CrossViT-ReID: Cross-Attention Vision Transformer for Occluded Cloth-Changing Person Re-Identification

Vuong D. Nguyen, Pranav Mantini, and Shishir K. Shah

Quantitative Imaging Lab, Dept. of Computer Science, University of Houston
dnguy222@cougarnet.uh.edu

Abstract. Real-world Person Re-Identification (Re-ID) presents severe challenges like occlusions and clothing changes, making traditional Re-ID methods fail. Existing occluded Re-ID methods struggle with cloth-changing scenarios, while current cloth-changing Re-ID methods do not explicitly address occlusions. To this end, we propose CrossViT-ReID, the first framework for the challenging yet practical Occluded Cloth-Changing Person Re-ID task. We perform occlusion synthesis to expose the model to real-world occlusion variations, and capture cloth-invariant body shape modality from silhouettes. The key to success of CrossViT-ReID lies in our novel **cross-modality collaborative training** strategy which is capable of *mining the complementary relationship between appearance and shape* adaptively under occlusions, clothing changes, or bad lighting conditions. Specifically, we devise two identical ViT-based branches. One branch takes in *holistic appearance* and *occluded shape*, aiming to *focus on appearance* when *shape is noisy*. Meanwhile, *occluded appearance* and *holistic shape* are inputs to the other branch, aiming to *attend to shape* when *appearance is partly unobservable*. Cross-attention fusion then makes the two branches exchange beneficial information and complement each other. After being trained, our framework is able to amplify the most informative cues when facing ambiguity caused by in-the-wild Re-ID challenges, thus significantly enhancing Re-ID accuracy. Extensive experiments demonstrate the superiority of CrossViT-ReID on both cloth-changing Re-ID and occluded Re-ID datasets.

Keywords: Occluded Cloth-Changing Person Re-Identification · Self-attention · Cross-attention · Vision Transformer

1 Introduction

Person Re-Identification (Re-ID) involves matching the same person in a non-overlapping camera system. This is a critical computer vision task and has a wide range of applications in surveillance and public safety. Since the emergence of deep learning models [24, 31, 38], research in person Re-ID has made significant advancements [44, 50, 70]. These works assume a simplistic Re-ID scenario where the target person reappears after a short span of time with limited change

Fig. 1: Samples from DeepChange, a cloth-changing Re-ID dataset [57], which also present severe occlusions besides clothing changes.

in appearance and within similar environment. However, in real-world scenarios, occlusions and clothing changes are frequent, leading to unreliable appearance. Traditional Re-ID methods that rely on appearance suffer a dramatic degradation in performance. [37].

To approach Occluded Re-ID, matching-based methods [18,19,49] divide the image into partitions and design matching strategies at partition level, or extract features and perform matching at body-part level [13,43,69]. Generative models have also been utilized to synthesize the occluded parts in the images [25, 58, 65]. However, they tend to hallucinate irrelevant or unnecessary details. Despite comparable performance on occluded Re-ID datasets [30,71], these methods [15, 45,56,60] *fail under cloth-changing scenario*, shown by their inferior performance on cloth-changing Re-ID datasets (see Sec. 5.1).

For Cloth-Changing Re-ID (CCRe-ID), two main approaches have been proposed: single-modality and multi-modality. Single-modality methods [7,9,11,16, 62] rely primarily on appearance or cloth labels and templates from the RGB images to learn person representations. These texture-based models require large-scale cloth-changing data with explicit clothing labels, which are rarely available in current CCRe-ID datasets. Thus, multi-modality approach remains more effective, where cloth-invariant modalities such as silhouettes [14, 20, 33, 35], contour sketches [4, 61], or skeleton-based pose [32, 34, 39, 47] are captured to be coupled with appearance for distinguishing individuals. However, previous works *ignore the influence of occlusions,* which is a common occurrence in CCRe-ID datasets as shown in Figure 1. This leads to their performance drop in occluded Re-ID environment as shown in Sec. 5.2.

There are *no method that explicitly tackle occlusions and clothing changes simultaneously*. To this end, we propose CrossViT-ReID, the **first framework** for the challenging yet practical Re-ID problem called Occluded Cloth-Changing Person Re-ID (OCCRe-ID). Inspired by [5, 10], given an original RGB image (holistic), we first synthesize occlusion on it to produce a pedestrian occluded RGB image (occluded) where the pedestrian is occluded (see Figure 3). This aims to enhance model's robustness by exposing it to large variations of occlusion. Second, to mitigate clothing changes, besides appearance modality from the RGB image, we capture body shape modality from silhouette masks estimated from the RGB images, which has shown effectiveness in [14, 20, 26]. Body shape cues also helps to enhance Re-ID accuracy under occlusions as will be shown in Table 4. Having the holistic/occluded RGB images and their corresponding silhouette images as inputs, the key to success of our framework is our novel

cross-modality collaborative training strategy which can *adaptively mine the complementary relationship between appearance and shape* under occlusions, clothing changes, and other Re-ID challenges as described below.

Our strategy is based on two intuitive observations: (1) if the visual appearance of the pedestrian is clear, semantic features from RGB image are stronger for Re-ID than shape; (2) however, shape becomes more competitive when appearance is barely observable under occlusions or bad lighting condition. Thus, our idea is that when one modality is occluded or ambiguous, the model should attend to informative cues from the other modality. To achieve this, we construct two branches in CrossViT-ReID. The *appearance-guided* branch takes in the pair of *holistic RGB* and *occluded silhouette* as inputs, which mimics observation (1), while the *shape-guided* branch takes in the pair of *occluded RGB* and *holistic silhouette* to represent observation (2). Given that each branch is trained to focus more on one modality, we need to train the entire framework to *adaptively amplify the stronger modality and reduce the influence of the other noisy modality* under challenging Re-ID scenarios. This is achieved by using cross-attention [2] to fuse the two branches multiple times, enabling them to exchange beneficial information in a collaborative manner.

We choose Vision Transformer [8] as the backbone for our framework since operating on patches enables the finer extraction of local features like face, hairstyle, or body parts, which are beneficial for Re-ID against clothing changes and occlusion. Given that the input RGB image is divided into patches and projected to appearance patch tokens. The input silhouette image is encoded using a shape encoder to obtain shape embedding, which is then coupled with appearance tokens along with positional tokens, forming the input token sequence.

To summarize, our contribution is that we propose CrossViT-ReID, the first solution for Occluded Cloth-Changing Person Re-ID, which is capable of extracting features from appearance and body shape in a collaborative manner, serving as strong cues to tackle occlusions, clothing changes, and other Re-ID challenges. We perform extensive experiments to demonstrate the superiority of CrossViT-ReID over both CCRe-ID and occluded Re-ID methods.

2 Related Work

2.1 Person Re-ID

Under simplistic scenarios without clothing changes or occlusions, person Re-ID has seen remarkable progress. Deep learning models based on Convolutional Neural Networks have been widely adopted as deep feature extractors [44, 50, 70]. Several works have tackled spatial misalignment posed by Re-ID challenges such as viewpoint changes [67, 68], or pose variations [6, 21, 41]. However, these methods are not applicable in real-world scenarios since they rely heavily on appearance features for individual matching, which is unreliable under clothing-change or occlusion situations.

2.2 Occluded Person Re-ID

To tackle the spatial misalignment issue of occluded Re-ID, several matching-based methods [19, 49, 60] have been proposed. Jin *et al.* [19] propose to assign more gradient weights to foreground visible human parts while discarding noisy patches. LTWS [18] obtains matching elements from visual patterns captured along the channel dimension of a CNN backbone. Several methods [5, 45, 54, 56] employ attention mechanisms within Transformers backbone to obtain position information for alignment or noise suppression. Other methods exploit auxiliary information from skeleton-based pose [30, 53] to mitigate challenges brought by occlusions. These methods necessitate a separate network to extract body part based features using graph constructed from pose, leading to inefficient computation cost. Importantly, they assume a cloth-consistent scenario, which does not always hold in real-world situations [37].

2.3 Cloth-Changing Person Re-ID

Since the recent release of cloth-changing Re-ID datasets [39,61], several methods have been proposed to tackle this challenge, in which two main categories can be observed: single-modality and multi-modality. The single-modality methods only leverage the original RGB image to extract biometric features via regional component reconstruction [7], or regional feature augmentation [11, 28]. Clothing labels and templates are mined as auxiliary labels for discriminative learning in [9, 16, 62]. Relying solely on RGB modality remains ineffective due to illumination and occlusion. The multi-modality methods leverage clothes-irrelevant modalities that are more stable in long-term such as contour sketches [4, 61], silhouettes [14, 20, 33, 35], skeleton-based pose [32, 34, 39, 47], or 3D body structure [3, 36]. However, the ambiguity in these 2D human geometric cues caused by viewpoint variations and occlusion has not been well tackled. Moreover, the complementary relationship across modalities has not been mined.

3 Method

3.1 Overview

An overview of the proposed framework is given in Figure 2. Denote the input batch of B RGB images as $X = \{I_i^{rgb}\}_{i=1}^{B}$. For each image $I^{rgb} \in X$, the synthetic occluded image I_{occ}^{rgb} is produced by the occlusion synthesis module. We then use a semantic segmentation model to estimate the corresponding silhouette mask images I^{sil}/I_{occ}^{sil} of I^{rgb}/I_{occ}^{rgb}, respectively. The proposed framework consists of two ViT-based branches: (1) the appearance-guided branch (**A-branch**), which takes in the holistic RGB I^{rgb} and the occluded silhouette I_{occ}^{sil}, aiming to amplify appearance when shape is noisy, and (2) the shape-guided branch (**S-branch**) which takes in the occluded RGB I_{occ}^{rgb} and the holistic silhouette I^{sil}, aiming to attend more to shape cues when appearance is not fully observable. For each branch, the RGB image is divided into patches and projected into

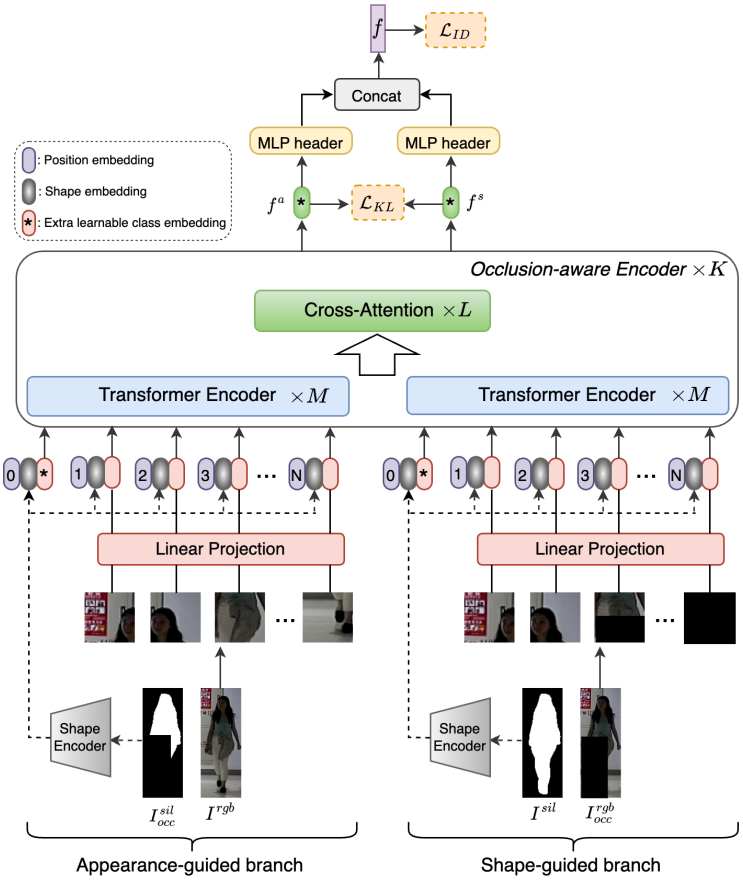

Fig. 2: Overview of CrossViT-ReID, which is described in Sec. 3.1.

patch tokens. Meanwhile, a shape encoder encodes the silhouette image into shape embedding, which is then coupled with patch tokens and positional embeddings, forming input token sequence to M-stacked Transformer encoders [48]. Two branches are then fused using cross-attention. Output classification tokens of the two branches are finally concatenated for the person embedding f.

During inference, the model only consists of the appearance-guided branch (A-branch), while the shape-guided branch and cross-attention are discarded. Taking the RGB and silhouette images as inputs, the CLS token output by the A-branch is used as person embedding for matching, where similarity scores are computed between the query embedding and every gallery embedding, then ranked to obtain the matched ID. The reasons behind our inference model are three-fold. First, this saves significant computation cost and model complexity for effective real-world deployment. Second, our strategy helps to make the best use of the ViT backbone, which is stronger in extracting local visual features from

Fig. 3: Visualization of eight types of occlusion synthesis.

RGB images than from binary masks. Third, during training, cross-attention fusion has enabled the S-branch to provide beneficial shape-guided cues to the A-branch. Moreover, the CLS token output by A-branch is also supervised by identification loss to ensure its discriminative power for Re-ID, which will be described in Sec. 3.5.

3.2 Occlusion Synthesis

Inspired by [5, 10], we perform occlusion synthesis to expose the model to practical variations of occlusion to enhance model's robustness. Specifically, we simulate occlusion by putting a consistent black patch on the RGB image I^{rgb}. The size of the occluded patch (the amount of occlusion) is randomly chosen within the range of height and width of I^{rgb}. Illustration of occlusion synthesis is provided in Figure 3. For each image, we randomly perform one of these eight types of occlusion: any of the four corners (#1 - #4), or top or bottom (#5 and #6), or left half or right half (#7 and #8).

3.3 ViT-based Feature Extraction

Each branch in CrossViT-ReID is built on top of Vision Transformer (ViT) [8]. The key difference between two branches is their inputs, while their architectures are identical. Thus, without loss of generality, we describe the A-branch below.

Given the input RGB image I^{rgb}, ViT first breaks I^{rgb} into N patches by dividing it with a certain patch size. The sequence of patch tokens $\mathbf{x}_{patch} \in \mathbb{R}^{N \times C}$ is then obtained by linearly projecting each flattened patch into tokens using linear or convolutional projection. C is the dimension of the embedding. An additional classification (CLS) token $\mathbf{x}_{cls} \in \mathbb{R}^{1 \times C}$ is added to the sequence as in the original ViT [8]. A-branch consists of M transformer encoders to encode information from the token sequence. Since self-attention in the transformer encoder is position-agnostic, position embeddings $\mathbf{x}_{pos} \in \mathbb{R}^{(1+N) \times C}$ is added into each token in the sequence (including the CLS token) to provide spatial information and encode the relative positions of the patches within the image.

Different from previous occluded Re-ID methods which directly employ ViT baseline for Re-ID [5, 17, 27] and fail under clothing changes, in this work, we incorporate cloth-invariant body shape information captured from the silhouette image to address this. The silhouette mask image is passed through a shape encoder $\mathcal{F}^S$ to output shape embedding $\mathbf{s} \in \mathbb{R}^{1 \times C}$. Architecure of shape encoder

$\mathcal{F}^S$ is given in Table 1b. $\mathbf{s}$ is then injected into each token in the sequence in an additive manner to obtain the token sequence $\mathbf{x}_0$ as follows:

$$\mathbf{x}_0 = \mathbf{x}_{pos} + \mathbf{s} + [\mathbf{x}_{cls} \,||\, \mathbf{x}_{patch}]. \tag{1}$$

$\mathbf{x}_0$ serves as input to the M-stacked transformer encoders. Each transformer encoder consists of a sequence of blocks. Each block contains multihead self-attention ($\mathcal{H}$) with a feed-forward network ($\mathcal{F}$). $\mathcal{F}$ contains two multilayer perceptron (MLP) layers. Expanding ratio r is applied at the hidden layer while GELU activation is applied after the first linear layer for non-linearity. Before every block, layer normalization (LN) and residual shortcuts are applied. The processing of the k^{th} block can be expressed as:

$$\mathbf{y}_k = \mathbf{x}_{k-1} + \mathcal{H}(LN(\mathbf{x}_{k-1})), \tag{2}$$

$$\mathbf{x}_k = \mathbf{y}_k + \mathcal{F}(LN(\mathbf{y}_k)). \tag{3}$$

3.4 Cross-branch Feature Fusion

Effective information fusion between two branches is important for the model to be capable of adaptively attending to the most informative cues for Re-ID in-the-wild. Four fusion strategies were proposed [2], namely all-attention fusion, class token fusion, pairwise fusion, and cross-attention fusion. Experimental comparison reveals that cross-attention helps to yield the best performance (see Table 6a). The idea behind cross-attention fusion is to make the CLS token of one branch interact with patch tokens of the other branch. The CLS token already captures abstract information among patch tokens in its own branch, thus can serve as an agent to exchange information with the other branch. After cross-attention fusion with the other branch, the CLS token is passed through the next transformer encoders of its own branch to transfer the learned information to its own patch tokens.

Cross-attention fusion procedure is the same for both branches. Here, we describe the cross-attention for A-branch (denoted by index a). Patch tokens from S-branch are concatenated with CLS token of A-branch as input for fusion:

$$\hat{\mathbf{x}}^a = \left[\mathbf{x}^a_{cls} \,||\, \mathbf{x}^s_{patch}\right]. \tag{4}$$

Multi-head cross-attention between $\mathbf{x}^a_{cls}$ and $\hat{\mathbf{x}}^a$ is then expressed as:

$$q = \mathbf{x}^a_{cls} W_q,\ k = \hat{\mathbf{x}}^a W_k,\ v = \hat{\mathbf{x}}^a W_v, \tag{5}$$

$$A = \text{softmax}\left(qk^T/\sqrt{C/h}\right), \tag{6}$$

$$\mathbf{y}^a_{cls} = \mathbf{x}^a_{cls} + Av, \tag{7}$$

$$\mathbf{z}^a = \left[\mathbf{y}^a_{cls} \,||\, \mathbf{x}^l_{patch}\right], \tag{8}$$

where $\mathbf{z}^a$ is the output of cross-attention. q, k, v denote query, key, and value, $W_q, W_k, W_v \in \mathbb{R}^{C\times(C/h)}$ are learnable parameters, C is the embedding dimension, h is the number of heads, and T denotes transpose. Note that patch size is identical for both branch, thus needing no projection for dimension alignment.

3.5 Loss functions

After K occlusion-aware encoders, where each occlusion-aware encoder comprises M Transformer encoders and L cross-attention layers, the CLS tokens f^a of A-branch and f^s of S-branch are used as branch-wise person embeddings. They are fed into MLP headers, then concatenated for the final person representation f. For identity supervision, we apply the identification losses $\mathcal{L}_{ID}$, $\mathcal{L}^a_{ID}$, and $\mathcal{L}^s_{ID}$ to f, f^a, and f^s, respectively, where each identification loss is the sum of a cross-entropy loss $\mathcal{L}_{ce}$ and a triplet loss $\mathcal{L}_{tri}$.

Moreover, to further enforce model's identity-awareness, we encourage a semantic consistency between the person embeddings of the two branches using the $\mathcal{L}_{KL}$ loss based on Kullback-Leibler (KL) divergence:

$$\mathcal{L}_{KL} = D_{KL}\left(\hat{p}^s \| \hat{p}^a\right) + D_{KL}\left(\hat{p}^a \| \hat{p}^s\right), \tag{9}$$

where $D_{KL}(\cdot)$ is the KL distance, $\hat{p}^a$ denotes output class probabilities of f^a from A-branch, and $\hat{p}^s$ denotes output class probabilities of f^s from S-branch. The total loss of the framework is:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{ID} + \lambda_2 \mathcal{L}^a_{ID} + \lambda_3 \mathcal{L}^s_{ID} + \lambda_4 \mathcal{L}_{KL}, \tag{10}$$

where $\lambda_1, ..., \lambda_4$ are scalars controlling the scale of each loss.

4 Experiments Setup

4.1 Datasets

Till date, there are no datasets for OCCRe-ID (explicitly contains both occlusions and clothing changes). Thus, we evaluate our method separately on occluded Re-ID datasets (Occluded-Duke [30] and Occluded-ReID [71]), and CCRe-ID datasets (LTCC [39], PRCC [61], and DeepChange [57]). Synthesizing datasets to facilitate OCCRe-ID is potential, which we leave for future works.

1. Occluded-Duke consists of 15,618 training images, 2,210 occluded query images and 17,661 gallery images. Though Duke-MTMC dataset [40] has been retracted, Occluded-Duke is available for use in research purposes [17,53,54]. Occluded-ReID contains 2,000 images of 200 identities. Each pedestrian has five occluded images.
2. *LTCC* contains a total of 17,138 images of 152 identities captured by 12 cameras. *PRCC* is a larger CCRe-ID dataset with a total of 33,698 images of 221 identities captured by three cameras. However, compared to PRCC, LTCC presents larger variations in clothing, pose, viewpoint, illumination, and especially occlusions, making it more challenging for Re-ID. Thus, it can serve as a strong benchmark dataset for comparison. *DeepChange* is a large-scale dataset which was collected under various outdoor scenes and weather conditions. It consists of 178k images from 1,121 identities captured by 17 cameras during twelve months. Since there is a significant distribution gap in data between DeepChange and LTCC/PRCC, it can be used to validate generalization ability of the Re-ID model.

Model	Patch Emb.	#heads	K	M	L	r
CrossViT-ReID-B	Linear	12	3	4	1	4
CrossViT-ReID-18	Linear	7	3	6	1	3
CrossViT-ReID-18*	3 Conv	7	3	6	1	3

(a) Model architecture configurations.

Layer	#kernels	BN	Activation	Pooling
Conv	32	True	ReLU	Max
Conv	32	True	ReLU	Max
FC-1024	None	True	ReLU	None
FC-512	None	False	None	None

(b) Architecture of shape encoder $\mathcal{F}^S$.

Table 1: Model architecture of a) CrossViT-ReID and b) Shape encoder

4.2 Evaluation Protocols

Previous methods are proposed for either Occluded Re-ID or CCRe-ID, and our method is the first solution for OCCRe-ID. Thus, we perform separate comparisons with SOTA Occluded Re-ID methods and SOTA CCRe-ID methods.

CMC at rank 1 (R-1 accuracy) and mean Average Precision (mAP) are used as evaluation metrics. For LTCC, two evaluation settings are set up: (1) Cloth-Changing (CC) where only cloth-changing samples are used for testing, and (2) Standard, where both cloth-consistent and cloth-changing samples form the test sets. For PRCC, only CC setting is performed. We did not perform Standard setting on PRCC because previous methods already reached a saturation in performance with 100% R-1 accuracy and mAP on this setting. This is due to the dataset's nature, where the test sets only contain cloth-consistent samples, and its simplistic Re-ID scenarios make this setting not challenging. For DeepChange, since no clothing labels are provided, we could only perform evaluation with Standard setting.

4.3 Implementation Details

Training. The framework is trained for 90 epochs on four 32GB V100 GPUs with a batch size of 128. We use Adam optimizer [23] with an initial learning rate of 0.0005, which is reduced by a factor of 0.1 after every 30 epochs, momentum of 0.9, and weight decay of 0.01. The images are resized to 256×128 as inputs. We set $\lambda_1 = 1$, $\lambda_2 = 0.7$, $\lambda_3 = 0.5$, and $\lambda_4 = 0.5$. λ_2 (associated with $\mathcal{L}_{ID}^a$ of A-branch) is set to be larger than λ_3 (associated with $\mathcal{L}_{ID}^s$ of S-branch) to provide A-branch with stronger identity supervision for more accurate inference (see Sec. 3.1 for more detailed rationale).

Model Architecture. A summary of model configurations used in our experiments is shown in Table 1a. K denotes the number of occlusion-aware encoders. M denotes the number transformer encoders in each branch. L denotes the number of cross-attention modules in one occlusion-aware encoder. r is the expanding ratio of feed-forward network in the transformer encoder. "*" denotes model using convolutional layers for patch embeddings instead of linear layers. The first model CrossViT-ReID-B sets each branch identical to the DeiT-base model [46]. For the remaining two models, the number 18 is equal to $K \times M$, which is the total number of transformer encoders in each branch. For all models, number of

Method	Backbone	Occluded-Duke		Occluded-ReID		LTCC			
		R-1	mAP	R-1	mAP	CC		Standard	
						R-1	mAP	R-1	mAP
PGFA [30]†	CNN	51.4	37.2	-	-	26.3	16.1	67.6	34.2
ASAN [19]	CNN	55.4	43.8	82.5	71.8	-	-	-	-
HOReID [49]†	CNN	55.1	43.8	80.3	70.2	29.1	18.0	69.3	36.2
OAMM [5]	CNN	62.6	46.1	-	-	-	-	-	-
LKWS [60]	CNN	62.2	46.3	81.0	71.0	-	-	-	-
OCNet [22]	CNN	64.3	54.4	-	-	-	-	-	-
QPM [51]	CNN	64.4	49.7	-	-	-	-	-	-
RTGTA [15]	CNN	61.0	50.1	71.8	51.0	-	-	-	-
FED [54]†	Trans	68.1	56.4	86.3	79.3	31.5	18.9	70.5	37.0
PFD [53]†	Trans	67.7	60.1	79.8	81.3	34.0	19.2	72.9	37.8
FRT [56]†	Trans	70.7	61.3	80.4	71.0	33.1	19.3	73.0	37.2
DPM [45]†	Trans	71.4	61.8	85.5	79.7	32.5	17.8	70.9	36.7
CrossViT-ReID-B	Trans	72.8	62.9	87.0	81.1	48.1	24.2	79.0	44.9
CrossViT-ReID-18	Trans	73.4	63.2	87.5	81.7	48.6	24.3	79.7	45.5
CrossViT-ReID-18*	Trans	**73.7**	**63.6**	**87.9**	**82.0**	**49.0**	**24.5**	**79.9**	**45.8**
CrossViT-ReID-18*‡	Trans	67.4	59.5	83.6	79.9	44.1	21.8	75.2	41.8

Table 2: Comparison with Occluded ReID methods Occluded-Duke and Occluded-ReID and a CCRe-ID dataset (LTCC). Best results are shown in **bold**, while second-best results (among other CCRe-ID methods, not including variations of CrossViT-ReID) are underlined. "†" denotes we reproduced results on LTCC from the available open-source code. "Trans" denotes model using Transformer-based backbone. "‡" denotes model trained without Occlusion Synthesis. "-" denotes results are not available.

heads (#heads) is the same for both branches. Silhouette masks are extracted using Detectron2 [55]. Architecture of shape encoder is shown in Table 1b. Patch size is 16 following [8].

5 Results

5.1 Comparison with Occluded Re-ID methods

We performed quantitative comparison of our method with previous Occluded Re-ID methods on two occluded datasets (Occluded-Duke and Occluded-REID). Moreover, we show that previous Occluded Re-ID methods struggled on a CCRe-ID dataset (LTCC) as they do not tackling clothing changes. This shows the superiority of our method over previous occluded Re-ID methods in both occluded and cloth-changing scenarios. Our model is robust to clothing changes via shape learning, along with its enhanced ability to tackle occlusions via our occlusion-aware cross-modality complementary training strategy.

As shown in Table 2, all settings of CrossViT-ReID outperform previous CNN-based and Transformer-based methods. Compared to the CNN-based approaches that use pose as auxiliary features [49, 51, 60] or leverage graph convolution [15], Transformer-based methods show remarkable superiority, which demonstrates the effectiveness of operating on patches using Transformer backbone for Re-ID. Compared to DPM [45] (a Transformer-based that performs second-best), CrossViT-ReID achieves a large improvement of 2.3%/1.8% and 2.4%/2.3% in R-1/mAP on Occluded-Duke and Occluded-ReID, respectively.

Methods	Modality	LTCC				PRCC		DeepChange		O-Duke	
		CC		Standard		CC		R-1	mAP	R1	mAP
		R-1	mAP	R-1	mAP	R-1	mAP				
PCB [44]†	RGB	23.5	10.0	65.1	30.6	41.8	38.7	38.1	10.5	42.6	33.7
TransReID [12]†	RGB	33.5	12.1	69.4	32.4	46.2	40.7	41.5	11.2	52.1	40.6
RCSANet [16]	RGB	-	-	-	-	50.2	48.6	-	-	-	-
CAL [9]†	RGB	40.1	18.1	74.2	40.8	55.2	55.8	45.2	12.3	58.1	51.4
ACID [63]	RGB	29.1	14.5	65.1	30.6	55.4	56.1	-	-	-	-
IGEP [66]	RGB	43.4	18.2	-	-	57.3	55.8	-	-	-	-
AFL [29]	RGB	42.1	18.4	74.4	39.1	57.4	56.5	-	-	-	-
LDF [1]	RGB	32.9	15.4	73.4	36.9	54.6	-	-	-	-	-
CCFA [11]	RGB	45.3	22.1	75.8	42.5	61.2	58.4	-	-	-	-
SAGE [61]†	RGB + sketch	-	-	-	-	34.4	-	43.5	10.9	48.0	42.9
CESD [39]	RGB + pose	26.2	12.4	71.4	34.3	-	-	-	-	-	-
GI-ReID [20]†	RGB + sil	23.7	10.4	63.2	29.4	33.3	-	43.9	11.0	55.2	46.6
3DSL [3]	RGB + pose + sil	31.2	14.8	-	-	51.3	-	-	-	-	-
FSAM [14]	RGB + pose + sil	38.5	16.2	73.2	35.4	54.5	-	-	-	-	-
CAMC [52]	RGB + pose	36.0	15.4	73.2	35.3	-	-	-	-	-	-
DCR-ReID [7]	RGB + sil	41.1	20.4	76.1	42.3	57.2	57.4	-	-	-	-
IRANet [42]	RGB + pose	-	-	-	-	54.9	53.0	-	-	-	-
UCAD [59]	RGB + sil	23.7	10.4	63.2	29.4	45.3	-	-	-	-	-
MBUNet [64]	RGB + pose	40.3	15.0	67.6	34.8	58.7	55.2	-	-	-	-
AIM [62]†	RGB + gray	40.6	19.1	76.3	41.1	57.9	58.3	45.3	12.1	57.6	50.0
CVSL [32]	RGB + pose	44.5	21.3	76.4	41.9	57.5	56.9	-	-	-	-
CCPG [34]	RGB + pose	46.2	22.9	77.2	42.9	61.8	58.3	-	-	-	-
CrossViT-ReID-B	RGB + sil	48.1	24.2	79.0	44.9	64.0	59.9	49.6	14.2	72.8	62.9
CrossViT-ReID-18	RGB + sil	48.6	24.3	79.7	45.5	64.3	60.0	50.4	14.9	73.4	63.2
CrossViT-ReID-18*	RGB + sil	**49.0**	**24.5**	**79.9**	**45.8**	**64.6**	**60.2**	**50.6**	**15.0**	**73.7**	**63.6**
CrossViT-ReID-18*‡	RGB	43.1	20.3	72.0	39.9	56.5	52.7	46.1	12.8	72.6	62.9

Table 3: Comparison with CCRe-ID methods on three CCRe-ID datasets (LTCC, PRCC, and DeepChange) and an Occluded Re-ID dataset (O-Duke). Best results are shown in **bold**, while second-best results (among other CCRe-ID methods, not including variations of CrossViT-ReID) are underlined. "†" denotes we reproduced results on Occluded-Duke the available open-source code. "‡" denotes not using shape embedding in model. "-" denotes results are not available.

The effectiveness comes from our comprehensive occlusion handling strategy via cross-modality collaborative learning technique. Importantly, clothing changes have not been considered in previous occluded Re-ID works, and we achieve significantly higher accuracy in this important real-world scenario. For example, CrossViT-ReID outperforms PFD [53] by 15%/6.3% in R-1/mAP in cloth-changing setting on LTCC. In terms of model configurations, using 18 transformer encoders in each branch raises the accuracy of DeiT-baseline by another 0.6%/0.3% in R-1/mAP on Occluded-Duke. Moreover, convolution-based patch tokenizer shows superiority over using linear-based tokenizer, which supports the hypothesis stated in the original ViT paper [8].

5.2 Comparison with CCRe-ID methods

In Table 3, we quantitatively compare our method with previous CCRe-ID methods on LTCC, PRCC, and DeepChange. Further, we reproduced results of CCRe-ID methods with public source code on Occluded-Duke, which shows lower performance on this dataset due to not explicitly tackling occlusions. Over-

A-branch	S-branch	Occluded-Duke		LTCC			
		R-1	mAP	CC		Standard	
				R-1	mAP	R-1	mAP
✓	-	68.3	58.0	31.7	15.3	60.1	32.7
-	✓	63.4	52.3	37.5	18.8	64.9	38.1
✓	✓	**73.7**	**63.6**	**49.0**	**24.5**	**79.9**	**45.8**

Table 4: Ablation study on the contribution of each branch in the framework.

all, we significantly outperform previous CCRe-ID methods by a large margin, where the improvements come from explicitly handling occlusion via occlusion synthesis and our cross-modality training strategy, making the model robust to occlusions, while enhancing it's discriminability against clothing changes.

Among CCRe-ID methods, results of the multi-modality methods [3, 14, 20, 32, 39, 52, 61, 62] show that this approach is effective in mitigating the influence of clothing changes. Although recent texture-based methods [9, 11] achieve comparable performance, they rely heavily on the availability of large clothing variations with expensive manual cloth labeling. Importantly, it can be seen that occlusions cause relatively inferior performance in CCRe-ID methods on Occluded-Duke since they ignore this scenario. Moreover, the discriminative fine-grained identity information across modalities has not been mined. By addressing these issues, all model settings of CrossViT-ReID outperform previous methods in both cloth-changing and occluded Re-ID environment. Specifically, CrossViT-ReID-18* achieves an improvement of 2.8%/1.6% and 2.7%/2.9% in R-1/mAP accuracy on LTCC and PRCC compared to the second-best results, respectively. Compared to TransReID which uses a simple ViT baseline for feature extraction, we achieve a significant gap in performance, showing the effectiveness of our dual-ViT-branch framework. On Occluded-Duke, a very remarkable gap of 15.6%/12.2% between CrossViT-ReID and the second-best method CAL [9] can be observed. Generalizability of the model is demonstrated via results on DeepChange. It can be seen that our method outperforms the second-best methods AIM [62] and CAL [9] by 5.6/2.7% in R-1/mAP.

5.3 Ablation Study

To further validate the effectiveness of our proposed method, we perform comprehensive ablation studies on: 1) contribution of each branch, 2) effectiveness of shape embedding, 3) occlusion synthesis, 4) cross-attention fusion schemes, 5) identity losses, and 6) model architecture configurations.

Appearance-guided vs Shape-guided branch. In Table 4, we run experiments with adding one branch and excluding the other to explore the contribution of each in occluded and cloth-changing Re-ID environments. It can be seen that using only S-branch can effectively guide the model on distinguishing identities under clothing variations. However, large performance drop can be seen on Occluded-Duke using only S-branch. This is because S-branch takes in the occluded RGB image, while for Occluded Re-ID visual patterns from body parts and appearance remain competitive for matching. In contrast, guiding the model

Model	$\mathcal{L}_{ID}^{a}$	$\mathcal{L}_{ID}^{s}$	$\mathcal{L}_{KL}$	Occluded-Duke		LTCC			
				R-1	mAP	CC		Standard	
						R-1	mAP	R-1	mAP
1	-	-	-	72.6	62.0	48.4	23.9	78.9	44.9
2	✓	-	-	73.0	62.5	48.7	24.2	79.2	45.5
3	✓	✓	-	73.2	62.9	48.8	24.3	79.6	45.7
4	✓	✓	✓	**73.4**	**63.1**	**49.0**	**24.5**	**79.9**	**45.8**

Table 5: Ablation study on the contribution of each loss.

by only appearance leads to inferior performance on LTCC, since the model relies heavily on unreliable texture information while body shape information can not be fully captured from occluded silhouette. It is demonstrated that overall, it is best to couple both branches, shown by an improvement in Re-ID performance on both datasets.

Shape Embedding. In the last rows of Table 3, we study how coupling body shape information with appearance can lead to a more discriminative person representation. It can be seen that shape embedding has a greater impact in cloth-changing environment, where effectiveness is clearly shown on CCRe-ID datasets. For example, shape leads to an improvement of 3.9%/3.2% in R-1/mAP in CC setting on LTCC. A larger gap can be seen on PRCC, which can be reasoned by its nature where most of its data is captured in frontal viewpoint, giving clear and informative body shape. On Occluded-Duke, shape embedding also leads to a slight boost in Re-ID performance, which indicates that coupling shape with appearance can also help to enhance discriminability against occlusion.

Occlusion Synthesis. From the last row of Table 2 where we performed experiment without occlusion synthesis, a significant performance drop can be seen. This step has a greater impact in occluded environment, shown by a larger performance gap on Occluded-Duke (6.3%/4.1 in R-1/mAP) when applying occlusion synthesis. Accuracy on LTCC is also boosted (2.9%/1.7% in R-1/mAP), demonstrating that occlusions commonly compound in CCRe-ID datasets and it is effective to address occlusions via augmentation.

Loss functions. Table 5 reports the contribution of each loss in enhancing the discriminability of our proposed framework. It can be seen that R-1 accuracy is improved by 0.4% on Occluded-Duke and 0.3% in cloth-changing setting on LTCC when adding the identification loss for A-branch $\mathcal{L}_{ID}^{a}$. Re-ID accuracy is further boosted on both datasets by applying the identification loss $\mathcal{L}_{ID}^{s}$ to S-branch. With Model 4, where we incorporated every loss, a clear quantitative improvement can be seen on the performance of our proposed framework.

Different Fusion Schemes. We compare different fusion schemes proposed in [2] in Table 6a. It can be seen that overall, fusion helps improve Re-ID performance compared to "None". For example, by using all-attention, R-1 accuracy is boosted by 0.7%/0.5% on Occluded-Duke/LTCC. Compared to class token which uses only CLS tokens for fusion, cross-attention which also leverages patch tokens shows superiority. This is because cloth-changing and occluded Re-ID patch tokens contain beneficial local features from face or body parts. Overall, cross-attention achieves the best accuracy, while there is not a significant gap among

Scheme	O-Duke		LTCC			
	R-1	mAP	CC		Standard	
			R-1	mAP	R-1	mAP
None	71.9	61.1	46.5	23.8	77.6	43.8
All-Att	72.6	61.9	47.0	24.3	78.1	44.2
CLS Token	72.8	62.4	47.7	24.8	78.8	44.9
Pairwise	72.7	62.6	47.9	25.0	78.6	45.0
Cross-Att	**73.7**	**63.6**	**49.0**	**24.5**	**79.9**	**45.8**

(a) Ablation study on the effectiveness of different fusion schemes.

Model	K	M	L	O-Duke		LTCC			
				R-1	mAP	CC		Standard	
						R-1	mAP	R-1	mAP
Base	3	6	1	73.4	63.2	48.1	24.2	79.0	44.9
A	3	<u>3</u>	1	69.3	59.1	46.2	22.4	77.2	43.9
B	3	<u>5</u>	1	72.1	62.3	47.2	23.7	78.1	44.1
C	3	6	<u>2</u>	73.4	63.1	47.7	23.8	78.5	44.5
D	<u>6</u>	6	1	73.3	63.2	47.9	24.1	78.8	44.7

(b) Different model configurations. <u>Underline</u> indicates changes from CrossViT-ReID-18

Table 6: Ablation study on fusion schemes and model configurations

the remaining three schemes, demonstrating its effectiveness in capturing the complementary relationship between branches to assist Re-ID.

Model configurations. We provide insights in different model configurations by using CrossViT-ReID-18 as baseline then modifying the layer architecture. Results are shown in Table 6b. From model A and B, it can be seen that larger number of transformer encoders per branch can lead to significantly better Re-ID results. This is because Re-ID is a fine-grained image retrieval task, and having deeper encoder further helps in extracting low-level features from limited identity-aware information of the person in the image. In terms of depth of cross-attention within an occlusion-aware encoder, we can stack more cross-attention modules (L). Results of model C shows that increasing frequency of information fusion across two branches does not help improve Re-ID performance. Effectiveness of deeper cross-attention is small since cross-attention is a linear operation without any nonlinearity function. Another way to increase frequency of information fusion is to increase the number of occlusion-aware encoders (model D). However, similarly, this does not help boost Re-ID accuracy. Moreover, both approaches will introduce more complexity in the model architecture.

6 Conclusion

We have introduced CrossViT-ReID, the first framework for addressing occlusions and clothing changes in Re-ID simultaneously. Leveraging shape learned from silhouette masks and appearance learned from RGB images, CrossViT-ReID integrates these modalities collaboratively, enhancing robustness against occlusions through augmentation and cross-attention fusion. Experiments on CCRe-ID and occluded Re-ID datasets demonstrate that our method outperforms current SOTA methods by a large margin. Our work paves the way for the challenging yet practical occluded cloth-changing Re-ID task.

Future works. We recently had some findings that occlusion type awareness is important to learn discriminative features from occluded inputs. Thus, in future work, we aim to plug in an occlusion detection module to produce occlusion-aware features, which can be used to guide feature extraction of the Re-ID backbone. Furthermore, we intend to develop an efficient multi-scale transformer-based framework to further tackle OCCRe-ID.

References

1. Chan, P.P.K., Hu, X., Song, H., Peng, P., Chen, K.: Learning disentangled features for person re-identification under clothes changing. ACM MM **19**(6) (2023)
2. Chen, C.F.R., Fan, Q., Panda, R.: Crossvit: Cross-attention multi-scale vision transformer for image classification. In: ICCV. pp. 357–366 (2021)
3. Chen, J., Jiang, X., Wang, F., Zhang, J., Zheng, F., Sun, X., Zheng, W.S.: Learning 3d shape feature for texture-insensitive person re-identification. In: CVPR. pp. 8142–8151 (2021). https://doi.org/10.1109/CVPR46437.2021.00805
4. Chen, J., Zheng, W.S., Yang, Q., Meng, J., Hong, R., Tian, Q.: Deep shape-aware person re-identification for overcoming moderate clothing changes. IEEE TMM **24**, 4285–4300 (2022). https://doi.org/10.1109/TMM.2021.3114539
5. Chen, P., Liu, W., Dai, P., Liu, J., Ye, Q., Xu, M., Chen, Q., Ji, R.: Occlude them all: Occlusion-aware attention network for occluded person re-id. In: ICCV. pp. 11833–11842 (2021)
6. Cho, Y.J., Yoon, K.J.: Pamm: Pose-aware multi-shot matching for improving person re-identification. IEEE TIP **27**(8), 3739–3752 (2018). https://doi.org/10.1109/TIP.2018.2815840
7. Cui, Z., Zhou, J., Peng, Y., Zhang, S., Wang, Y.: Dcr-reid: Deep component reconstruction for cloth-changing person re-identification. IEEE TCSVT (2023). https://doi.org/10.1109/TCSVT.2023.3241988
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
9. Gu, X., Chang, H., Ma, B., Bai, S., Shan, S., Chen, X.: Clothes-changing person re-identification with rgb modality only. In: CVPR. pp. 1050–1059 (2022). https://doi.org/10.1109/CVPR52688.2022.00113
10. Gupta, A., Chellappa, R.: You can run but not hide: Improving gait recognition with intrinsic occlusion type awareness. In: WACV. pp. 5893–5902 (2024)
11. Han, K., Gong, S., Huang, Y., Wang, L., Tan, T.: Clothing-change feature augmentation for person re-identification. In: CVPR. pp. 22066–22075 (2023)
12. He, S., Luo, H., Wang, P., Wang, F., Li, H., Jiang, W.: Transreid: Transformer-based object re-identification. In: CVPR. pp. 15013–15022 (2021)
13. He, T., Shen, X., Huang, J., Chen, Z., Hua, X.S.: Partial person re-identification with part-part correspondence learning. In: CVPR. pp. 9105–9115 (2021)
14. Hong, P., Wu, T., Wu, A., Han, X., Zheng, W.S.: Fine-grained shape-appearance mutual learning for cloth-changing person re-identification. In: CVPR. pp. 10508–10517 (2021). https://doi.org/10.1109/CVPR46437.2021.01037
15. Huang, M., Hou, C., Yang, Q., Wang, Z.: Reasoning and tuning: Graph attention network for occluded person re-identification. IEEE TIP **32**, 1568–1582 (2023)
16. Huang, Y., Wu, Q., Xu, J., Zhong, Y., Zhang, Z.: Clothing status awareness for long-term person re-identification. In: ICCV. pp. 11875–11884 (2021). https://doi.org/10.1109/ICCV48922.2021.01168
17. Jia, M., Cheng, X., Lu, S., Zhang, J.: Learning disentangled representation implicitly via transformer for occluded person re-identification. IEEE TMM **25**, 1294–1305 (2022)
18. Jia, M., Cheng, X., Zhai, Y., Lu, S., Ma, S., Tian, Y., Zhang, J.: Matching on sets: Conquer occluded person re-identification without alignment. In: AAAI. vol. 35, pp. 1673–1681 (2021)

19. Jin, H., Lai, S., Qian, X.: Occlusion-sensitive person re-identification via attribute-based shift attention. IEEE TCSVT **32**(4), 2170–2185 (2021)
20. Jin, X., He, T., Zheng, K., Yin, Z., Shen, X., Huang, Z., Feng, R., Huang, J., Chen, Z., Hua, X.S.: Cloth-changing person re-identification from a single image with gait prediction and regularization. In: CVPR. pp. 14258–14267 (2022). https://doi.org/10.1109/CVPR52688.2022.01388
21. Khaldi, K., Nguyen, V.D., Mantini, P., Shah, S.: Unsupervised person re-identification in aerial imagery. In: WACVW. pp. 260–269 (2024)
22. Kim, M., Cho, M., Lee, H., Cho, S., Lee, S.: Occluded person re-identification via relational adaptive feature correction learning. In: ICASSP. pp. 2719–2723 (2022)
23. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
24. Le, N., Pham, T., Do, T., Tjiputra, E., Tran, Q.D., Nguyen, A.: Music-driven group choreography. In: CVPR. pp. 8673–8682 (2023)
25. Li, S., Bak, S., Carr, P., Wang, X.: Diversity regularized spatiotemporal attention for video-based person re-identification. In: CVPR. pp. 369–378 (2018)
26. Li, Y.J., Weng, X., Kitani, K.M.: Learning shape representations for person re-identification under clothing change. In: WACV. pp. 2431–2440 (2021)
27. Li, Y., He, J., Zhang, T., Liu, X., Zhang, Y., Wu, F.: Diverse part discovery: Occluded person re-identification with part-aware transformer. In: CVPR. pp. 2898–2907 (2021)
28. Liu, F., Ye, M., Du, B.: Dual level adaptive weighting for cloth-changing person re-identification. IEEE TIP **32**, 5075–5086 (2023). https://doi.org/10.1109/TIP.2023.3310307
29. Liu, Y., Ge, H., Wang, Z., Hou, Y., Zhao, M.: Clothes-changing person re-identification via universal framework with association and forgetting learning. IEEE TMM **26**, 4294–4307 (2024)
30. Miao, J., Wu, Y., Liu, P., Ding, Y., Yang, Y.: Pose-guided feature alignment for occluded person re-identification. In: ICCV. pp. 542–551 (2019)
31. Nguyen, T.P., Pham, T.T., Nguyen, T., Le, H., Nguyen, D., Lam, H., Nguyen, P., Fowler, J., Tran, M.T., Le, N.: Embryosformer: Deformable transformer and collaborative encoding-decoding for embryos stage development classification. In: WACV. pp. 1981–1990 (2023)
32. Nguyen, V.D., Khaldi, K., Nguyen, D., Mantini, P., Shah, S.: Contrastive viewpoint-aware shape learning for long-term person re-identification. In: WACV. pp. 1041–1049 (2024)
33. Nguyen, V.D., Mantini, P., Shah, S.K.: Acml: Attention-based cross-modality learning for cloth-changing and occluded person re-identification. In: 2024 IEEE International Conference on Image Processing (ICIP). pp. 2396–2402 (2024). https://doi.org/10.1109/ICIP51287.2024.10647794
34. Nguyen, V.D., Mantini, P., Shah, S.K.: Contrastive clothing and pose generation for cloth-changing person re-identification. In: CVPRW. pp. 7541–7549 (2024)
35. Nguyen, V.D., Mantini, P., Shah, S.K.: Occluded cloth-changing person re-identification via occlusion-aware appearance and shape reasoning. In: 2024 IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). pp. 1–8 (2024). https://doi.org/10.1109/AVSS61716.2024.10672564
36. Nguyen, V.D., Mantini, P., Shah, S.K.: Temporal 3d shape modeling for video-based cloth-changing person re-identification. In: WACVW. pp. 173–182 (2024)
37. Nguyen, V.D., Mirza, S., Zakeri, A., Gupta, A., Khaldi, K., Aloui, R., Mantini, P., Shah, S.K., Merchant, F.: Tackling domain shift in person re-identification: A survey and analysis. In: CVPRW. pp. 4149–4159 (2024)

38. Pham, T.T., Brecheisen, J., Nguyen, A., Nguyen, H., Le, N.: I-ai: A controllable & interpretable ai system for decoding radiologists' intense focus for accurate cxr diagnoses. In: WACV. pp. 7850–7859 (2024)
39. Qian, X., Wang, W., Zhang, L., Zhu, F., Fu, Y., Xiang, T., Jiang, Y.G., Xue, X.: Long-term cloth-changing person re-identification. In: ACCV. pp. 71–88 (2021)
40. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: ECCV. pp. 17–35 (2016)
41. Sarfraz, M.S., Schumann, A., Eberle, A., Stiefelhagen, R.: A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In: CVPR. pp. 420–429 (2017)
42. Shi, W., Liu, H., Liu, M.: Iranet: Identity-relevance aware representation for cloth-changing person re-identification. Image and Vision Computing **117**, 104335 (2022)
43. Sun, Y., Xu, Q., Li, Y., Zhang, C., Li, Y., Wang, S., Sun, J.: Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification. In: CVPR. pp. 393–402 (2019)
44. Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S.: Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In: ECCV. p. 501–518 (2018)
45. Tan, L., Dai, P., Ji, R., Wu, Y.: Dynamic prototype mask for occluded person re-identification. In: ACM MM. pp. 531–540 (2022)
46. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: ICML. pp. 10347–10357. PMLR (2021)
47. Trinh, Q.H., Bui, N.T., Hoang, D.H., Thi, P.T.V., Nguyen, H.D., Jha, D., Bagci, U., Le, N., Tran, M.T.: Pgds: Pose-guidance deep supervision for mitigating clothes-changing in person re-identification. In: 2024 IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). pp. 1–8 (2024). https://doi.org/10.1109/AVSS61716.2024.10672607
48. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
49. Wang, G., Yang, S., Liu, H., Wang, Z., Yang, Y., Wang, S., Yu, G., Zhou, E., Sun, J.: High-order information matters: Learning relation and topology for occluded person re-identification. In: CVPR. pp. 6449–6458 (2020)
50. Wang, G., Yuan, Y., Chen, X., Li, J., Zhou, X.: Learning discriminative features with multiple granularities for person re-identification. In: ACM MM (2018). https://doi.org/10.1145/3240508.3240552
51. Wang, P., Ding, C., Shao, Z., Hong, Z., Zhang, S., Tao, D.: Quality-aware part models for occluded person re-identification. IEEE TMM **25**, 3154–3165 (2023)
52. Wang, Q., Qian, X., Fu, Y., Xue, X.: Co-attention aligned mutual cross-attention for cloth-changing person re-identification. In: ACCV. pp. 2270–2288 (2022)
53. Wang, T., Liu, H., Song, P., Guo, T., Shi, W.: Pose-guided feature disentangling for occluded person re-identification based on transformer. In: AAAI. vol. 36, pp. 2540–2549 (2022)
54. Wang, Z., Zhu, F., Tang, S., Zhao, R., He, L., Song, J.: Feature erasing and diffusion network for occluded person re-identification. In: CVPR. pp. 4754–4763 (2022)
55. Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R.: Detectron2. https://github.com/facebookresearch/detectron2 (2019)
56. Xu, B., He, L., Liang, J., Sun, Z.: Learning feature recovery transformer for occluded person re-identification. IEEE TIP **31**, 4651–4662 (2022)

57. Xu, P., Zhu, X.: Deepchange: A long-term person re-identification benchmark with clothes change. In: ICCV. pp. 11196–11205 (2023)
58. Yan, C., Pang, G., Jiao, J., Bai, X., Feng, X., Shen, C.: Occluded person re-identification with single-scale global representations. In: ICCV. pp. 11875–11884 (2021)
59. Yan, Y., Yu, H., Li, S., Lu, Z., He, J., Zhang, H., Wang, R.: Weakening the influence of clothing: Universal clothing attribute disentanglement for person re-identification. In: IJCAI. pp. 1523–1529 (2022)
60. Yang, J., Zhang, J., Yu, F., Jiang, X., Zhang, M., Sun, X., Chen, Y.C., Zheng, W.S.: Learning to know where to see: A visibility-aware approach for occluded person re-identification. In: ICCV. pp. 11885–11894 (2021)
61. Yang, Q., Wu, A., Zheng, W.S.: Person re-identification by contour sketch under moderate clothing change. IEEE TPAMI **43**(6), 2029–2046 (2021). https://doi.org/10.1109/tpami.2019.2960509
62. Yang, Z., Lin, M., Zhong, X., Wu, Y., Wang, Z.: Good is bad: Causality inspired cloth-debiasing for cloth-changing person re-identification. In: CVPR. pp. 1472–1481 (2023). https://doi.org/10.1109/CVPR52729.2023.00148
63. Yang, Z., Zhong, X., Zhong, Z., Liu, H., Wang, Z., Satoh, S.: Win-win by competition: Auxiliary-free cloth-changing person re-identification. IEEE TIP **32**, 2985–2999 (2023)
64. Zhang, G., Liu, J., Chen, Y., Zheng, Y., Zhang, H.: Multi-biometric unified network for cloth-changing person re-identification. IEEE TIP **32**, 4555–4566 (2023)
65. Zhao, C., Lv, X., Dou, S., Zhang, S., Wu, J., Wang, L.: Incremental generative occlusion adversarial suppression network for person reid. IEEE TIP **30**, 4212–4224 (2021)
66. Zhao, Z., Liu, B., Lu, Y., Chu, Q., Yu, N., Chen, C.W.: Joint identity-aware mixstyle and graph-enhanced prototype for clothes-changing person re-identification. IEEE TMM **26**, 3457–3468 (2024)
67. Zheng, R., Gao, C., Sang, N.: Viewpoint transform matching model for person re-identification. Neurocomputing **433**, 19–27 (2021). https://doi.org/10.1016/j.neucom.2020.12.100
68. Zhihui, Z., Jiang, X., Zheng, F., Guo, X., Huang, F., Sun, X., Zheng, W.: Viewpoint-aware loss with angular regularization for person re-identification. AAAI **34**, 13114–13121 (2020). https://doi.org/10.1609/aaai.v34i07.7014
69. Zhong, Y., Wang, X., Zhang, S.: Robust partial matching for person search in the wild. In: CVPR. pp. 6827–6835 (2020)
70. Zhou, K., Yang, Y., Cavallaro, A., Xiang, T.: Omni-scale feature learning for person re-identification. In: ICCV. pp. 3701–3711 (2019). https://doi.org/10.1109/ICCV.2019.00380
71. Zhuo, J., Chen, Z., Lai, J., Wang, G.: Occluded person re-identification. In: ICME. pp. 1–6 (2018)