DIFFER: Disentangling Identity Features via Semantic Cues for Clothes-Changing Person Re-ID

Xin Liang Yogesh S Rawat
Center for Research in Computer Vision, University of Central Florida
xin.liang@ucf.edu, yogesh@ucf.edu

Abstract

Clothes-changing person re-identification (CC-ReID) aims to recognize individuals under different clothing scenarios. Current CC-ReID approaches either concentrate on modeling body shape using additional modalities including silhouette, pose, and body mesh, potentially causing the model to overlook other critical biometric traits such as gender, age, and style, or they incorporate supervision through additional labels that the model tries to disregard or emphasize, such as clothing or personal attributes. However, these annotations are discrete in nature and do not capture comprehensive descriptions.

In this work, we propose DIFFER: Disentangle Identity Features From Entangled Representations, a novel adversarial learning method that leverages textual descriptions to disentangle identity features. Recognizing that image features inherently mix inseparable information, DIFFER introduces NBDetach, a mechanism designed for feature disentanglement by leveraging the separable nature of text descriptions as supervision. It partitions the feature space into distinct subspaces and, through gradient reversal layers, effectively separates identity-related features from non-biometric features. We evaluate DIFFER on 4 different benchmark datasets (LTCC, PRCC, CelebreID-Light, and CCVID) to demonstrate its effectiveness and provide state-of-the-art performance across all the benchmarks. DIFFER consistently outperforms the baseline method, with improvements in top-1 accuracy of 3.6% on LTCC, 3.4% on PRCC, 2.5% on CelebReID-Light, and 1% on CCVID. Our code can be found here.

1. Introduction

Person re-identification (ReID) [18, 23] is an important task in computer vision, aiming to match individuals across different camera views and environments. ReID plays a significant role in various real-world applications, including surveillance, security, and intelligent transportation systems. Traditional ReID methods primarily rely on visual cues such as appearance and clothing, assuming that individuals wear the same outfit across different observations. However, this assumption is often unrealistic, particularly in real-world settings where people frequently change their clothes. Clothes-changing re-identification (CC-ReID) studies a more challenging problem for real-world situations, where the goal is to recognize individuals despite changes in their clothing. In this work, we focus on CC-ReID to identify the person over a long time period and in a more dynamic environment.

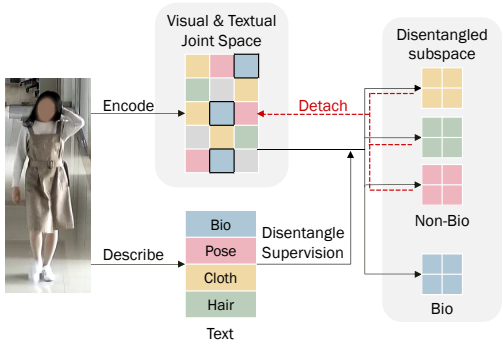

Figure 1. ***Motivation behind DIFFER***. Encoded image features often entangle a multitude of information. By leveraging the separable nature of text descriptions, we guide the model to disentangle the feature space into distinct subspaces. This allows us to detach non-biometric factors and preserve the crucial biometric features.

Existing works on CC-ReID have introduced various strategies to tackle the challenges posed by significant changes in appearance. Many multi-modal approaches rely on external modalities, such as pose [35], silhouette [19], clothing removal [31], or body shape modeling [25], to capture identity-related features. While these methods can improve performance, they introduce additional computational overhead during inference and are often prone to errors due to the dependency on external models, which may fail in complex or unconstrained environments. Moreover, by largely relying on body shape information, these models may not adequately capture other pertinent biometric infor-

mation such as race, gender, and clothing style.

To overcome these challenges, some recent works rely only on RGB images and leverage clothing labels [14] or personal attributes [22, 26, 34] to learn biometric features. Such discrete annotations are useful for learning biometric features, but certain non-biometric factors, such as clothing, hairstyles, or postures, which exhibit infinite variability, are not amenable to precise categorical labeling. Furthermore, relying only on the elimination of clothes information from the encoded feature could let the model neglect other distracting factors such as hairstyle, pose, or environment. In this work, we aim to address some of these limitations by proposing a method that disentangles biometric and non-biometric features using descriptive pseudo-labels without relying on additional external modalities during inference, enhancing flexibility.

We propose DIFFER, a framework leveraging multi-modal foundation models to disentangle non-biometric features from identity-specific representations. DIFFER builds on the capabilities of visual-language models (VLMs) to align visual and language semantics, effectively separating biometric and non-biometric features through separable textual descriptions generated from a VLM. Our proposed NBDetach module uses gradient reversal to remove non-biometric information during training, preserving only identity-specific features. Importantly, semantic descriptions are only needed during training and are not required at inference, enhancing practicality. Evaluated on four benchmark datasets for clothes-changing person re-identification (CC-ReID), DIFFER consistently outperforms baseline methods without needing additional modalities, maintaining both flexibility and robustness.

The contributions of this work are as follows: (1) We propose DIFFER, a novel multimodal approach for disentangling non-biometric information while preserving biometric features via the guidance of separated semantic descriptions. (2) We introduce the use of semantic descriptions as pseudo-labels for biometric and non-biometric attributes, generated using VLMs. (3) We develop NBDetach, a gradient reversal-based technique for feature disentanglement that enables the model to ignore non-biometric features. (4) We conduct extensive evaluations across multiple datasets to validate the robustness and scalability of DIFFER.

2. Related Works

Clothes Changing Person Re-Identification. CC-ReID proposes a unique problem to identify the same person in the long term, various outfits, and dynamic environments. An increasing amount of research contributes to this problem, which can be roughly categorized into single-modality and multi-modality methods.

Methods using only RGB modality usually introduce additional supervision signals to extract the cloth-irrelevant features, such as clothing labels [4, 14, 16, 45], person attributes [34] and descriptions [26]. Multi-modality methods usually introduce human parsing [10, 15, 31], 3D shapes [5, 25], or silhouette [19, 33] to enforce the model to ignore the cloth-irrelevant information and focus on the body shape. These methods transform the problem from eliminating interference factors into cross-modality feature alignment, but these additional modalities may eliminate crucial visual information such as age, gender, or style.

Foundation Models. Foundation Models are large-scale models that are pre-trained on vast amounts of diverse data and can be adapted to a wide range of downstream tasks through fine-tuning or prompt-based techniques. Large language models (LLMs), such as GPT4 [1], have demonstrated impressive generalization across a wide range of natural language processing tasks. Vision-language models (VLM) are pre-trained to match image and language on large scales of data, allowing them to perform tasks involving both texts and images in a joint feature space, such as LLaVA [28], InternVL [6] and CogVLM [42]. CLIP (Contrastive Language-Image Pretraining) models [36] employ contrastive learning to learn joint representations of images and texts. For example, EVA-CLIP-18B [39] successfully scales up to 18 billion parameters and is the largest and most powerful open-source CLIP model to date.

Extensive research has successfully applied the pre-trained VLMs to fine-grained image re-identification tasks, including [18, 23, 44]. However, some existing methods failed to utilize the description ability of VLM from DIFFERent perspectives and focus only on one aspect of the VLM's outputs, such as body shape [26]; other methods[23] require prompt learning and a two-stage training process, limiting their efficiency and flexibility.

Feature disentanglement. Feature disentanglement [43] aims to isolate and separate distinct factors within the embedded representation space. Traditional methods often rely on generative models like VAEs and GANs [29] to split the latent space into various generative factors, aiding in structured factor separation. Recent advancements have extended feature disentanglement beyond generation tasks to areas such as image classification [12, 37], natural language processing [8, 46], and multimodal applications [30]. Techniques vary widely, including multi-task adversarial networks that use min-max optimization for simultaneous encoder and style discriminator training [29], metric learning-based approaches like DFR [7] utilizing gradient reversal layers [13], and orthogonal linear projections for separating visual and textual features in CLIP [30].

In person re-identification (Re-ID), feature disentanglement has been effectively utilized to separate identity from non-identity factors [2–4, 24]. Some Re-ID methods integrate a generative task to separately learn identity and non-

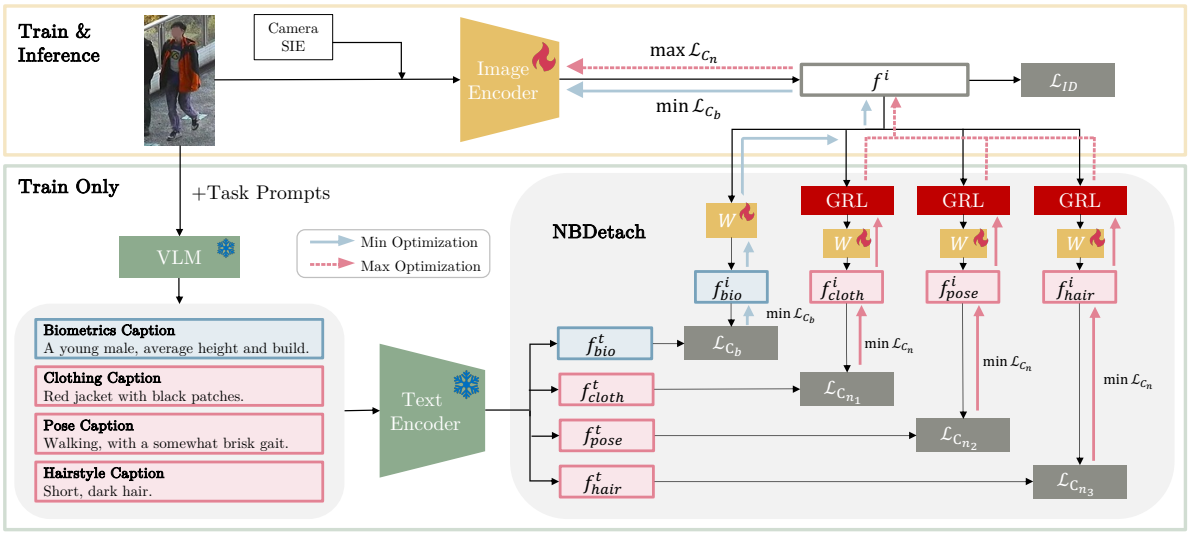

Figure 2. ***Overview of DIFFER:*** First, we use VLM with DIFFERent task prompts to generate DIFFERent textual captions in the perspective of biometrics and non-biometrics factors, which are encoded as textual features, $f^t_b, f^t_{n_1}, ..., f^t_{n_N}$. Next, the image is encoded as image feature f^i in an entangled feature space, and camera side information embedding (Camera SIE) is added during positional embedding. Finally, **NBDetach** is proposed to detach the non-biometric information and preserve the biometric information. Fully linear projection layers W are used to project the image feature to DIFFERent subspaces. Then **GLR** (Gradient Reverse Layer) is added to disentangle the non-biometric information from the entire feature space supervised by non-biometric contrastive losses $\mathcal{L}_{C_n}$.

identity features [4], while others in domains like vehicle Re-ID use networks like DFLNet [3] to learn orientation-specific and general features simultaneously. For occluded person Re-ID, ProFD [9] uses text prompts to disentangle body part features. In clothes-changing Re-ID, dual-stream frameworks have been adopted to isolate identity features from appearance changes and camera variations [24]. However, such methods often require generative models [4] or human annotations [3], highlighting the need for more flexible and annotation-free disentanglement techniques in Re-ID. So we propose DIFFER which uses text descriptions from VLM to guide the model in disentangling the biometric feature from the visual feature space.

3. Methodology

3.1. Overview

Given an input image, our goal is to learn a representation that can reliably identify an individual across variations in clothing, poses, environments, and hairstyles. To achieve this, we propose DIFFER: Disentangle Identity Features From Entangled Representations via separated semantic descriptions. DIFFER isolates identity-related (biometric) features from non-identity (non-biometric) factors, as shown in Fig. 2. It utilizes a multi-modal approach where semantic information from textual descriptions is used to disentangle biometric and non-biometric information. DIFFER's architecture comprises three main components: (1) a visual encoder, (2) semantic learning, and (3) identity feature disentanglement.

A person's image $\mathbf{x}$ is first processed by the visual encoder, producing image features $\mathbf{f}^i$ used for identification. To extract semantic features, we generate biometric and non-biometric descriptions from $\mathbf{x}$ with corresponding prompts from a pre-trained visual-language model (VLM). The resulting textual descriptions are then encoded by a text encoder $\mathcal{T}$, producing textual features $\mathbf{f}^t_b$ for biometrics and $\mathbf{f}^t_{n_1}, ..., \mathbf{f}^t_{n_N}$ for N distinct non-biometric factors.

To further disentangle non-biometric factors, we introduce the non-biometric detach module (NBDetach), which detaches non-biometric components while preserving biometric information. NBDetach projects image features $\mathbf{f}^i$ into separate visual biometric features $\mathbf{f}^i_b$ and visual non-biometric features $\mathbf{f}^i_{n_1}, ..., \mathbf{f}^i_{n_N}$, aligning them with corresponding semantic features using contrastive losses $\mathcal{L}_C$. The gradient reversal layers are added before the non-biometric projection heads, serving as a forgetting mechanism that helps the image encoder remove non-biometric information. During inference, only the visual encoder is needed to obtain biometric features, while the text encoder and NBDetach module are not required.

3.2. Visual encoder

The visual encoder $\mathcal{I}$ is responsible for extracting visual features from the input image. It encodes the image to generate an entangled representation that contains a mixture of information, including biometric, non-biometric and envi-

Example Conversation with VLM

Biometric Description

Input: *Describe the individual's overall **physical appearance**, including estimated age, gender, height (e.g., short, average, tall based on surroundings if applicable), and build(e.g., slender, average, robust).*
VLM: *The individual appears to be a young male, possibly in his late teens to early twenties. He has short, dark hair and glasses. Based on the surroundings, he seems to be of average height and build.*

Non-Biometric Description

Input: *Detail the type of **clothing** the person is wearing(the style, colors, and any visible logos or patterns), shoes and any accessories (e.g., glasses, watches, jewelry).*
VLM: *The individual is wearing a red jacket with black and white patches. Underneath the jacket, he has a blue shirt with a white logo or emblem on it. He is also wearing purple pants and black and white shoes. He is wearing glasses.*

ronmental information. This feature representation serves as the foundation for further disentanglement and identification. To facilitate disentanglement using semantic descriptions, it is important that the latent feature space of visual encoder aligns with the text encoder. We use ViT, a transformers-based visual encoder from a multimodal model that has been pre-trained on image-language pairs [39]. The classification token of ViT is used to represent image features $\mathbf{f}^i$.

3.3. Semantic learning

Existing clothing-changing ReID methods usually adopt a supervised method to guide the model in learning the desired biometric information, such as the 3D body shape, silhouettes, or personal attributes, while forgetting unnecessary non-biometric information. However, some non-biometrics labels are hard to acquire since we cannot categorize the infinite possibilities into finite categories, such information including varying clothing styles, poses, or hairstyles. Besides, it is hard to differentiate the image feature encoded by a neural network from different factors in an embedded space. These difficulties have motivated us to utilize the strength of a large visual-language foundation model, which has the ability to describe a person's appearance in an image, and it is easy to separate the text descriptions into different aspects.

In this study, we adopt a context-objective framework to generate our task prompts for the biometric and non-biometric information with a large language model GPT4 [1]. Then we use an open VLM model, CogVLM [42], to extract the text captions for each image in the training dataset. Our biometric information includes body build, age, gender, etc. This is included in one sentence and is later encoded by a text encoder $\mathcal{T}$ as a biometrics text feature $\mathbf{f}_b^t$. Moreover, the non-biometric descriptions can include hairstyle, clothing, pose, and environment captions with specially designed prompts to extract different information from the image (examples in supplementary material). The non-biometric textual feature is encoded as $\mathbf{f}_n^t$ by the text encoder $\mathcal{T}$. Our task prompts and example descriptions for biometrics and non-biometrics corresponding to the image in Fig. 2 are shown in the text box above.

After generating all the text descriptions, we use a summarization prompt to combine all the biometric descriptions for one individual into one summarising text. So all the images belonging to the same individual should have the same biometric text feature. GPT4 is used to summarize the textual descriptions.

3.4. NBDetach: Feature disentanglement

To detach the non-biometric information from the image feature space while preserving biometric information, we propose a Non-Biometric Detach (**NBDetach**) module. The aim is to preserve the biometric features and ignore the non-biometric features. It relies on semantic features from the textual descriptions to guide this disentanglement. After the image $\mathbf{x}$ is encoded as the image feature $\mathbf{f}^i$ by the image encoder $\mathcal{I}$, we project the image feature into different subspaces to match the biometric and N different non-biometric textual features. The resulting image features are denoted as $\mathbf{f}_b^i$ for biometric feature and $\mathbf{f}_{n_k}^i$ for non-biometric features, where $k \in \{1, ..., N\}$. The biometric projection head is denoted as H_b, and the non-biometric projection heads are denoted as H_{n_k}, where linear projection layers W are used to perform the projections. We use an image-to-text contrastive loss [36] between textual and visual features for both biometric and non-biometric features.

$$\mathcal{L}_{C_h} = -\frac{1}{B}\sum_{k=1}^{B}\log\frac{\exp(\cos(\mathbf{f}_{h,k}^i, \mathbf{f}_{h,k}^t))}{\sum_{j=1}^{B}\exp(\cos(\mathbf{f}_{h,j}^i, \mathbf{f}_{h,j}^t))}, \quad (1)$$

where h is either the biometric $\mathcal{L}_{\mathrm{C}_b}$ or non-biometric $\mathcal{L}_{\mathrm{C}_n}$, B is the batch size, $\mathbf{f}_{h,k}^i$ is the projected image feature for instance k, and $\mathbf{f}_{h,k}^t$ is the corresponding text feature in domain h. Our goal is to preserve biometric information in the

image feature $\mathbf{f}^i$ and forget non-biometric information. A contrastive loss will help in preserving biometric information in the image features $\mathbf{f}^i$, but we want to disentangle the non-biometric information. Therefore we rely on gradient reversal for non-biometric features, which will help detach the non-biometric information from image feature $\mathbf{f}^i$. The training objectives are formulated as,

$$\min_{\mathcal{I}, H_b} \mathcal{L}_{\mathrm{C}_b} \tag{2}$$

$$\max_{\mathcal{I}} \min_{H_n} \mathcal{L}_{\mathrm{C}_n}. \tag{3}$$

As shown above, $\mathcal{I}$ and H_b are optimized jointly to minimize the biometric contrastive loss such that the image feature preserves more biometric information. Conversely, $\mathcal{I}$ and H_n play an adversarial role such that the non-biometric projection tries to extract the corresponding non-biometric information as much as possible, while the image encoder tries to detach the non-biometric information from the encoded image feature space.

To perform the min-max optimizations simultaneously, a gradient reverse layer (GRL) [13] is added between the image encoder $\mathcal{I}$ and the non-biometric projection heads H_n. GRL multiplies the gradient with a negative coefficient to reverse the gradient direction, ensuring that the training objectives before and after applying the layer are opposite. So it is possible to jointly train the encoder and non-biometric heads with opposite objectives.

3.5. Loss Function

ID Loss For the identification loss, we used a standard cross entropy classification loss $\mathcal{L}_{cls}$ and triplet loss $\mathcal{L}_{tri}$. For image feature $\mathbf{f}_i^i$ and its corresponding identity label y_i, the ID loss is calculated as:

$$\mathcal{L}_{cls} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{c=1}^{C} y_{i,c}\log(\hat{y}_{i,c}), \tag{4}$$

$$\mathcal{L}_{tri} = \sum_{i=1}^{B}\left[\max(0,\, euc(\mathbf{f}_i^a, \mathbf{f}_i^p) - euc(\mathbf{f}_i^a, \mathbf{f}_i^n) + \alpha)\right], \tag{5}$$

where B is the batch size, i is the instance index, $y_{i,c}$ is the one hot label of class c, $\hat{y}_{i,c}$ is the softmax probability score for class c. In equation (5), $euc(\mathbf{x}, \mathbf{y})$ is the euclidean distance between feature $\mathbf{x}$ and $\mathbf{y}$, $\mathbf{f}_i^a$ is the anchor image feature, $\mathbf{f}_i^p$ is the feature of the positive paired instance with the maximum distance for all the positive instances in the batch, and $\mathbf{f}_i^n$ is the feature of the negative paired instance with the minimum distance for all the negative instances in the batch, and α is the margin for the triplet loss. The total ID loss is a linear combination of the classification and triplet loss with control coefficients λ_c and λ_t:

$$\mathcal{L}_{ID} = \lambda_c\mathcal{L}_{cls} + \lambda_t\mathcal{L}_{tri}. \tag{6}$$

Overall Loss Our total loss is a linear combination of all the losses as formulated below,

$$\mathcal{L}_{total} = \lambda_{id}\mathcal{L}_{ID} + \lambda_b\mathcal{L}_{C_b} + \lambda_n\sum_{k=1}^{N}\mathcal{L}_{C_{n_k}}, \tag{7}$$

where λ's are constant coefficients to control the contribution of each loss function. Without loss of generality, we set all $\lambda = 1$ in all of our experiments.

4. Experiment

Dataset and metrics We train and evaluate our method on four CC-ReID datasets, including LTCC [35], PRCC [40], Celeb-reID-light [20] and CCVID [14]. The LTCC contains 17,119 images of 152 identities across multiple cameras, and the PRCC comprises 33,698 images from 221 identities with two outfits each. Celeb-reID-light, a subset of Celeb-reID, includes 10,842 images of 9,021 individuals from diverse sources. CCVID, a video re-ID dataset with 2,856 sequences and 226 identities, is converted to image re-ID by using 10% of frames for training and the first frame for testing. We evaluate model performance using Rank-1 accuracy and mean average precision (mAP) on both standard and clothing-changing re-identification tasks.

Implementation details We use the EVA02-CLIP-L [38] model of patch size 14 as our visual encoder. Following prior work [17], we use camera encoding as side information embedding (SIE) during positional encoding to capture the difference of viewpoints. CogVLM [42] is used to extract the text description of the person in the image. We use GPT4 [1] to generate the language prompts for each task and summarize the biometric descriptions from the same class. The input image is resized to 224×224 for all datasets. The batch size is set to 8 for LTCC and 16 for other datasets. The base learning rate is set as 2×10^{-6}. A warming-up learning strategy with an initial learning rate of 8.42×10^{-7} and the cosine learning rate decay are used to train 60 epochs. The SGD optimizer is employed in the optimization process. Moreover, the weight decay for the experiment is 0.05. Two Nvidia Quadro RTX 5000 GPUs are used for training and testing.

For loss function, DIFFER uses ID loss, biometric contrastive loss, and non-biometric contrastive loss in all the experiments unless specified. For the selection of non-biometric factors, the specific factors chosen for each dataset are as follows: in LTCC and PRCC, hair and clothing are used; Celeb-reID-L uses hair and pose, while CCVID focuses on clothing alone. These selections help ensure that DIFFER effectively disentangles relevant non-biometric features across varied dataset conditions.

Method	LTCC		PRCC		CCVID		Celeb-L	
	top1	mAP	top1	mAP	top1	mAP	top1	mAP
Baseline	54.6	30.4	65.1	63.0	86.9	86.7	72.7	53.3
DIFFER	**58.2**	**31.6**	**68.5**	**64.7**	**87.9**	**87.4**	**75.6**	**54.3**

Table 1. Comparison DIFFER with baseline method across different datasets. The LCTT, PRCC, and CCVID are all evaluated under the clothes-changing setting. Celeb-L stands for Celeb-reID-light dataset.

4.1. Comparison with the baseline method

We compare DIFFER with a baseline method that uses the same backbone architecture but is trained solely with ID loss, i.e. DIFFER without the NBDetach module. We evaluate the performance on four benchmark datasets: LTCC, PRCC, CCVID under cloth changing setting and Celeb-reID-light in Tab. 1. For LTCC, DIFFER achieves a top-1 accuracy improvement from 54.6% to 58.2% and an increase in mAP from 30.4% to 31.6%. Similarly, in PRCC, DIFFER raises top-1 accuracy from 65.1% to 68.5% and mAP from 63.0% to 64.7%. Celeb-reID-light and CCVID also show notable improvements, with top-1 and mAP gains for both datasets. These results highlight its effectiveness DIFFER in enhancing re-identification performance.

4.2. Ablations and analysis

We perform ablations on different model architectures, loss terms, VLMs, and varying non-biometric factors, along with subspace dimensions.

4.2.1. Model Architectures

Tab. 2 presents the performance comparison between the baseline method and our proposed DIFFER approach across four model architectures: ClipViT-B/16, ClipViT-L/14[36], EVA2-CLIP-B/16, and EVA2-CLIP-L/14[38]. We evaluate the models on LTCC under two settings. Across all model configurations, DIFFER consistently outperforms the baseline methods. DIFFER can increase the CLIPViT model top1 accuracy by 2% and more than 3% under the CC setting. The EVA2-CLIP-L model with DIFFER yields the highest performance, demonstrating the strength of our approach when combined with larger models. This overall trend indicates that DIFFER is highly effective at enhancing CC-ReID performance across varied model architectures.

4.2.2. Loss functions

We study the effect of different loss functions on PRCC and LTCC datasets under the cloth-changing setting (Tab. 3). The loss functions include identity loss $\mathcal{L}_{ID}$, biometric contrastive loss $\mathcal{L}_{C_b}$, and non-biometric loss $\mathcal{L}_{C_n}$. For the non-biometric loss, hair and clothing attributes are used. As shown in the table, both biometric loss and non-biometric loss improve the performance on both datasets, and the combination of both losses achieves the best accuracy, improving the Rank-1 accuracy from 54.6% to 58.2% on LTCC dataset and, 65.1% to 68.5% on PRCC dataset.

Model	Method	CC		General	
		top1	mAP	top1	mAP
ClipViT-B	Baseline	34.2	15.0	67.5	31.9
	DIFFER	36.5	15.6	72.6	35.2
ClipViT-L	Baseline	42.6	19.5	75.7	38.0
	DIFFER	44.4	20.0	78.3	39.7
EVA2-CLIP-B	Baseline	37.5	18.3	76.9	40.4
	DIFFER	40.6	21.7	74.0	41.0
EVA2-CLIP-L	Baseline	54.6	31.2	81.9	50.4
	DIFFER	**58.2**	**31.6**	**85.0**	**52.8**

Table 2. Different model architectures performance on LTCC under both CC and general setting.

$\mathcal{L}_{ID}$	$\mathcal{L}_{C_b}$	$\mathcal{L}_{C_n}$	LTCC		PRCC	
			top1	mAP	top1	mAP
✓			54.6	30.4	65.1	63.0
✓	✓		55.4	30.8	66.7	63.6
✓		✓	55.6	31.0	67.2	63.8
✓	✓	✓	**58.2**	**31.6**	**68.5**	**64.7**

Table 3. Ablation study of loss functions. The table evaluates the impact of combining different losses on the LTCC and PRCC cloth-changing scenario.

NBF	LTCC		PRCC		Celeb-L	
	top1	mAP	top1	mAP	top1	mAP
None	55.4	30.8	66.7	63.6	74.1	54.7
Hair	**56.6**	31.3	**68.4**	64.0	**75.2**	**55.0**
Cloth	**56.6**	30.4	67.3	**64.2**	74.2	54.8
Pose	56.1	**31.5**	66.5	63.6	75.1	54.9

Table 4. Analysis of different non-biometric factors (NBF). 'None' indicates that only the biometric contrastive loss is used. (Celeb-L: Celeb-reID-light)

4.2.3. Single non-biometric factor

DIFFER could use different non-biometric factors in the NBDetach module, including clothing, pose, and hairstyle. Here, we use a single non-biometric projection head to evaluate which non-biometric factor has the most negative effect on the model performance in Tab. 4. We perform these experiments on LTCC, PRCC, and Celeb-reID-light under the clothes-changing setting. As shown in Tab. 4, different non-biometric factors play various roles in different datasets. Disentangling factors like hair or cloth can significantly improve the model's performance on the LTCC and PRCC datasets, whereas hairstyle and pose play a slightly more important role in the Celeb-reID-light dataset. So the choice of non-biometric features should be dataset-specific and different selections of non-biometric factors should be used in different user cases.

Hair	Cloth	Pose	1024 dim		512 dim	
			top1	mAP	top1	mAP
✓			55.8	29.5	56.6	31.3
	✓		**57.1**	31.4	56.6	30.4
		✓	**57.1**	30.4	56.1	31.5
✓	✓		54.1	**32.0**	**58.2**	**31.6**
	✓	✓	56.1	31.3	56.6	31.5
✓		✓	56.1	31.8	57.4	31.4
✓	✓	✓	55.1	31.0	57.7	31.2

Table 5. Performance comparison with multiple non-biometric factors and subspace feature dimensions. The results are evaluated on the LTCC dataset under the clothes-changing setting.

VLM Model		LTCC	
Model	# Parameters	Rank-1	mAP
LLama1.6	7B	57.9	31.7
CogVLM	18B	**58.2**	31.6
InternVL2.5	26B	57.7	**32.6**

Table 6. Experiment on different VLMs. The performance is evaluated on LTCC under the clothes-changing setting.

4.2.4. Multiple non-biometric factors

We explore whether incorporating multiple non-biometric factors could enhance the model's performance on LTCC in Tab. 5. Our results reveal that the effect of multiple non-biometric factors depends on the relationship between the image feature dimension and the biometric/non-biometric subspace dimensions. When the image feature dimension (1024) matches the subspace dimensions (1024), using multiple non-biometric factors reduces performance, with Rank-1 accuracy dropping from 57.1% to 54.1%. Conversely, when the biometric and non-biometric subspaces are smaller than the image feature dimension (512 vs. 1024), adding multiple factors enhances Rank-1 accuracy from 56.1% to 58.2%. While higher-dimensional textual feature spaces offer greater representational capacity, they do not necessarily facilitate effective feature disentanglement. This aligns with prior findings [30], which show that optimal visual and textual feature disentanglement is achieved in lower-dimensional subspaces. Furthermore, adding all three non-biometric factors reduces performance even in the 512-dimensional subspace, likely because this dimension is too large to effectively disentangle the 1024-dimensional space into three subspaces.

4.2.5. Experiment using different VLMs

We conducted experiments using different VLMs, including LLama1.6 [27] and InternVL2.5 [6], in addition to CogVLM-18B in Tab. 6. The results indicate that all three models show similar performance, suggesting that these large-scale visual-language models are equally good at giving low-level person appearance descriptions.

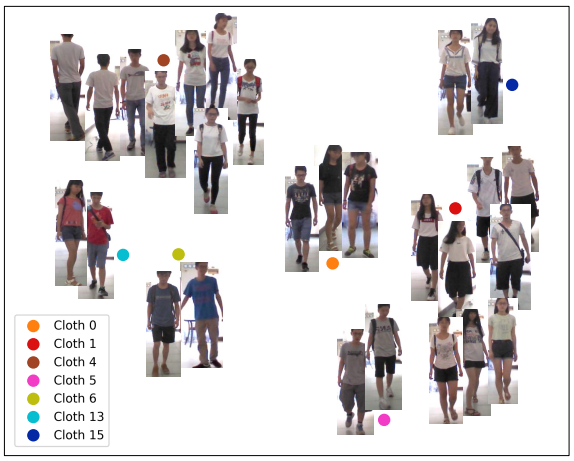

Figure 3. Clothing textual feature cluster visualization results. We visualize the cluster results of different images based on the clothing textual feature. As shown in the figure, images with similar outfits are successfully grouped together.

4.3. Comparison with state-of-the-art methods

We compare DIFFER with existing CNN-based clothing-changing person REID methods, including 3DSL [5], GI-ReID [21], CAL [14], AIM [45], LDF [4], 3DInvarReID [25], SCNet [15], CVSL [32], CLIP3DReID [26], CCPG [33], and ViT based methods including TransReID [17], Instruct-ReID[18], MADE [34]. We evaluate the performance on two datasets, PRCC and LTCC, in Tab. 7.

As shown in the table, DIFFER outperforms other methods across most evaluation metrics in both datasets, benefiting from the robust capability of the multimodal foundation model and the proposed NBDetach module. For the PRCC dataset, DIFFER achieves the highest top-1 accuracy (68.5%) and mAP (64.7%) in the cloth-changing scenario; in the LTCC dataset, DIFFER again leads with top-1 accuracy of 58.2% and mAP of 31.6% in the cloth-changing scenario, demonstrating its effectiveness in long-term clothes changing re-identification tasks.

4.4. Effectiveness of textual features

To demonstrate the effectiveness of the textual features, we visualize the textual feature clustering results in Fig. 3. We use the clothes textual feature for PRCC datasets and cluster the clothing feature with the DBSCAN [11] cluster algorithm to label different clothes with different IDs. Then t-SNE [41] is used to visualize the cluster results and we overlay the corresponding images on top of the feature representation. As shown in Fig. 3, the clothing textual feature successfully clusters the outfits in the image regardless of gender or body shape. The results demonstrate that our clothing textual features could measure the outfit similarity between different people, while previous clothes labels only focused on differentiating the clothes for one individual.

Method	Venue/Year	Modality	Backbone	PRCC				LTCC			
				CC		SC		CC		General	
				top1	mAP	top1	mAP	top1	mAP	top1	mAP
3DSL[5]	CVPR 21	RGB+pose+sil.+3D	ResNet-50x2	-	51.3	-	-	31.2	14.8	-	-
GI-ReID [21]	CVPR 22	RGB+pose	OSNet	-	37.5	-	-	28.11	13.17	63.2	29.4
CAL [14]	CVPR 22	RGB	ResNet-50	55.2	55.8	**100**	99.8	40.1	18	74.2	40.8
AIM [45]	CVPR 23	RGB	ResNet-50x2	57.9	58.3	**100**	**99.9**	40.6	19.1	76.3	41.1
LDF [4]	ACM 23	RGB	ResNet-50	58.4	58.6	**100**	99.7	32.9	15.4	73.4	-
3DInvarReID [25]	ICCV 23	RGB+3D	-	56.5	57.2	-	-	40.9	18.9	-	-
SCNet [15]	ACM 23	RGB+parsing/ RGB(Inference)	ResNet-50	58.7	59.9	**100**	97.8	47.5	25.5	76.3	43.6
CVSL [32]	WACV 24	RGB + pose	ResNet-50	57.5	56.9	97.5	99.1	44.5	21.3	76.4	41.9
CLIP3DReID [26]	CVPR 24	RGB+desc.+3D/ RGB(Inference)	ResNet-50	60.6	59.3	-	-	42.1	21.7	-	-
CGPG [33]	CVPR 24	RGB+sil.	ResNet-50	61.8	58.3	**100**	99.6	46.2	22.9	77.2	42.9
TransReID[17]	CVPR 21	RGB	ViT-B	46.6	44.8	**100**	99	34.4	17.1	70.4	37.0
Instruct-ReID[18]	CVPR 24	RGB + instruct	ViT-B+ALBEF	54.2	52.3	-	-	-	-	75.8	52.0
MADE [34]	IEEE TMM 24	RGB+attributes	EVA2-CLIP-L	64.3	59.1	**100**	98.6	47.4	24.4	84.2	48.2
DIFFER		RGB+description/ RGB(Inference)	EVA2-CLIP-L	**68.5**	**64.7**	99.9	99.5	**58.2**	**31.6**	**85.0**	**52.8**

Table 7. Comparison of state-of-the-art methods on PRCC and LTCC datasets. CC stands for clothes-changing, SC stands for same-clothes, and General stands for both CC and SC. The best results are highlighted with bold and the second best results are underlined.

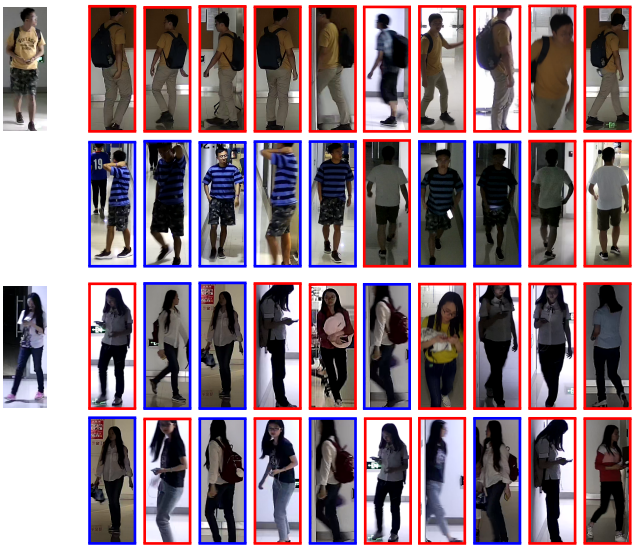

Figure 4. The Top-10 retrieval results from LTCC dataset with baseline and DIFFER. Each row displays the ranked results for a query image (left), followed by Rank-1 to Rank-10 retrieval images from left to right. The first row shows baseline results, and the second shows DIFFER results, with correct matches in blue boxes and incorrect matches in red.

4.5. Retrieval result visualization

We visualize the top 10 matched gallery images for some query images from LTCC in Fig. 4. As illustrated, our model successfully identifies individuals despite variations in clothing and pose, while the baseline method struggles to differentiate between individuals with similar clothing and hairstyles, even when body shapes differ. However, we observe that DIFFER encounters large incorrect matches in the second sample, where most retrieved results share similar body shapes, hairstyles, and occlusion. This indicates the model still faces challenges with occlusions and visual similarities in body angle and shape. Incorporating finer-grained biometric features could improve its ability to distinguish between visually similar individuals.

5. Conclusion

In this work, we propose DIFFER, a novel method for clothes-changing person re-identification that disentangles biometric and non-biometric features using a multimodal approach. By leveraging pseudo-labels generated from large visual-language models, DIFFER effectively separates identity-related information from non-biometric attributes. Unlike many existing approaches, DIFFER does not rely on additional external modalities during inference, making it more reliable and flexible. Our extensive evaluation on multiple benchmark datasets demonstrates that DIFFER consistently outperforms the baseline method.

Limitations While DIFFER shows strong performance in disentangling non-biometric features, limitations remain. Combining multiple non-biometric aspects can destabilize the gradient reversal process, but future advancements in large VLMs with extended context lengths may enable us to input all non-biometric descriptions simultaneously, enhancing stability. We also recognize the ethical implications of person re-identification (ReID), as misuse could lead to privacy violations. We are committed to responsible development and align with the ReID community in advocating for safeguards against misuse.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2, 4, 5

[2] 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*, pages 287–296, 2024. 2

[3] Yan Bai, Yihang Lou, Yongxing Dai, Jun Liu, Ziqian Chen, and Ling-Yu Duan. Disentangled Feature Learning Network for Vehicle Re-Identification. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 474–480, Yokohama, Japan, 2020. International Joint Conferences on Artificial Intelligence Organization. 3

[4] Patrick PK Chan, Xiaoman Hu, Haorui Song, Peng Peng, and Keke Chen. Learning disentangled features for person re-identification under clothes changing. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(6):1–21, 2023. 2, 3, 7, 8

[5] 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*, pages 8146–8155, 2021. 2, 7, 8

[6] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 2, 7

[7] Hao Cheng, Yufei Wang, Haoliang Li, Alex C Kot, and Bihan Wen. Disentangled feature representation for few-shot image classification. *IEEE transactions on neural networks and learning systems*, 2023. 2

[8] Pengyu Cheng, Martin Renqiang Min, Dinghan Shen, Christopher Malon, Yizhe Zhang, Yitong Li, and Lawrence Carin. Improving disentangled text representation learning with information-theoretic guidance. *arXiv preprint arXiv:2006.00693*, 2020. 2

[9] Can Cui, Siteng Huang, Wenxuan Song, Pengxiang Ding, Min Zhang, and Donglin Wang. Profd: Prompt-guided feature disentangling for occluded person re-identification. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1583–1592, 2024. 3

[10] 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. 2

[11] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, pages 226–231, 1996. 7

[12] Zeyu Feng, Chang Xu, and Dacheng Tao. Self-supervised representation learning by rotation feature decoupling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10364–10374, 2019. 2

[13] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. 2, 5

[14] 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. 2, 5, 7, 8

[15] Peini Guo, Hong Liu, Jianbing Wu, Guoquan Wang, and Tao Wang. Semantic-aware consistency network for cloth-changing person re-identification. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 8730–8739, 2023. 2, 7, 8

[16] 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

[17] 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. 5, 7, 8

[18] 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, 2023. arXiv:2306.07520 [cs]. 1, 2, 7, 8

[19] Peixian Hong, Tao Wu, Ancong Wu, Xintong Han, and Wei-Shi Zheng. Fine-grained shape-appearance mutual learning for cloth-changing person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10513–10522, 2021. 1, 2

[20] Yan Huang, Qiang Wu, Jingsong Xu, and Yi Zhong. Celebrities-reid: A benchmark for clothes variation in long-term person re-identification. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019. 5

[21] 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*, pages 14278–14287, 2022. 7, 8

[22] Kyung Won Lee, Bhavin Jawade, Deen Mohan, Srirangaraj Setlur, and Venu Govindaraju. Attribute de-biased vision transformer (ad-vit) for long-term person re-identification. In *2022 18th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8. IEEE, 2022. 2

[23] Siyuan Li, Li Sun, and Qingli Li. CLIP-ReID: Exploiting Vision-Language Model for Image Re-Identification without Concrete Text Labels, 2022. arXiv:2211.13977 [cs]. 1, 2

[24] Yubo Li, De Cheng, Chaowei Fang, Changzhe Jiao, Nannan Wang, and Xinbo Gao. Disentangling identity features from interference factors for cloth-changing person re-identification. In *ACM Multimedia 2024*, 2024. 2, 3

[25] 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*, pages 19617–19626, 2023. 1, 2, 7, 8

[26] 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*, pages 256–266, 2024. 2, 7, 8

[27] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 7

[28] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 2

[29] Yang Liu, Zhaowen Wang, Hailin Jin, and Ian Wassell. Multi-task Adversarial Network for Disentangled Feature Learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3743–3751, Salt Lake City, UT, USA, 2018. IEEE. 2

[30] Joanna Materzynska, Antonio Torralba, and David Bau. Disentangling visual and written concepts in CLIP. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16389–16398, New Orleans, LA, USA, 2022. IEEE. 2, 7

[31] Jingyi Mu, Yong Li, Jun Li, and Jian Yang. Learning clothes-irrelevant cues for clothes-changing person re-identification. In *BMVC*, page 337, 2022. 1, 2

[32] 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*, pages 1041–1049, 2024. 7, 8

[33] 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*, pages 7541–7549, 2024. 2, 7, 8

[34] Chunlei Peng, Boyu Wang, Decheng Liu, Nannan Wang, Ruimin Hu, and Xinbo Gao. Masked attribute description embedding for cloth-changing person re-identification. *IEEE Transactions on Multimedia*, 2024. 2, 7, 8

[35] Xuelin Qian, Wenxuan Wang, Li Zhang, Fangrui Zhu, Yanwei Fu, Tao Xiang, Yu-Gang Jiang, and Xiangyang Xue. Long-term cloth-changing person re-identification. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 1, 5

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 4, 6

[37] Eduardo Hugo Sanchez, Mathieu Serrurier, and Mathias Ortner. Learning disentangled representations via mutual information estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 205–221. Springer, 2020. 2

[38] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 5, 6

[39] Quan Sun, Jinsheng Wang, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, and Xinlong Wang. EVA-CLIP-18B: Scaling CLIP to 18 Billion Parameters, 2024. arXiv:2402.04252 [cs]. 2, 4

[40] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European conference on computer vision (ECCV)*, pages 480–496, 2018. 5

[41] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 7

[42] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. Cogvlm: Visual expert for pretrained language models, 2023. 2, 4, 5

[43] Xin Wang, Hong Chen, Si'ao Tang, Zihao Wu, and Wenwu Zhu. Disentangled Representation Learning, 2024. arXiv:2211.11695 [cs]. 2

[44] Shan Yang and Yongfei Zhang. MLLMReID: Multimodal Large Language Model-based Person Re-identification, 2024. arXiv:2401.13201 [cs]. 2

[45] 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*, pages 1472–1481, 2023. 2, 7, 8

[46] Xiongyi Zhang, Jan-Willem van de Meent, and Byron Wallace. Disentangling Representations of Text by Masking Transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 778–791, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. 2