Causality-Inspired Invariant Representation Learning for Text-Based Person Retrieval

Yu Liu[1,2], **Guihe Qin**[1,2], **Haipeng Chen**[1,2*], **Zhiyong Cheng**[3], **Xun Yang**[4]
[1]College of Computer Science and Technology, Jilin University, China
[2]Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, China
[3]Qilu University of Technology (Shandong Academy of Sciences), JiNan, China
[4]University of Science and Technology of China, HeFei, China
yul20@mails.jlu.edu.cn, {qingh, chenhp}@jlu.edu.cn, jason.zy.cheng@gmail.com, xyang21@ustc.edu.cn

Abstract

Text-based Person Retrieval (TPR) aims to retrieve relevant images of specific pedestrians based on the given textual query. The mainstream approaches primarily leverage pre-trained deep neural networks to learn the mapping of visual and textual modalities into a common latent space for cross-modality matching. Despite their remarkable achievements, existing efforts mainly focus on learning the statistical cross-modality correlation found in training data, other than the intrinsic causal correlation. As a result, they often struggle to retrieve accurately in the face of environmental changes such as illumination, pose, and occlusion, or when encountering images with similar attributes. In this regard, we pioneer the observation of TPR from a causal view. Specifically, we assume that each image is composed of a mixture of causal factors (which are semantically consistent with text descriptions) and non-causal factors (retrieval-irrelevant, *e.g.,* background), and only the former can lead to reliable retrieval judgments. Our goal is to extract text-critical robust visual representation (*i.e.*, causal factors) and establish domain invariant cross-modality correlations for accurate and reliable retrieval. However, causal/non-causal factors are unobserved, so we emphasize that ideal causal factors that can simulate causal scenes should satisfy two basic principles: **1) Independence**: being independent of non-causal factors, and **2) Sufficiency**: being causally sufficient for TPR across different environments. Building on that, we propose an Invariant Representation Learning method for TPR (IRLT), that enforces the visual representations to satisfy the two aforementioned critical properties. Extensive experiments on three datasets clearly demonstrate the advantages of IRLT over leading baselines in terms of accuracy and generalization.

Introduction

Text-based Person Retrieval (TPR) (Li et al. 2017) aims to retrieve images of a target person with high semantic relevance to a given linguistic description from a gallery of images. In recent years, there has been a growing interest in TPR (Ding et al. 2021; Suo et al. 2022; Jiang and Ye 2023) since the textual query can provide more natural and comprehensive descriptions of pedestrians in practical applications such as crime search and missing person search. This task is important but also challenging as it requires accurately modeling of the visual-linguistic alignments.

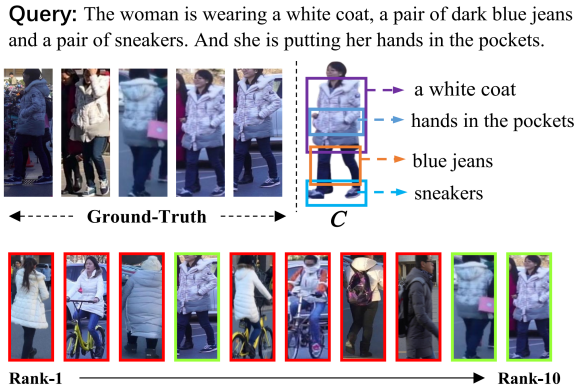

Figure 1: A toy example of the TPR task on the RSTPReid dataset. Matched/Mismatched images are marked by green/red rectangles. "C" denotes the causal cues.

Towards this end, many TPR models have emerged (Zhu et al. 2021; Suo et al. 2022; Chen et al. 2022; Jiang and Ye 2023). They usually employ the identical structural design of "image/text backbone + feature embedding". This design involves first extracting image/text features using the image/text backbone, and then using feature embedding (method-specific) to embed the extracted image and text features into a joint space that enables cross-modal alignment. Most recent methods usually utilize unimodal pre-trained networks (*e.g.*, ResNet (He et al. 2016) and ViT (Dosovitskiy et al. 2021), BERT (Devlin et al. 2018)) to initialize the image/text backbone. In addition, recent studies (Yan et al. 2022; Jiang and Ye 2023) show that Contrastive Language-Image Pre-training (CLIP) model, as a representative work in visual-language pre-training (VLP), has successfully been adopted for the TPR task. These methods leverage the abundant multi-modal correspondence information provided by CLIP, resulting in superior performance compared to unimodal pre-training methods. In summary, recent achievements of TPR should be mainly attributed to the strong ability of pre-trained deep networks in representation learning.

Despite the remarkable achievement in retrieval accuracy, the typical learning objective of recent efforts usually relies

*Corresponding Author.

on the empirical risk minimization (ERM), and thus tends to over-exploit the spurious correlations between textual query and person images due to the inherent dataset bias while easily ignoring the intrinsic cross-modality causal correlation. As a result, existing efforts may easily suffer from unreliable person retrieval. For instance, as depicted in Fig.1, we find that the first two images in Ground-Truth cannot be retrieved correctly, since they have significant illumination and pose variations from the other three images. Most ERM-based works are usually unable to make correct judgments due to the change of non-causal factors. In addition, we also observe that most of the incorrect results in Fig.1 have the attribute "white" which is similar to Ground-Truth. We know that deep feature learning follows a low-level to high-level paradigm. ERM-based statistical models typically stop learning when they learn "simple" features (*e.g.*, "white") at low layers that are sufficient to minimize the loss, losing critical information (*e.g.*, "hands in the pockets" and "blue jeans") at high layers to distinguish similar pedestrians. *How to develop a robust text-based person retrieval method that can effectively capture the causally invariant visual-linguistic correlations is the critical research question, but receiving less attention so far.*

In this paper, taking a causal look at TPR (Pearl, Glymour, and Jewell 2016), we split the visual scenes into two parts: 1) the causal factors, which contains the text-critical cues (*e.g.*, "white coat", "blue jeans", and "sneakers"), and 2) the non-causal components, which is retrieval-irrelevant (*e.g.*, "illumination style", "pose", and "occlusion"). Our goal is to extract semantically consistent causal factors with text descriptions from raw input images as the visual representation of pedestrians. Unfortunately, the absence of annotations for causal/non-causal factors makes causal reasoning particularly challenging. To uncover the stable causal relationship between text descriptions and person images, we emphasize that the causal factors should satisfy two properties: **1) Independence**: being independent from the non-causal factors, *i.e.*, being robust to the change of domain-specific factors (*e.g.*, *background* and *illumination*), and **2) Sufficiency**: capturing causally sufficient information for stable retrieval. Drawing inspiration from causal inference (Krueger et al. 2021), we propose an Invariant Representation Learning for TPR (IRLT) method to enforce the model to learn robust person representation that satisfies the two properties. IRLT equips the existing TPR backbone model with two additional modules: a style intervener and a scene simulator. Specifically, the style intervener simulates the variation of non-causal components by modeling the uncertainty of features, forcing the model to learn causally independent representation. In addition, the scene simulator places pedestrian images in similar and dissimilar environments, ensuring the sufficiency of retrieval by discovering subtle but discriminative causal factors. Note that our IRLT is model-agnostic and can be compatible with existing TPR methods flexibly.

Our technical contributions are briefly summarized as:

- We investigate the TPR task from the causal view, which aims to learn causally-invariant person representation for not only accurate but also reliable cross-modal retrieval.
- We propose a novel and model-agnostic IRLT method that enforces the visual representations to satisfy the *Independence* and *Sufficiency* properties. It shows good compatibility with existing TPR methods.
- We demonstrate the superiority of IRLT by extensive experiments and analysis on three benchmark datasets.

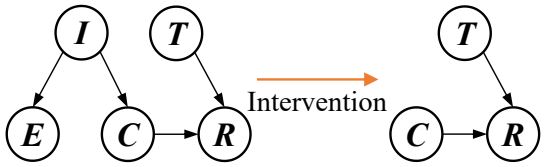

Figure 2: A causal graph of TPR.

TPR from the Causal View

Problem Formulation: Given a textual query and gallery person images, our task is to learn the cross-modal similarity for retrieval. Formally, the cross-modal dataset D consists of N image-text pairs, denoted as $D = \{x_i, t_i\}_{i=1}^{N}$. Each pair includes a pedestrian image x_i captured by a specific surveillance camera and its corresponding text description query t_i. In addition, the M pedestrians in the dataset all correspond to M specific identity labels, marked as $Y = \{y_i\}_{i=1}^{N}$ with $y_i \in \{1, \cdots, M\}$. The TPR backbone usually includes an image encoder and a text encoder, which can output image global representation $\mathbf{I}_g \in \mathbb{R}^d$ and text global representation $\mathbf{T}_g \in \mathbb{R}^d$, respectively. Since the pre-trained language models for text representation have achieved great progress in recent years, the main goal of this work is to effectively learn robust visual representation of person images.

During training, the basic learning objective of TPR models is to minimize the cross-modal matching loss based on the constructed positive/negative image-text pairs in the training set. During inference, we rank candidate images based on the estimated similarity scores between $\mathbf{I}_g$ and $\mathbf{T}_g$.

Causal Look at TPR: Here, we believe that disclosing "which part of the image is crucial for being retrieved by this text?" is the key to clearly presenting visual-linguistic alignment. To this end, we re-examine the TPR methods from the perspective of causal theory (Pearl 2009), then formalize it into a Structural Causal Model (SCM) (Pearl, Glymour, and Jewell 2016) by studying the causal relationships among five variables: input image I, text T, causal factor C, non-causal factor E, ground-truth retrieval R. Fig 2 illustrates the causal graph, where each link notes the causality between two nodes: cause→effect:

- $C \leftarrow I \rightarrow E$. Image I can be divided into two parts: 1) the causal factor C, which is consistent with the semantics of text T, and 2) the non-causal part E, *e.g.*, background, which is sensitive to the environmental change.
- $C \rightarrow R \leftarrow T$. The retrieval R is determined by T and C, reflecting visual-linguistic alignments. The example in Fig.1 demonstrates the critical visual cues (*i.e.*, C) for retrieval, which can lead to the true causal effect.

Looking closely at the causal graph, we find that non-causal factor E and ground truth retrieval R can be spuriously cor-

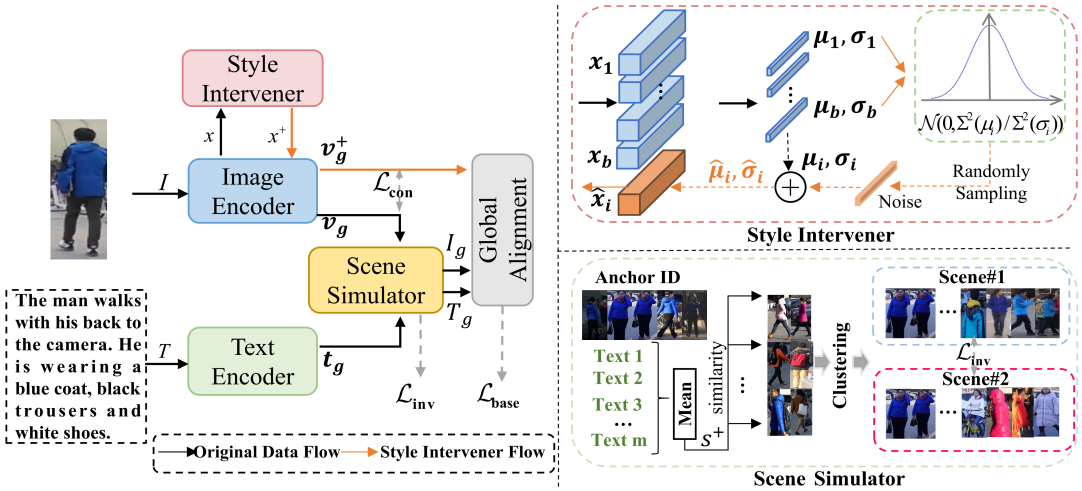

Figure 3: Overview of our proposed IRLT method. It mainly aims to learn causally-invariant person representation, which can satisfy two critical properties: *Independence* and *Sufficiency*. It can also be flexibly integrated into existing TPR framework.

related by $E \leftarrow I \rightarrow C \rightarrow R \leftarrow T$. Even though there is no direct causal path from E to R. This spurious association is known as a backdoor path (Goldberg 2019) and is usually ignored by traditional methods. Due to the existence of the backdoor path, statistical correlation-based models blindly capture the causal correlation between E and R. Therefore, determining causal factor C is the key to addressing these limitations and then improving the model's reliability.

Methodology

In order to determine causal factor C, inspired by the Independent Causal Mechanisms (ICM) Principle (Scholkopf et al. 2021) and invariant learning (Arjovsky et al. 2019), we argue that causal factors required for simulating the causal scene must adhere to the principles of **Independence** and **Sufficiency**. Therefore, we propose Invariant Representation Learning for TPR (**IRLT**). As shown in Fig.3, IRLT adds a style intervener and a scene simulator to TPR backbone for learning the independence and sufficiency of causal factors.

Style Intervener for Independence

As discussed before, causal factor C and non-causal factor E are often intertwined, and we first aim to separate C and E through causal intervention. Although the explicit form of C is usually unknown to the image encoder F_I, based on prior knowledge, we know that C should remain invariant to the intervention of E change, *i.e.*, $P(C|do(E)) = P(C|E)$. In this work, here we assume that the person image consists of two components: one is semantic ***content***, corresponding to causal factor C, and the rest part is ***style*** information, which is content-unrelated and corresponds to non-causal factor E, such as background. Therefore, the task of determining causal factor C is transformed as learning style-invariant semantic content representation of person image.

It can be captured by domain generalization and invariant learning techniques (Li et al. 2022b; Lv et al. 2022). For example, DSU (Li et al. 2022a) makes the model pay more attention to the content of the image by simulating the uncertainty of domain changes, to learn more invariant features. Inspired by this, we argue that non-causal scene E represents different domain-specific style information. Therefore, we can obtain more invariant information by interfering with the style information (*i.e.*, E). Specifically, let $\mathbf{x} \in \mathbb{R}^{B\times C\times H\times W}$ denotes the encoded features in the intermediate layers of the F_I, the channel-wise mean $\mu \in \mathbb{R}^{B\times C}$ and standard deviation $\sigma \in \mathbb{R}^{B\times C}$ of each feature map in a mini-batch as the original style information.

To simulate potential style changes, we assume that the distribution of each feature statistic follows a multi-variate Gaussian distribution. The mean and standard deviation of the feature statistic follow $\mathcal{N}(\mu, \Sigma_\mu^2)$ and $\mathcal{N}(\sigma, \Sigma_\sigma^2)$, respectively. We estimate the uncertainty of feature mean μ and standard deviation σ, using the variance of the feature statistics. This represents the potential range of style changes (non-causal scene changes) and is denoted by $\Sigma_\mu \in \mathbb{R}^C$ and $\Sigma_\sigma \in \mathbb{R}^C$. After obtaining uncertainty estimates for each feature channel, new feature statistics can be sampled randomly from the corresponding distribution:

$$\beta(\mathbf{x}) = \mu(\mathbf{x}) + \epsilon_\mu \Sigma_\mu(\mathbf{x}), \qquad \epsilon_\mu \sim \mathcal{N}(\mathbf{0}, \mathbf{1}), \tag{1}$$

$$\gamma(\mathbf{x}) = \sigma(\mathbf{x}) + \epsilon_\sigma \Sigma_\sigma(\mathbf{x}), \qquad \epsilon_\sigma \sim \mathcal{N}(\mathbf{0}, \mathbf{1}), \tag{2}$$

where $\beta(\mathbf{x})$ and $\gamma(\mathbf{x})$ represent the mean and standard deviation of the random style statistics, respectively. ϵ_μ and ϵ_σ both follow the standard Gaussian distribution. Referring to AdaIN (Huang and Belongie 2017), we replace the original style information with randomly generated style statistics:

$$\mathbf{x}^+ = \gamma(\mathbf{x}) \left(\frac{\mathbf{x} - \mu(\mathbf{x})}{\sigma(\mathbf{x})} \right) + \beta(\mathbf{x}). \tag{3}$$

Eq. (1) and (2) can be viewed as a type of do-operation, $do(E)$, which modifies the original style information. Eq. (3) is a type of **causal intervention** that augments the original image in the feature-level. We change the image encoder F_I to two branches: original data branch f_o and style intervener branch f_s. The difference between f_s and f_o is that f_s

adds the style intervener. The style intervener is placed after all the top-four ConvBlocks, respectively, in ResNet-50 and before the first transformer encoder in ViT. The global-level image features extracted by the two branches are respectively as: $\mathbf{v}_g = f_o(\mathbf{x}) \in \mathbb{R}^d$ and $\mathbf{v}_g^+ = f_s(\mathbf{x}^+) \in \mathbb{R}^d$. The text $T = \{t_i\}_{i=1}^q$ with word length q is fed into the text encoder F_T to obtain text features $\mathbf{t}_g \in \mathbb{R}^d$. We employ InfoNCE (Oord, Li, and Vinyals 2018) to make $\mathbf{v}_g$ independent by learning the consistency between $\mathbf{v}_g$ and $\mathbf{v}_g^+$:

$$\mathcal{L}_{\text{con}} = -\log \frac{\exp(\mathbf{v}_g^{\mathrm{T}} \mathbf{v}_g^+)}{\exp(\mathbf{v}_g^{\mathrm{T}} \mathbf{v}_g^+) + \sum_{i=1}^n \exp(\mathbf{v}_g^{\mathrm{T}} \mathbf{v}_g^-)}, \tag{4}$$

where the intervener of other images constitutes n negatives $\mathbf{v}_g^-$. Therefore, we can obtain **independent** (*i.e.*, $C \perp E$) visual representations $\mathbf{v}_g$ by minimizing the loss in Eq. (4).

Scene Simulator for Sufficiency

Even though the style intervener gives independence to the image feature $\mathbf{v}_g$, however, due to the complex alignment relationship between heterogeneous data, statistical models still tend to utilize simple alignment between image feature $\mathbf{v}_g$ and text feature $\mathbf{t}_g$, losing tiny but discriminative features. As shown in Fig.1, the text word "white" is associated with the "white" attribute in the images of other pedestrians, ignoring its own small but distinctive features (*e.g.*, hands in the pockets, blue jeans, and sneakers). This phenomenon is known as shortcut bias in causal inference (Pearl 2009). Recently, Invariant Risk Minimization (IRM) (Arjovsky et al. 2019; Krueger et al. 2021) cuts this spurious association by learning invariance across environments. Specifically, REx (Krueger et al. 2021) splits the training data into multiple environments $e \in \mathcal{E}$ and minimizes the variance of the risk across environments as regularization to find robust features. Inspired by this, we use environment-invariant learning to learn features that are **sufficient** (*i.e.*, $C \rightarrow R \leftarrow T$) for accurate and reliable retrieval across different scenes.

However, traditional IRM requires annotation of the environment, which is often not achievable in practice. To this end, we first propose a scene simulator that automatically builds the environment $\mathcal{E}$. As depicted in Fig.3, we regard each ID as an anchor environment mode, and divide the images of the remaining IDs into two groups: whether they are similar to the corresponding query of the anchor ID. Specifically, for each anchor ID that contains l images, **Scene#1** regards these l samples as positives, and "similar" samples from other IDs as negative; **Scene#2** contains the same positive samples, while "dissimilar" samples from other IDs are negative. Cosine function $S(\cdot,\cdot)$ is used to calculate the similarity $\mathbf{S} \in \mathbb{R}^{m \times n}$ between the m text representation $\mathbf{t}_{id} \in \mathbb{R}^{m \times d}$ of the anchor ID and n images representation $\mathbf{v}_o \in \mathbb{R}^{n \times d}$ sampled from other IDs. Then, we average the similarity matrix $\mathbf{S}$ along the axis of anchor ID to obtain $\mathbf{s}^+$. After ranking $\mathbf{s}^+$, we can easily construct two scenes: the "similar" samples grouped in Scene#1 (the high half value in $\mathbf{s}^+$) and the "dissimilar" samples grouped in Scene#2 (the lower half value in $\mathbf{s}^+$). We introduce an ID-wise IRM learning objective, $\mathcal{L}_{\text{inv}}$, to learn causally sufficient features. Specifically, we respectively compute the triplet-loss of $\mathbf{v}_g^e$ and $\mathbf{t}_g^e$ in each environment $e \in \mathcal{E}_k$, and make the training risk in different environments (*i.e.*, Scene#1 and Scene#2) as consistent as possible. Formally, the learning objective is :

$$\begin{aligned} \ell(e \in \mathcal{E}_k) = &\max(\alpha - S(\mathbf{v}_p, \mathbf{t}_p) + S(\mathbf{v}_p, \mathbf{t}_n), 0) \\ &+ \max(\alpha - S(\mathbf{v}_p, \mathbf{t}_p) + S(\mathbf{v}_n, \mathbf{t}_p), 0), \end{aligned} \tag{5}$$

$$\mathcal{L}_{\text{inv}} = \sum\nolimits_{k=1}^{M} (\lambda_1 \text{Var}(\{\ell(1), ..., \ell(e)\}) + \lambda_2 \sum\nolimits_{e \in \mathcal{E}_k} \ell(e)), \tag{6}$$

where $\mathbf{v}_p$ and $\mathbf{t}_p$ are drawn from a matching image-text pair. $\mathbf{t}_n$ and $\mathbf{v}_n$ denote the hardest negative text for $\mathbf{v}_p$ and the hardest negative image for $\mathbf{t}_p$ in e, respectively. Var stands for variance. So far, we have obtained text representation $\mathbf{T}_g \in \mathbb{R}^d$ and causally invariant person representations $\mathbf{I}_g \in \mathbb{R}^d$, so the cross-modal similarity is calculated as the cosine similarity $S_g = \mathbf{I}_g^{\mathrm{T}} \mathbf{T}_g / (\|\mathbf{I}_g\| \, \|\mathbf{T}_g\|)$:

Learning

To this end, we aim to integrate the $\mathcal{L}_{\text{con}}$ loss in Eq. (4) and the $\mathcal{L}_{\text{inv}}$ loss in Eq. (6) with the learning objective of existing methods to further improve the accuracy and reliability of TPR in a plug-and-play manner. The loss functions used by existing methods usually consist of $\mathcal{L}_{\text{id}}$ (Suo et al. 2022), $\mathcal{L}_{\text{cr}}$ (Ding et al. 2021), and $\mathcal{L}_{\text{sdm}}$ (Jiang and Ye 2023). $\mathcal{L}_{\text{id}}$ is the cross-entropy loss, which classifies images or texts into different groups based on their IDs. $\mathcal{L}_{\text{cr}}$ is the Compound Ranking (CR) loss which overcomes the problem of large intra-class variance in the text descriptions. $\mathcal{L}_{\text{sdm}}$ is the Similarity Distribution Matching (SDM) loss. It strengthens the correlation between matched pairs. We collectively refer to their various combinations in existing methods as $\mathcal{L}_{\text{base}}$. Thus, the overall loss $\mathcal{L}_{\text{Total}}$ is denoted as:

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{base}} + \underbrace{\mathcal{L}_{\text{inv}} + \lambda_3 \mathcal{L}_{\text{con}}}_{\text{Key components in IRLT}} \tag{7}$$

Summary: we have introduced two key components in IRLT. One is the Style Intervener in the Sec. 3.1 that aims to learn style-invariant visual representation by minimizing the $\mathcal{L}_{\text{con}}$ loss. The other one is the Scene Simulator in the Sec. 3.2 that aims to learn environment-invariant visual representation by minimizing the $\mathcal{L}_{\text{inv}}$ loss. The two parts can well complement each other, thus easily achieving our goal of learning causally invariant visual representation for TPR.

Experiments

This section conducts experiments to answer the following questions. **RQ1**: How effective is IRLT in improving the accuracy, cross-domain generalization, and robustness of existing SOTA TPR works?**RQ2**: How does the style intervener and scene simulator affect the performance? **RQ3**: What are the learning patterns and insights of IRLT?

Settings

Datasets. CUHK-PEDES(Li et al. 2017) is a pioneering dataset specifically designed for text-to-image person retrieval. It includes a total of 40,206 images and 80,412 text descriptions, covering 13,003 distinct identities. **ICFG-PEDES** (Ding et al. 2021) consists of a total of 54,522 images representing 4,102 distinct identities. Each image is

Methods	Ref	CUHK-PEDES			ICFG-PEDES			RSTPReid		
		Rank-1	Rank-5	Rank-10	Rank-1	Rank-5	Rank-10	Rank-1	Rank-5	Rank-10
NAFS	arXiv21	59.36	79.13	86.00	-	-	-	-	-	-
DSSL	MM21	59.98	80.41	87.56	-	-	-	39.05	62.60	73.95
LBUL	MM22	64.04	82.66	87.22	-	-	-	45.55	68.20	77.85
TIPCB	Neuro22	64.26	83.19	89.10	54.96	74.72	81.89	-	-	-
CAIBC	MM22	64.43	82.87	88.37	-	-	-	47.35	69.55	79.00
LGUR	MM22	65.25	83.12	89.00	59.02	75.32	81.56	-	-	-
IVT	ECCVW22	65.59	83.11	89.21	56.04	73.60	80.22	46.70	70.00	78.80
CFine	arXiv22	69.57	85.93	91.15	60.83	76.55	82.42	50.55	72.50	81.60
Baseline	-	58.47	78.69	85.32	52.27	71.83	79.13	40.54	66.48	78.31
+ IRLT (Ours)	-	60.53	80.11	87.23	54.64	73.76	80.47	44.56	71.18	81.98
Baseline-CLIP	-	70.49	87.63	92.08	61.04	78.21	83.99	58.68	80.72	87.13
+ IRLT (Ours)	-	73.13	89.07	93.31	62.76	80.01	85.37	59.81	81.79	88.56
SSAN	arXiv21	61.37	80.15	86.73	54.23	72.63	79.53	43.50	67.80	77.15
+ IRLT (Ours)	-	62.64	81.13	87.02	55.31	73.15	80.59	46.14	68.93	78.79
SRCF	ECCV22	64.04	82.99	88.81	57.18	75.01	81.49	45.05	70.50	80.25
+ IRLT (Ours)	-	66.05	83.52	89.54	58.32	75.82	82.20	47.14	72.21	81.01
IRRA	CVPR23	73.38	89.83	93.71	63.46	80.25	85.82	60.20	81.30	88.20
+ IRLT (Ours)	-	74.46	90.19	94.01	64.72	81.35	86.31	61.49	82.26	89.23

Table 1: Performance comparisons with SOTA methods on the CUHK-PEDES, ICFG-PEDES and RSTPReid datasets.

associated with a single text description. **RSTPReid** (Zhu et al. 2021) comprises 20,505 images depicting 4,101 unique identities captured by 15 cameras. Each identity is associated with five images captured by distinct cameras, and each image is annotated with two text descriptions. For all three datasets, we follow their official data splits for experiments and utilize the Rank-k metrics (with k values of 1, 5, and 10) as the principal evaluation metrics.

Baselines. We verify the effectiveness of IRLT on two types of TPR backbone networks: 1) **non-CLIP-driven:** NAFS (Gao et al. 2021), DSSL (Zhu et al. 2021), SCAN (Lee et al. 2018), SSAN (Ding et al. 2021), LBUL (Wang et al. 2022b), TIPCB (Chen et al. 2022), CAIBC (Wang et al. 2022a), LGUR (Shao et al. 2022), SRCF (Suo et al. 2022); 2) **CLIP-driven:** IVT (Shu et al. 2022), CFine (Yan et al. 2022), IRRA (Jiang and Ye 2023). To fully investigate the effectiveness of IRLT in boosting the TPR performance, we first build two simple TPR baseline methods, termed **Baseline** and **Baseline-CLIP**, and also use three existing SOTA TPR methods as our baselines: **SSAN**, **SRCF** and **IRRA**. Our IRLT is compatible with the five baselines. The experimental results are shown in Table 1.

Implementation Details. The same settings of Baseline and Baseline-CLIP: 1) The input image is resized to 384 × 128. 2) Adam (Kingma and Ba 2014) is used as the optimizer to train for 60 epochs and set the batchsize to 64. 3) The hyper-parameters λ_2 and λ_3 are fixed as 1 and 0.1, respectively. The different settings of Baseline and Baseline-CLIP: 1) Baseline follows SRCF (Suo et al. 2022), using ResNet-50 as the image encoder and BETR as the text encoder. Baseline-CLIP follows IRRA (Jiang and Ye 2023), using CLIP-ViT as the image encoder and CLIP-Xformer as the text encoder. 2) The text length is set to 64 and 77, respectively. 3) The dimension d of the representation is 1024 and 512, respectively. 4) The initial learning rate is $5e$-4 and $1e$-5, respectively. 5) The learning rate decay strategies are fixed-step decay (0.1 times decay every 10 epochs) and cosine learning rate decay, respectively. 6) The hyper-parameter λ_1 is 0.5 and 0.1, respectively. The two baselines are simple and efficient which use independent image/text encoder to extract the image/text representation and perform cosine similarity computation in a common space for fast retrieval. Furthermore, when IRLT is combined with SSAN, SRCF, and IRRA, the original experimental settings of these methods remain unaltered. We conduct our experiments using a single RTX 3090 GPU with 24GB of memory.

Main Result (RQ1)

Comparisons with SOTA Methods. Table 1 presents the comparison between our approach and the SOTAs on three benchmark datasets.We have the following observations:**1)** On all three benchmark datasets, the proposed IRLT clearly outperforms the Baseline and Baseline-CLIP with a significant advantage (+1.13%∼4.02%) on Rank-1. The retrieval results are more effective than some methods (*e.g.,* DSSL, CFine) using complex cross-modal alignment. Moreover, IRLT also achieves a stable improvement (+1.08%∼1.29%) on Rank-1 over SOTA (IRRA). This consistent performance demonstrates the overall effectiveness of applying invariant representation learning to TPR and supports the theoretical soundness of the invariance principle. **2)** Narrowing down the analysis to each of the three existing SOTA TPR methods, IRLT demonstrates model-agnostic property by delivering significant gains (+1.08%∼2.64%) on Rank-1 for each backbone model across all benchmark datasets. We observe that the improvements on SRCF and SSAN are more significant compared to those on IRRA. This is mainly because IRRA is finetuned based on CLIP, allowing the transfer of semantic alignment knowledge obtained from large-scale cross-modal datasets to downstream tasks. This, in turn, partially explains the contributions of IRLT. **3)** Comparing the average improvements on Rank-1 across differ-

	Methods	Rank-1	Rank-5	Rank-10
C→I	SCAN	21.27	39.26	48.83
	SSAN	29.24	49.00	58.53
	LGUR	34.25	52.58	60.85
	SRCF	33.47	52.63	61.05
	SRCF+IRLT	35.26	54.08	62.73
	IRRA	42.21	61.92	70.13
	IRRA+IRLT	**43.46**	**62.53**	**71.26**
I→C	SCAN	13.63	28.61	37.05
	SSAN	21.07	38.94	48.54
	LGUR	25.44	44.48	54.39
	SRCF	25.13	43.96	54.68
	SRCF+IRLT	26.79	45.62	54.89
	IRRA	36.89	58.02	67.46
	IRRA+IRLT	**38.03**	**59.14**	**68.51**

Table 2: Cross-dataset generalization of IRLT. "C" denotes CUHK-PEDES, while "I" represents ICFG-PEDES.

ent benchmarks, we observe that IRLT achieves the best improvement on RSTPReid (+1.13%∼4.02%), while obtaining relatively moderate improvements on CUHK-PEDES (+1.08%∼2.64%) and ICFG-PEDES (1.08%∼2.37%). The reason for this discrepancy is that RSTPReid has a relatively smaller size, which limits the inference capability of the backbone model. However, this limitation aligns with the focus of IRLT, enabling it to perform better in a less generalized situation, thus leading to more favorable growth.

Cross-dataset generalization of IRLT. Our LRLT effectively reduces spurious associations between text and visual features, and it is natural to assume that the model generalizes well to other domains. Therefore, we test the effect of IRLT on the cross-dataset generalization task. Specifically, we employ a model trained on the source domain to assess its performance on the target domain. We utilize CUHK-PEDES and ICFG-PEDES as the source domain and the target domain in turn. As shown in Table 2, comparisons with SRCF and IRRA, which are the best performers in the cross-dataset generalization task, IRLT ("+IRLT") still achieves a notable improvement (+1.14%∼1.79%) on Rank-1. This is mainly because IRLT can uncover the intrinsic cross-modality causal correlation for better generalization.

Robustness of IRLT. We verify the robustness and reliability of IRLT by applying different types of perturbations to the images. As shown in Table 3, when IRLT is combined with SRCF and IRRA, IRLT ("+IRLT") can progressively improve the anti-jamming ability (*i.e.,* the percentage drop under all variations) of both models. This suggests that our emphasis on the independence and sufficiency that visual representations should have can effectively improve the robustness and reliability of the models.

In-Depth Study (RQ2)

What is the impact of the IRLT component? To fully understand the inference mechanism of LRLT, we carefully analyze its structure. Specifically, we explore the effectiveness of the proposed invariant learning by analyzing the performance of the style intervener and scene simulator with different backbones on three benchmarks. We report the performance in Table 4 and summarize our findings as follows: 1) **The Effect of the Style Intervener.** We first demonstrate the efficacy of the style intervener (*i.e.*, $\mathcal{L}_{con}$) by comparing its permanence ("+Intervener") to SRCF and IRRA. The style intervener can force the model to learn the independence between causal and non-causal factors. This can obtain style-invariant visual representation, which provides significant gains (+0.56%∼1.41%) on Rank-1 for each backbone model across all benchmark datasets. 2) **The Effect of the Scene Simulator.** We validate the substantial efficacy of the scene simulator (*i.e.*, $\mathcal{L}_{inv}$) by investigating its performance ("+Simulator"). IRLT consistently improves Rank-1 (+0.51%∼1.66%) on all benchmark datasets for each model, suggesting that emphasizing the sufficiency of causal factors can enhance visual-linguistic alignment.

Methods	**R**	**H**	**V**	**R**	**C**
NAFS	59.94	55.52	54.84	49.90	54.89
SSAN	61.37	57.91	56.24	50.80	58.11
SRCF	64.04	61.89	60.33	59.23	61.50
SRCF+IRLT	**66.05**	**64.30**	**63.04**	**62.10**	**63.84**
IRRA	73.38	71.05	69.26	68.11	70.77
IRRA+IRLT	**74.46**	**72.63**	**70.79**	**69.81**	**72.32**

Table 3: Robustness of IRLT. The 2nd column is examples of original images from CUHK-PEDES. Columns 3-6, represent the four disruption settings we implemented, including random horizontal translation, random vertical translation, random rotating, and random cropping.

Ablation	CUHK-PEDES	ICFG-PEDES	RSTPReid
SRCF	64.04	57.18	45.05
+ Intervener	65.23	58.15	46.46
+ Simulator	65.46	57.91	46.71
+IRLT	**66.05**	**58.32**	**47.14**
IRRA	73.38	63.46	60.20
+ Intervener	73.94	64.26	61.13
+ Simulator	73.89	63.81	60.74
+IRLT	**74.46**	**64.72**	**61.49**

Table 4: Evaluation on the effectiveness of sub-modules.

What are the effects of hyper-parameters? λ_1 and λ_2 in Eq. (6) denote the strength of environment-invariant, while λ_3 in Eq. (7) controls the strength of the disentanglement for the causal/non-causal features. To explore their impacts, we conduct experiments on CUHK-PEDES by combining IRLT with SRCF and IRRA. We fix two coefficients as 1 and change the other one in$\{0.05, 0.1, 0.5, 1, 2\}$. The peaks of λ_1 are 0.5 and 0.1 in Fig.4 (a) and Fig.4 (b), respectively. This is consistent with our finding that baselines with weak feature extraction require greater strength of environmental-invariant supervision. λ_2 and λ_3 achieve the best results in the region of 1 and 0.1, respectively. Overall, the change in hyper-parameters does not bring significant performance degradation, which validates the stability of our method.

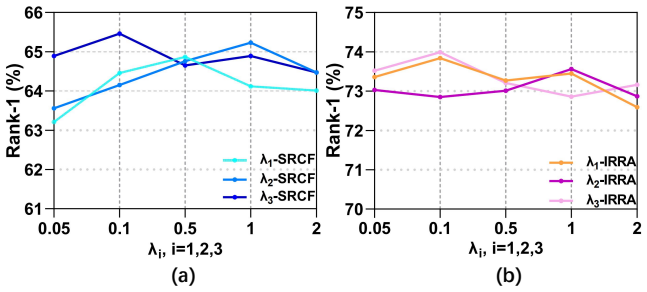

Figure 4: Study of the three hyper-parameters, λ_1, λ_2 and λ_3 using two baselines: SRCF (a) and IRRA (b).

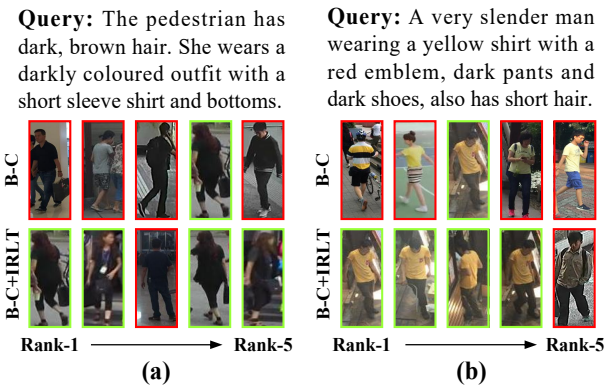

Figure 5: Top-5 retrieval results on two examples in CUHK-PEDES. The "B-C" row and the "B-C+IRLT" row show the results of Baseline-CLIP and IRLT ("+IRLT"), respectively. Matched/Mismatched images are marked in green/red.

Quantitative Study (RQ3)

Looking closely at Fig. 5 (a), we find that IRLT can improve the retrieval accuracy of Baseline-CLIP when facing images of one pedestrian with various environments and postures. This is thanks to the Style Intervener focuses on learning the characteristics of the pedestrians themselves. In addition, as displayed in Fig. 5 (b), when "yellow shirt" appears in the query, Baseline-CLIP indiscriminately retrieves other pedestrian images with the attribute of "yellow", whereas IRLT captures sufficient causal factors to perform correct similar pedestrian retrieval. In conclusion, by exploiting the independence and sufficiency of the visual representation, accurate and reliable retrieval can be accomplished.

Related Work

Text-based person retrieval (TPR) is a cross-modal retrieval task (Li et al. 2017; Chang et al. 2023). It is different from classic person re-identification task (Yang et al. 2017; Yang, Wang, and Tao 2018; Yang, Zhou, and Wang 2018). The primary challenge lies in fine-grained alignment of visual scenes and text descriptions. Existing approaches can be broadly classified into two categories: 1) focusing on cross-modal alignment strategies (Suo et al. 2022; Chen et al. 2022) and 2) focusing on powerful representation learning (Jiang and Ye 2023; Yan et al. 2022). For cross-modal alignment, methods usually design unique modules or strategies. For example, (Suo et al. 2022) designs denoising filters and dictionary filters to extract critical features. (Shu et al. 2022) proposes multi-level alignment (MLA) and bidirectional mask modeling (BMM) to obtain fine-grained alignment. For representation learning, some approaches enhanced the model's feature representation capabilities by incorporating multiple auxiliary tasks. These tasks included attribute alignment (Wang et al. 2020), foreground segmentation (Zhu et al. 2021), image-to-text generation (Zeng et al. 2021), *etc.* Recently, Contrastive Language Image Pre-training (CLIP) (Radford et al. 2021) as the most representative work of visual-language pre-training (VLP), has achieved significant success, due to its rich multi-modal representation. Therefore, many works (Yan et al. 2022; Jiang and Ye 2023) have introduced CLIP into TPR to enhance cross-modal understanding and matching. Different from the above methods, we do not design complex cross-modal alignment strategies or introduce a new powerful backbone network. Instead, we eliminate the spurious associations between cross-modal representations from the causal view.

Causal Inference. Multi-modal datasets inevitably introduce undesired correlations between inputs and ground truth annotations during the collection process. Consequently, causal inference (Pearl, Glymour, and Jewell 2016) has attracted increasing attention as a tool for removing bias from domain-specific datasets in information retrieval and multimedia, such as video moment retrieval (Yang et al. 2021), video question answering (Li et al. 2022b), video relation detection (Li et al. 2021), and visual dialog (Qi et al. 2020), *etc.* To overcome inherent bias in datasets, IRM (Arjovsky et al. 2019) advances causal inference by proposing environment-invariant priors and reducing cross-environment variance, thus revealing the underlying causal patterns. However, conventional IRM usually lacks environment annotations. Different from the above methods, we learn visual-linguistic alignment by intervening in the distribution of non-causal variables and relaxing the constraints on the required environmental annotations. To the best of our knowledge, IRLT is the first work that introduces causal inference as a model-agnostic framework for TPR.

Conclusion

In this paper, we pointed out that existing TPR methods are easily susceptible to the unstable statistical correlations between textual query and the non-causal factors in person images, thus leading to unreliable retrieval results. We propose a causality-inspired and model-agnostic IRLT method to learn causally invariant representation for TPR. Extensive experimental results clearly validated the effectiveness and reliability of IRLT across diverse experimental settings. In the future, we will try to explicitly disentangle the causal factors from the natural person images for more rigorous causal inference and also extend our effort to other cross-modal visual reasoning tasks (Yang et al. 2020, 2022; Zhou et al. 2023a,b; Dong et al. 2021; Yang et al. 2023).

Acknowledgments

This research is supported by the National Natural Science Foundation of China (62276112, 62272435, and U22A2094) and Jilin Province Science and Technology Development Plan Key RD Project (20230201088GX) 871245).

References

Arjovsky, M.; Bottou, L.; Gulrajani, I.; and Lopez-Paz, D. 2019. Invariant Risk Minimization. *CoRR*, abs/1907.02893.

Chang, T.; Yang, X.; Luo, X.; Ji, W.; and Wang, M. 2023. Learning Style-Invariant Robust Representation for Generalizable Visual Instance Retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*, 6171–6180.

Chen, Y.; Zhang, G.; Lu, Y.; Wang, Z.; and Zheng, Y. 2022. TIPCB: A simple but effective part-based convolutional baseline for text-based person search. *Neurocomputing*, 494: 171–181.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ding, Z.; Ding, C.; Shao, Z.; and Tao, D. 2021. Semantically self-aligned network for text-to-image part-aware person re-identification. *arXiv preprint arXiv:2107.12666*.

Dong, J.; Li, X.; Xu, C.; Yang, X.; Yang, G.; Wang, X.; and Wang, M. 2021. Dual encoding for video retrieval by text. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8): 4065–4080.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Gao, C.; Cai, G.; Jiang, X.; Zheng, F.; Zhang, J.; Gong, Y.; Peng, P.; Guo, X.; and Sun, X. 2021. Contextual Non-Local Alignment over Full-Scale Representation for Text-Based Person Search. *arXiv preprint arXiv:2101.03036*.

Goldberg, L. R. 2019. The Book of Why: The New Science of Cause and Effect. *Notices of the American Mathematical Society*, 1.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Huang, X.; and Belongie, S. J. 2017. Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, 1510–1519. IEEE Computer Society.

Jiang, D.; and Ye, M. 2023. Cross-Modal Implicit Relation Reasoning and Aligning for Text-to-Image Person Retrieval. *CoRR*, abs/2303.12501.

Kingma, D.; and Ba, J. 2014. Adam: A Method for Stochastic Optimization. *arXiv: Learning,arXiv: Learning*.

Krueger, D.; Caballero, E.; Jacobsen, J.; Zhang, A.; Binas, J.; Zhang, D.; Priol, R. L.; and Courville, A. C. 2021. Out-of-Distribution Generalization via Risk Extrapolation (REx). In *ICML*, volume 139, 5815–5826.

Lee, K.; Chen, X.; Hua, G.; Hu, H.; and He, X. 2018. Stacked Cross Attention for Image-Text Matching. In Ferrari, V.; Hebert, M.; Sminchisescu, C.; and Weiss, Y., eds., *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part IV*, volume 11208 of *Lecture Notes in Computer Science*, 212–228. Springer.

Li, S.; Xiao, T.; Li, H.; Zhou, B.; Yue, D.; and Wang, X. 2017. Person search with natural language description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1970–1979.

Li, X.; Dai, Y.; Ge, Y.; Liu, J.; Shan, Y.; and Duan, L. 2022a. Uncertainty Modeling for Out-of-Distribution Generalization. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Li, Y.; Wang, X.; Xiao, J.; Ji, W.; and Chua, T. 2022b. Invariant Grounding for Video Question Answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 2918–2927. IEEE.

Li, Y.; Yang, X.; Shang, X.; and Chua, T.-S. 2021. Interventional video relation detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, 4091–4099.

Lv, F.; Liang, J.; Li, S.; Zang, B.; Liu, C. H.; Wang, Z.; and Liu, D. 2022. Causality Inspired Representation Learning for Domain Generalization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 8036–8046. IEEE.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Pearl, J. 2009. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2nd edition.

Pearl, J.; Glymour, M.; and Jewell, N. P. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.

Qi, J.; Niu, Y.; Huang, J.; and Zhang, H. 2020. Two Causal Principles for Improving Visual Dialog. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, 8748–8763. PMLR.

Scholkopf, B.; Locatello, F.; Bauer, S.; Ke, N. R.; Kalchbrenner, N.; Goyal, A.; and Bengio, Y. 2021. Toward Causal Representation Learning. *Proceedings of the IEEE*, 612–634.

Shao, Z.; Zhang, X.; Fang, M.; Lin, Z.; Wang, J.; and Ding, C. 2022. Learning Granularity-Unified Representations for Text-to-Image Person Re-identification. In Magalhães, J.; Bimbo, A. D.; Satoh, S.; Sebe, N.; Alameda-Pineda, X.; Jin, Q.; Oria, V.; and Toni, L., eds., *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, 5566–5574. ACM.

Shu, X.; Wen, W.; Wu, H.; Chen, K.; Song, Y.; Qiao, R.; Ren, B.; and Wang, X. 2022. See Finer, See More: Implicit Modality Alignment for Text-Based Person Retrieval. In Karlinsky, L.; Michaeli, T.; and Nishino, K., eds., *Computer Vision - ECCV 2022 Workshops - Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part V*, volume 13805 of *Lecture Notes in Computer Science*, 624–641. Springer.

Suo, W.; Sun, M.; Niu, K.; Gao, Y.; Wang, P.; Zhang, Y.; and Wu, Q. 2022. A Simple and Robust Correlation Filtering Method for Text-Based Person Search. In *European Conference on Computer Vision*, 726–742. Springer.

Wang, Z.; Fang, Z.; Wang, J.; and Yang, Y. 2020. ViTAA: Visual-Textual Attributes Alignment in Person Search by Natural Language. *arXiv preprint arXiv:2005.07327*.

Wang, Z.; Zhu, A.; Xue, J.; Wan, X.; Liu, C.; Wang, T.; and Li, Y. 2022a. CAIBC: Capturing All-round Information Beyond Color for Text-based Person Retrieval. In Magalhães, J.; Bimbo, A. D.; Satoh, S.; Sebe, N.; Alameda-Pineda, X.; Jin, Q.; Oria, V.; and Toni, L., eds., *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, 5314–5322. ACM.

Wang, Z.; Zhu, A.; Xue, J.; Wan, X.; Liu, C.; Wang, T.; and Li, Y. 2022b. Look Before You Leap: Improving Text-based Person Retrieval by Learning A Consistent Cross-modal Common Manifold. In Magalhães, J.; Bimbo, A. D.; Satoh, S.; Sebe, N.; Alameda-Pineda, X.; Jin, Q.; Oria, V.; and Toni, L., eds., *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, 1984–1992. ACM.

Yan, S.; Dong, N.; Zhang, L.; and Tang, J. 2022. CLIP-Driven Fine-grained Text-Image Person Re-identification. *arXiv preprint arXiv:2210.10276*.

Yang, X.; Dong, J.; Cao, Y.; Wang, X.; Wang, M.; and Chua, T.-S. 2020. Tree-augmented cross-modal encoding for complex-query video retrieval. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, 1339–1348.

Yang, X.; Feng, F.; Ji, W.; Wang, M.; and Chua, T.-S. 2021. Deconfounded Video Moment Retrieval with Causal Intervention. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Yang, X.; Wang, M.; Hong, R.; Tian, Q.; and Rui, Y. 2017. Enhancing person re-identification in a self-trained subspace. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 13(3): 1–23.

Yang, X.; Wang, M.; and Tao, D. 2018. Person re-identification with metric learning using privileged information. *IEEE Transactions on Image Processing*, 27: 791–805.

Yang, X.; Wang, S.; Dong, J.; Dong, J.; Wang, M.; and Chua, T.-S. 2022. Video moment retrieval with cross-modal neural architecture search. *IEEE Transactions on Image Processing*, 31: 1204–1216.

Yang, X.; Zhou, P.; and Wang, M. 2018. Person reidentification via structural deep metric learning. *IEEE transactions on neural networks and learning systems*, 30(10): 2987–2998.

Yang, Y.; Chen, H.; Liu, Z.; Lyu, Y.; Zhang, B.; Wu, S.; Wang, Z.; and Ren, K. 2023. Action Recognition with Multi-stream Motion Modeling and Mutual Information Maximization. *arXiv preprint arXiv:2306.07576*.

Zeng, P.; Jing, S.; Song, J.; Fan, K.; Li, X.; We, L.; and Guo, Y. 2021. Relation-aware aggregation network with auxiliary guidance for text-based person search. *World Wide Web*.

Zhou, S.; Guo, D.; Li, J.; Yang, X.; and Wang, M. 2023a. Exploring Sparse Spatial Relation in Graph Inference for Text-Based VQA. *IEEE Transactions on Image Processing*.

Zhou, S.; Guo, D.; Yang, X.; Dong, J.; and Wang, M. 2023b. Graph Pooling Inference Network for Text-Based VQA. *ACM Transactions on Multimedia Computing, Communications and Applications*.

Zhu, A.; Wang, Z.; Li, Y.; Wan, X.; Jin, J.; Wang, T.; Hu, F.; and Hua, G. 2021. DSSL: Deep Surroundings-person Separation Learning for Text-based Person Retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia*, 209–217.