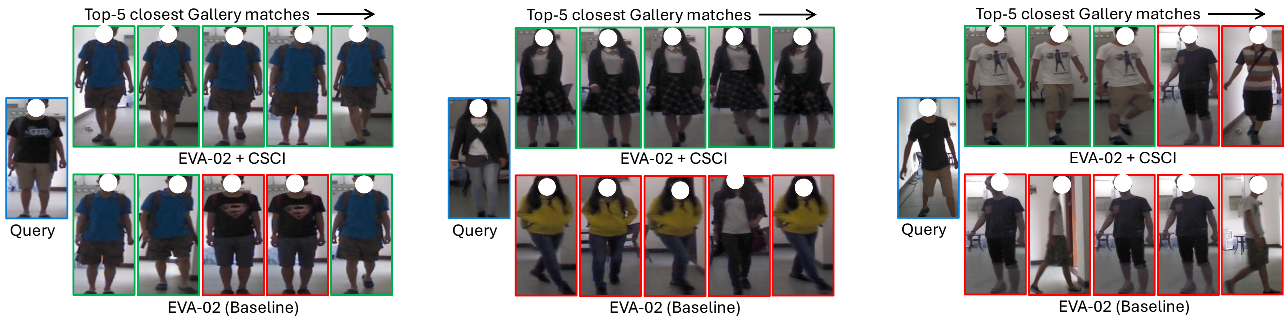
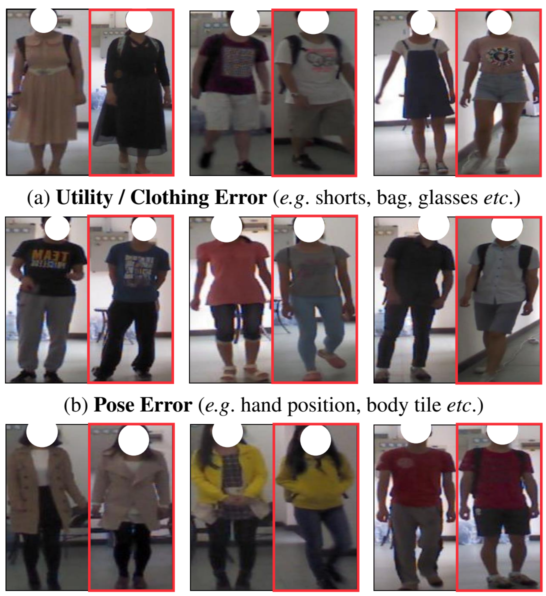


Figure 7. **CSCI vs Baseline**: Column of images are top 5 gallery matches for the query, taken from PRCC dataset **(clothes changing)**, with green (red) indicating correct (wrong) matches. *Left* & *Middle* are cases where baseline mistakes when the gallery wears similar clothing (*left* has black t-shirt, and *middle* has blue jeans). *Right* shows both models making mistakes, likely because of pose similarity (orientation/tilt of hand/body). The baseline matches two different poses likely because of similar background & illumination to the query.



(a) **Utility / Clothing Error** (*e.g.* shorts, bag, glasses *etc.*)

(b) **Pose Error** (*e.g.* hand position, body tile *etc.*)

(c) **Color Error** (*e.g.* grey coat, yellow jacket, red top *etc.*)

Figure 8. **CSCI Error analysis:** Top-1 mismatches on PRCC.

similar clothes are clustered together (5,6), indicating that f_{CO} effectively identifies similar colors. However, noisy clusters (7,8) hint at a potential information leak between f_{ReID} features and f_{CO} as similar poses (7) and faces (8) (hidden for privacy) are grouped together. This issue might be solved by completely separating color embedding from ReID features (separate branch), which involves a computational efficiency trade-off (*left for future research*).

5.3. CSCI comparison with Vanilla Image Model

The Top-5 matches for the PRCC test set were computed using our method, CSCI, and the vanilla image baseline, EVA-02, as shown in Fig. 7. For each query (blue), we display cases where our method correctly matches (green) versus incorrect matches (red) by the baseline. In Fig. 7 *(left & middle)*, the baseline makes mistakes when the gallery images share similar clothing colors with the query, such as a black t-shirt on the *left* and blue jeans in the *middle*. Our method, CSCI, avoids this pitfall by correctly matching distinctly different colored clothing. In Fig. 7 *(right)*, both our model and the baseline make mistakes, likely because of similar poses such as the positioning/orientation of the hand. However, the baseline has two cases (2$^{\text{nd}}$ & 5$^{\text{th}}$) where the mismatch is likely due to similarity in background and illumination (rather than pose).

5.4. Error Analysis

Figure 8 categorizes the Top-1 errors made by the CSCI model on the PRCC dataset into 3 types: (a) **Utility / Clothing Errors**: Model incorrectly matches two different-looking individuals who share unique accessories like glasses (hidden with the face for privacy), a bag, or rarely appearing clothing like shorts or a one-piece skirt in the dataset. (b) **Similar Pose**: When query and gallery show individuals in the same pose (*e.g.*, hand position, body tilt *etc.*), they are mistakenly matched, indicating a need for pose robustness. (c) **Color Error**: Rarely appearing colors, *e.g.* yellow or red, sometimes lead to incorrect matches, highlighting the need for improved robustness against a wide spectrum of color (clothing) appearances.

6. Conclusion

"Colors See Colors Ignore (CSCI)" offers an RGB-only solution, based on colors as a proxy for appearance bias in real-world CC-ReID without clothing labels. CSCI is simple and efficient, avoiding the complexity of external models, attributes, and annotations by learning to ignore colors through the learned color embeddings. CSCI is shown to be superior to existing methods on both Image and Video ReID. We show that learned color embeddings correlate with ground-truth clothing labels and their separability under different environmental scenarios. CSCI hopes to offer a viable, cheaper proxy to clothing attributes/annotations. Limitation & ethical concerns in *Supplementary*.

7. Acknowledgement

The authors gratefully acknowledge Steven Dick (UCF High-Performance Computing) for his invaluable assistance in resolving compute-related queries, even during late-night hours. This material is based upon work supported by the National Science Foundation (NSF) Accelerating Research Translation (ART) program under *Grant No. 2331319*. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

References

[1] Mahmoud Afifi and Michael S Brown. Sensor-independent illumination estimation for dnn models. In *British Machine Vision Conference (BMVC)*, 2019. 2, 4

[2] Mahmoud Afifi, Marcus A. Brubaker, and Michael S. Brown. HistoGAN: Controlling colors of GAN-generated and real images via color histograms. In *CVPR*, 2021. 4

[3] Shahzad Ahmad, Sukalpa Chanda, and Yogesh S Rawat. Ez-clip: Efficient zeroshot video action recognition. *arXiv preprint arXiv:2312.08010*, 2023. 4, 5, 6

[4] Shahzad Ahmad, Sukalpa Chanda, and Yogesh S Rawat. T2l: Efficient zero-shot action recognition with temporal token learning. *Transactions on Machine Learning Research*, 2025. 4, 6

[5] Daniel Arkushin, Bar Cohen, Shmuel Peleg, and Ohad Fried. Geff: improving any clothes-changing person reid model using gallery enrichment with face features. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 152–162, 2024. 2, 6

[6] Shehreen Azad and Yogesh Singh Rawat. Activity-biometrics: Person identification from daily activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 287–296, 2024. 2

[7] Shehreen Azad and Yogesh S Rawat. Disenq: Disentangling q-former for activity-biometrics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 2

[8] Vaibhav Bansal, Gian Luca Foresti, and Niki Martinel. Cloth-changing person re-identification with self-attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pages 602–610, 2022. 2

[9] Jiaxing Chen, Xinyang Jiang, Fudong Wang, Jun Zhang, Feng Zheng, Xing Sun, and Wei-Shi Zheng. Learning 3d shape feature for texture-insensitive person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8146–8155, 2021. 1

[10] Weihua Chen, Xianzhe Xu, Jian Jia, Hao Luo, Yaohua Wang, Fan Wang, Rong Jin, and Xiuyu Sun. Beyond appearance: a semantic controllable self-supervised learning framework for human-centric visual tasks. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2

[11] Ying Chen, Shixiong Xia, Jiaqi Zhao, Yong Zhou, Qiang Niu, Rui Yao, Dongjun Zhu, and Dongjingdian Liu. Rest-reid: Transformer block-based residual learning for person re-identification. *Pattern Recognition Letters*, 157:90–96, 2022. 2

[12] David Cornett, Joel Brogan, Nell Barber, Deniz Aykac, Seth Baird, Nicholas Burchfield, Carl Dukes, Andrew Duncan, Regina Ferrell, Jim Goddard, et al. Expanding accurate person recognition to new altitudes and ranges: The briar dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 593–602, 2023. 1

[13] Zhenyu Cui, Jiahuan Zhou, Yuxin Peng, Shiliang Zhang, and Yaowei Wang. Dcr-reid: Deep component reconstruction for cloth-changing person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(8):4415–4428, 2023. 5

[14] Daniel Davila, Dawei Du, Bryon Lewis, Christopher Funk, Joseph Van Pelt, Roderic Collins, Kellie Corona, Matt Brown, Scott McCloskey, Anthony Hoogs, and Brian Clipp. Mevid: Multi-view extended videos with identities for video person re-identification. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023. 5

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, 2019. 3

[16] 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. 5

[17] Chao Fan, Junhao Liang, Chuanfu Shen, Saihui Hou, Yongzhen Huang, and Shiqi Yu. Opengait: Revisiting gait recognition towards better practicality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9707–9716, 2023. 2

[18] Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. *Image and Vision Computing*, page 105171, 2024. 3, 5

[19] 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*, pages 1060–1069, 2022. 1, 2, 4, 5

[20] Ayush Gupta and Rama Chellappa. You can run but not hide: Improving gait recognition with intrinsic occlusion type awareness. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5893–5902, 2024. 1

[21] Ke Han, Shaogang Gong, Yan Huang, Liang Wang, and Tieniu Tan. Clothing-change feature augmentation for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22066–22075, 2023. 2, 5

[22] 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*, pages 15013–15022, 2021. 2, 4, 5, 6

[23] Weizhen He, Yiheng Deng, Shixiang Tang, Qihao Chen, Qingsong Xie, Yizhou Wang, Lei Bai, Feng Zhu, Rui Zhao, Wanli Ouyang, Donglian Qi, and Yunfeng Yan. Instruct-reid: A multi-purpose person re-identification task with instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17521–17531, 2024. 1, 2, 5, 6

[24] Mengxi Jia, Xinhua Cheng, Shijian Lu, and Jian Zhang. Learning disentangled representation implicitly via transformer for occluded person re-identification. *IEEE Transactions on Multimedia*, 25:1294–1305, 2022. 2

[25] Xin Jin, Tianyu He, Kecheng Zheng, Zhiheng Yin, Xu Shen, Zhen Huang, Ruoyu Feng, Jianqiang Huang, Zhibo Chen, and Xian-Sheng Hua. Cloth-changing person re-identification from a single image with gait prediction and regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14278–14287, 2022. 1

[26] Jiachen Li, Menglin Wang, and Xiaojin Gong. Transformer based multi-grained features for unsupervised person re-identification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pages 42–50, 2023. 5, 6

[27] Junjie Li, Guanshuo Wang, Fufu Yu, Yichao Yan, Qiong Jia, Shouhong Ding, Xingdong Sheng, Yunhui Liu, and Xiaokang Yang. Rethinking clothes changing person reid: Conflicts, synthesis, and optimization. *arXiv preprint arXiv:2404.12611*, 2024. 5, 6

[28] Wen Li, Cheng Zou, Meng Wang, Furong Xu, Jianan Zhao, Ruobing Zheng, Yuan Cheng, and Wei Chu. Dc-former: Diverse and compact transformer for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1415–1423, 2023. 3

[29] Yichen Li, Yufei Yin, Wengang Zhou, and Houqiang Li. Refining video-based person re-identification: An integrated framework with facial and body cues. In *Proceedings of the 1st ICMR Workshop on Multimedia Object Re-Identification*, New York, NY, USA, 2024. Association for Computing Machinery. 6

[30] Yu-Jhe Li, Xinshuo Weng, and Kris M. Kitani. Learning shape representations for person re-identification under clothing change. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2432–2441, 2021. 7

[31] Xin Liang and Yogesh S Rawat. Differ: Disentangling identity features via semantic cues for clothes-changing person re-id. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 13980–13989, 2025. 2

[32] Feng Liu, Minchul Kim, ZiAng Gu, Anil Jain, and Xiaoming Liu. Learning clothing and pose invariant 3d shape representation for long-term person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19617–19626, 2023. 2, 5

[33] Feng Liu, Minchul Kim, Zhiyuan Ren, and Xiaoming Liu. Distilling clip with dual guidance for learning discriminative human body shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 256–266, 2024. 1, 2, 5, 6

[34] Jingyi Mu, Yong Li, Jun Li, and Jian Yang. Learning clothes-irrelevant cues for clothes-changing person re-identification. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*, page 337. BMVA Press, 2022. 2

[35] Vuong D. Nguyen, Khadija Khaldi, Dung Nguyen, Pranav Mantini, and Shishir Shah. Contrastive viewpoint-aware shape learning for long-term person re-identification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1041–1049, 2024. 2, 5

[36] Vuong D. Nguyen, Pranav Mantini, and Shishir K. Shah. Contrastive clothing and pose generation for cloth-changing person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 7541–7549, 2024. 5

[37] Vuong D Nguyen, Pranav Mantini, and Shishir K Shah. Temporal 3d shape modeling for video-based cloth-changing person re-identification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 173–182, 2024. 5

[38] Hao Ni, Yuke Li, Lianli Gao, Heng Tao Shen, and Jingkuan Song. Part-aware transformer for generalizable person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11280–11289, 2023. 2, 5, 6

[39] Priyank Pathak. Fine-grained re-identification. *arXiv preprint arXiv:2011.13475*, 2020. 1

[40] Priyank Pathak and Yogesh S Rawat. Coarse attribute prediction with task agnostic distillation for real world clothes changing reid. In *36th British Machine Vision Conference 2025, BMVC 2025, Sheffield, UK, November 24-27, 2025*. BMVA Press, 2025. 4, 5

[41] Priyank Pathak, Amir Erfan Eshratifar, and Michael Gormish. Video person re-id: Fantastic techniques and where to find them (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13893–13894, 2020. 1

[42] 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. *arXiv preprint arXiv:2005.12633*, 2020. 4

[43] Xiujun Shu, Ge Li, Xiao Wang, Weijian Ruan, and Qi Tian. Semantic-guided pixel sampling for cloth-changing person re-identification. *IEEE Signal Processing Letters*, 28:1365–1369, 2021. 2

[44] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021. 3

[45] Fangbin Wan, Yang Wu, Xuelin Qian, Yixiong Chen, and Yanwei Fu. When person re-identification meets changing clothes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2020. 1

[46] Likai Wang, Xiangqun Zhang, Ruize Han, Jialin Yang, Xiaoyu Li, Wei Feng, and Song Wang. A benchmark of video-based clothes-changing person re-identification. *arXiv preprint arXiv:2211.11165*, 2022. 6

[47] Qizao Wang, Xuelin Qian, Bin Li, Yanwei Fu, and Xiangyang Xue. Image-text-image knowledge transferring for lifelong person re-identification with hybrid clothing states. *arXiv preprint arXiv:2405.16600*, 2024. 2

[48] Qizao Wang, Xuelin Qian, Bin Li, Xiangyang Xue, and Yanwei Fu. Exploring fine-grained representation and recomposition for cloth-changing person re-identification. *IEEE Transactions on Information Forensics and Security*, 2024. 2, 5, 6

[49] Zepeng Wang, Xinghao Jiang, Ke Xu, and Tanfeng Sun. A transformer-based cloth-irrelevant patches feature extracting method for long-term cloth-changing person re-identification. In *Advances in Computer Graphics: 39th Computer Graphics International Conference, CGI 2022, Virtual Event, September 12–16, 2022, Proceedings*, pages 278–289. Springer, 2023. 5, 6

[50] 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. 4

[51] Zhengwei Yang, Meng Lin, Xian Zhong, Yu Wu, and Zheng Wang. Good is bad: Causality inspired cloth-debiasing for cloth-changing person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1472–1481, 2023. 2, 4, 5

[52] Mang Ye, Shuoyi Chen, Chenyue Li, Wei-Shi Zheng, David Crandall, and Bo Du. Transformer for object re-identification: A survey. *arXiv preprint arXiv:2401.06960*, 2024. 2

[53] Haidong Zhu, Wanrong Zheng, Zhaoheng Zheng, and Ram Nevatia. Sharc: Shape and appearance recognition for person identification in-the-wild. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6290–6300, 2024. 5