VLUReID: Exploiting Vision-Language Knowledge for Unsupervised Person Re-Identification

Dongmei Zhang[1], Ray Zhang[2], Fan Yang[1], Yuan Li[1], Huizhu Jia[2†], Xiaodong Xie[1], Shanghang Zhang[1]
[1] *National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University*, Beijing, China
[2] *Shanghai AI Lab*, Shanghai, China
dmzhang@stu.pku.edu.cn

2024 IEEE International Conference on Multimedia and Expo (ICME) | 979-8-3503-9015-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICME57554.2024.10688349

***Abstract*—The superior performances of pre-trained vision-language models on various downstream tasks demonstrate the effectiveness of integrating cross-modal vision-language knowledge into visual tasks. However, this knowledge is hardly used for visual-based person re-identification (re-ID) because the datasets lack textual descriptions. Existing efforts require manual annotations for training, which can be time-consuming. We propose VLUReID, a framework that improves visual-based person re-ID using vision-language knowledge without requiring manual annotations from datasets. Specifically, the Vision-to-Text Association (VTA) module uses designed textual prompts to prompt the vision-language model in generating pseudo-semantic labels for visual inputs. Subsequently, within the Dual-Branch Asymmetric Training (DBAT) module, we propose an asymmetric training strategy to extract cross-modal knowledge from pseudo-semantic labels and integrate it into the person re-ID model. The experimental results on two widely-used benchmarks for unsupervised video-based person re-ID demonstrate the effectiveness of our framework.**

***Index Terms*—prompt, unsupervised person re-ID, cross-modal vision-language knowledge**

I. Introduction

Person re-identification (re-ID) task aims to identify and match individuals across various camera views, achieved by training models to learn discriminative features that remain consistent against appearance variations stemming from factors like pose, lighting, camera views, and occlusions. This leads to different methods of delivering information about identifying and matching individuals. Examples include datasets that are manually labeled using person IDs used in supervised methods [1], or camera system layout and filming time [2], among other examples.

Although a variety of prior knowledge is available, cross-modal vision-language knowledge is not frequently used for visual-based person re-ID which offers clearer assistance for identification by drawing inspiration from human approaches to person re-ID. A person's appearance description, such as *"she is wearing a red top and blue pants."*, is more reliable than a direct match to visual characteristics that might change in appearance. Furthermore, vision-language knowledge's higher results on a range of downstream tasks highlight the importance of such knowledge in visual tasks.

†Corresponding author.

This project is supported by the Joint R&D Fund of Beijing Smartchip Microelectronics Technology Co., Ltd.

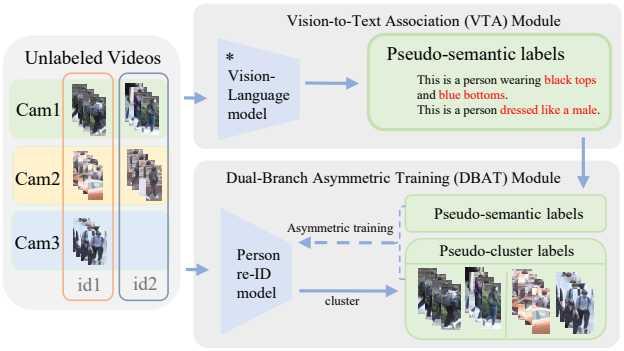


Fig. 1. **The overview of VLUReID framework**. Identity labels 'id*' for videos are not supplied. The pseudo-semantic labels for each video are generated in the VTA module. These labels are used in the DBAT module to supervise the person re-ID model, along with the pseudo-cluster labels. '*' represents that the vision-language model is frozen.

However, explicit textual labels are frequently used in visual tasks where vision-language knowledge has already been applied successfully. Text-based person re-ID [3] is an example. Textual descriptions are always absent from datasets used for visual-based person re-ID tasks. Furthermore, the person re-ID tasks require fine-grained matching, which makes it difficult to manually give appropriate textual descriptions. In person re-ID, Li *et al.* [1] makes the initial effort to apply CLIP. However, it necessitates manual annotations and a two-stage training technique with high computational and temporal overhead to update parameters.

In light of the above analysis, We propose a framework, **VLUReID**, to integrate cross-modal vision-language knowledge into the person re-ID model, without the need for manual annotations. We use textual prompts designed for person re-ID and asymmetric training techniques to extract cross-modal knowledge from the vision-language model. Our framework has two modules. In the Vision-to-Text Association (VTA) module, the pseudo-semantic labels are generated by prompting the vision-language model using textual prompts. In the Dual-Branch Asymmetric Training (DBAT) module, we use both a dual-branch asymmetric architecture and an asymmetric training strategy. Cross-modal vision-language knowledge improves the person re-ID model. During testing, no extra

computational load is applied—only the person re-ID model is employed. In conclusion, our contributions are as follows:

- To the best of our knowledge, We are the first to incorporate cross-modal vision-language knowledge into visual-based person re-ID without the need for manual annotations.
- We propose the VLUReID framework. To improve the person re-ID model, the cross-modal vision-language knowledge is extracted through the use of the textual prompts designed for person re-ID tasks and the asymmetric training strategy.
- Experimental results on two popular benchmarks for video-based person re-ID, MARS, and DukeMTMC-VideoReID, show the efficacy of our framework.

II. Related Works

We introduce methods for vision-language pre-training models and visual-based person re-identification (re-ID).

Vision-Language Pre-training Models. The study and use of vision-language pre-training models demonstrate the benefit of cross-modal knowledge. The rich training data and well-designed training tasks enable Contrastive Language image pre-training (CLIP) [4] to extract high-level semantic information from images. The CLIP model has been applied to many downstream tasks. Yan *et al.* [3] firstly introduces CLIP to text-based person re-ID.

Visual-based Person Re-ID Methods. Visual-based person re-ID models learn discriminative features that maintain consistency against appearance variations caused by factors like pose, lighting, camera viewpoints, and occlusions. Thus a range of knowledge is offered in different ways that allows the model to learn the intrinsic traits of individuals, either directly or implicitly. For example, datasets that are manually labeled using person IDs used in supervised methods [1], or camera system layout and filming time [2], etc.

Useful prior information can also come from vision-language knowledge. Li *et al.* [1] and Lin *et al.* [5] apply cross-modal knowledge from vision-language models to person re-ID tasks and yield promising results. However, their method uses human-annotated datasets in a two-stage training procedure, increasing the time and resource costs involved in gathering data and training the model. Our framework enables the application of vision-language knowledge to person re-ID without requiring human annotations or changing the vision-language model's parameters.

III. Methodology

Our VLUReID framework, as illustrated in Fig. 1, includes two key modules, the Vision-to-Text Association (VTA) module and the Dual-Branch Asymmetric Training (DBAT) module. Their details are depicted in Fig. 2.

A. Vision-to-Text Association Module

We design textual prompts for person re-identification (re-ID) to fully use the cross-modal knowledge included in the vision-language model. The vision-language model—the CLIP model in our framework—can then be prompted to generate the pseudo-semantic labels for each visual input.

Visual inputs. CLIP is trained to process image $I \in \mathbb{R}^{H\times W\times C}$. In video data $V = \{I_1, I_2, \ldots, I_T\}$, where $T = |V|$, every frame is input to CLIP's vision encoder. From there, their features $f_V \in \mathbb{R}^{T\times 512}$ are extracted. Preprocessing each frame using the CLIP model enables a more thorough use of video data than simply training the person re-ID model with sampled frames.

Textual inputs. Since visual-based person re-ID datasets do not contain textual descriptions, textual prompts must be carefully designed. There are some factors to be considered when doing the person re-ID process. Many elements taken as a whole affect how well videos match. However, on the one hand, it's difficult to provide a thorough description for every video manually. On the other hand, vision-language models are usually trained at the image level, they face difficulties in distinguishing finer concepts like spectacles, jewelry, and so on. Thus, it becomes crucial to carefully choose and refine the important components.

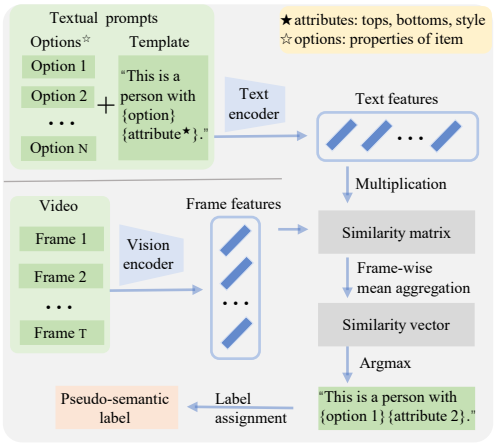
We select three essential attributes for the textual prompts, which include things like top and bottom color as well as image style. Since the tops and bottoms of a person cover the most surface area, CLIP may successfully associate visual elements and textual concepts. Similarly, we predict that CLIP performs well in describing general visual inputs as it is trained via contrastive learning at the image level. In addition to clothing color, image style is a useful addition that addresses other aspects of appearance such as body type, hair length, clothing type, and more. To make things easier, we divide the image style into two main categories: male and female. To verify CLIP's ability, we predict the image style of videos in MARS [6] and DukeMTMC-VideoReID [7] datasets. The CLIP model has a fair level of accuracy according to the labels in [8] , attaining 92% and 77% respectively.

We design the template and options for each attribute to create the textual prompts. The template is *"This is an image of a person with* { *option* } {*attribute*}*."*. The choices for the *tops* and *bottoms* attributes come from chromatography, which has eight different colors. Additionally, the *style* attribute has two options. It is flexible to modify the options and attributes following the features of the visual inputs. Given the CLIP's zero-shot classification ability, such adjustments don't add any more work. A set of textual prompts is formed by associating each attribute with its options. The resulting textual features are represented as $f_P \in \mathbb{R}^{N\times 512}$ for each attribute, where N is the number of attribute's options.

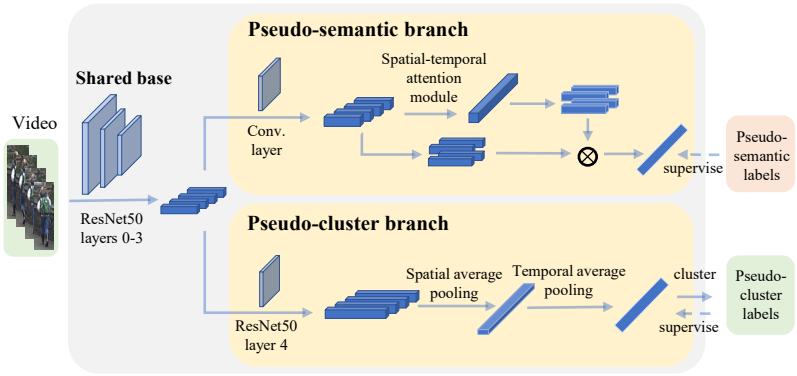
Generation of pseudo-semantic labels. The pseudo-semantic labels can be assigned by computing the similarity between the visual feature f_V and the textual features f_P.

$$s = \arg\max_{n} \sum_{t=1}^{T} \mathrm{S}(f_{V_t}, f_{P_n}) \tag{1}$$

where the index of chosen option is denoted by $s \in \mathbb{R}$. $\mathrm{S}(\cdot)$ represents the cosine similarity measure. The t_{th} visual feature



(a) Vision-to-Text Association (VTA)



(b) Dual-Branch Asymmetric Training (DBAT)

Fig. 2. **The two essential VLUReID framework modules.** (a) shows the VTA module, where the vision and text encoders are used to generate pseudo-semantic labels. Textual prompts are designed for person re-ID tasks. (b) shows the model architecture in the DBAT module, including a pseudo-semantic branch and a pseudo-cluster branch. During testing, the shared base and the pseudo-cluster branch form the person re-ID model.

and the n_{th} textual characteristic of a movie are denoted by the variables f_{V_t} and f_{P_n}.

B. Dual-Branch Asymmetric Training Module

We design a dual-branch network and an asymmetric training strategy in the DBAT module, taking into account that the pseudo-semantic labels and the pseudo-cluster labels describe person videos in different ways and at different levels, where the former is for predicting the specific attributes of a person and the latter is for differentiating between different people.

Dual-branch network. Our dual-branch network is composed of a pseudo-cluster branch and a pseudo-semantic branch. The pseudo-semantic labels are generated in the VTA module and the pseudo-cluster labels are generated through clustering.

The widely used backbones, ResNet50 [9] and BNNeck [10] are adopted in the pseudo-cluster branch. The pseudo-semantic branch uses the spatial and temporal attention (STA) module as in [11] because occlusion may make it more difficult to distinguish between the colors of the tops and bottoms.

$$F_{in} = \sigma(W * \mathrm{Res}(V) + b) \tag{2}$$

$$F_{out} = \mathrm{avgpool}(\sigma(\mathrm{Tem}(\mathrm{Spa}(F_{in}))) \otimes F_{in}) \tag{3}$$

where $F_{in} \in \mathbb{R}^{T\times H\times W\times C_{in}}$ represents the feature map of video $V \in \mathbb{R}^{T\times H'\times W'\times 3}$. This map is extracted using the shared base $Res(\cdot)$ and convolution operation $*$ with weight W and bias b. σ represents ReLU activation function. The output of the STA module is element-wise multiplied with F_{in} using $\otimes$. $\mathrm{Spa}(\cdot)$ represents spatial attention and $\mathrm{Tem}(\cdot)$ represents temporal attention. $\mathrm{Spa}(\cdot)$ takes F_{in} as input and output the spatial attention vector which is processed by a two-dimensional convolutional layer. $\mathrm{Tem}(\cdot)$ then uses a 1-d convolutional layer to aggregate the temporal information and outputs the temporal attention vector. Finally after average pooling the output feature $F_{out} \in \mathbb{R}^{1024}$ is used to predict the pseudo-semantic labels.

Even with the inclusion of a new branch, testing only uses the person re-ID model, which includes the shared base and the pseudo-cluster branch, so there is no appreciable computational cost on inference. The person re-ID model is improved by cross-modal vision-language knowledge in pseudo-semantic labels by optimizing the shared base.

Asymmetric training strategy. The pseudo-semantic labels will lag in subsequent training even if the vision-language model has far more cross-modal information since they are produced by a pre-trained model that stays frozen, whereas the pseudo-cluster labels are steadily improved throughout training. Consequently, we merely train both branches during the first phase; after that, the pseudo-semantic branch is frozen. Relevant experiments are shown in Sec. IV-C.

We use the commonly used cross-entropy loss L_{id} and the triplet loss L_{tri} in the pseudo-cluster branch. Based on the pseudo-semantic labels, the pseudo-semantic branch uses the cross-entropy loss $L_{semantic}$ to classify the video into predetermined options for all attributes. The overall loss is:

$$L_{total} = \begin{cases} L_{semantic} + L_{id} + L_{tri} & ep < K \\ L_{id} + L_{tri} & \text{others} \end{cases} \tag{4}$$

where ep represents the training epoch and the threshold K is the frozen point for asymmetric training.

IV. Experiments

A. Experimental Setup

Datasets. We evaluate our framework on two unsupervised video person re-ID benchmarks. The MARS [6] dataset contains 1,261 identities captured by six cameras. The DukeMTMC-VideoReID (DukeV) [7] dataset is captured by 8 cameras and contains 4,832 tracklets of 1,404 identities.

TABLE I
PERFORMANCE COMPARISONS ON **MARS** AND **DUKEMTMC-VIDEOREID**.

Method	Type	MARS				DukeMTMC-VideoReID			
		Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
TAUDL [12]	tracklet-based	43.8	59.9	72.8	29.1	-	-	-	-
CCM [13]		65.3	77.8	81.3	41.2	76.5	89.6	91.9	68.7
TSSL [14]	others	56.3	-	-	30.5	73.9	-	-	64.6
uPMnet [15]		-	-	-	-	83.6	93.1	**97.2**	76.9
EUG [7]	cluster-based	62.7	74.9	82.6	42.5	72.8	84.2	91.5	63.2
SSL [16]		62.8	77.2	80.1	43.6	76.4	88.7	91.0	69.3
NHAC [17]		61.8	75.3	79.9	40.1	82.8	92.7	95.6	76.0
SRC [18]		62.7	76.1	80.0	40.5	83.0	93.3	95.0	76.5
TMC [19]		72.4	85.9	89.3	61.8	-	-	-	-
AuxUSLReID [2]		80.9	-	-	**72.4**	85.3*	-	-	81.1*
VLUReID	cluster-based	**82.2**	**91.7**	**94.3**	72.2	**87.5**	**96.3**	**97.2**	**85.1**
Improvement	v.s. second-best	+1.3	+5.8	+5.0	-0.2	+2.2	+3.2	+0.0	+4.0

*AuxUSLReID with auxiliary information exploiting modules omitted is used in DukeMTMC-VideoReID for fair comparison.

Evaluation Metrics. Mean average precision (mAP) and Rank-*k* accuracy are adopted as metrics.

Implementation Details. The pseudo-cluster labels are clustred based on density. The clustering settings are the same as [2]. The options for attributes tops and bottoms colors include 'white', 'purple', 'black', 'red', 'green', 'orange', 'yellow', and 'blue'. For the image style attribute, the options are set to 'male' and 'female'. The frozen point of asymmetric training K is set to 10 for DukeV and 50 for MARS. On two NVIDIA GeForce RTX 3080 GPUs, the model is trained. The number of training epochs is set to 60. The learning rate is first initialized with 0.00035 and is divided by 10 at the 50^{th} epoch. The training batch size is set to 32 video tracklets and four images are randomly sampled for each tracklet. We use the past temporal average model of our model for stable training, and its outputs also supervise our model, referred to as the soft version of each loss. The weights assigned to the soft version of each loss are denoted as λ_{softID}, $\lambda_{softTri}$, and λ_{softSe}. The former two are set to 0.5.

B. Comparisons with State-of-the-Art Methods

We compare our framework with unsupervised video person re-ID methods in Tab. I, including cluster-based methods [2], [7], [16]–[19]. Apart from cluster-based methods, tracklet-based methods [12], [13] leveraging camera information to establish the association between videos are also popular in unsupervised settings. Wu *et al.* [14] and Zang *et al.* [15] are also compared. It is worth mentioning that the distances between vectors are directly calculated in [7], [16], and we also classify it as cluster-based methods.

On MARS dataset, Teng *et al.* [2] propose to incorporate auxiliary information and training tricks. Our VLUReID framework achieves a 1.3% improvement in Rank-1 accuracy, clearly outperforming this method. The results show that our framework is proficient in finding the highest-ranked matches. Although our framework shows a slight decline in mAP accuracy of 0.2%, the larger improvement in Rank-1 accuracy highlights our framework's ability to extract the most pertinent and accurate matches, which is especially useful in situations where the accuracy of top-ranked matches is important.

TABLE II
RESULTS ON KEY MODULES ON MARS.

Model	Rank-1	Rank-5	Rank-10	mAP
SB	79.9	89.6	91.9	69.3
DB w/ RN50	81.0	90.4	92.2	70.1
DB w/ STA	**82.2**	**91.7**	**94.3**	**72.2**

TABLE III
RESULTS ON FROZEN POINT FOR ASYMMETRIC TRAINING (K).

K	Rank-1	Rank-5	Rank-10	mAP
0	86.5	93.9	95.7	82.3
10	**87.7**	95.0	96.7	**84.6**
20	**87.7**	95.4	96.6	84.3
30	87.2	**95.6**	96.7	83.7
40	84.9	94.7	96.6	81.8
50	85.0	94.9	**97.0**	82.9

On DukeV dataset, Teng *et al.* [2]'s approach also demonstrates a noteworthy performance. Even with these significant improvements, our framework still achieves an improvement of 4.0% mAP and 2.2% Rank-1 accuracy. When one takes into account the constant performance gains shown both in the MARS and Duke datasets, all of these findings highlight the effectiveness of our framework.

C. Ablation Study

We examine how well the dual-branch network design with the STA module and cross-modal vision-language knowledge function. We also do ablation research on the information distillation method and the value of the frozen point K. Experiments are conducted on DukeV unless otherwise noted.

Effectiveness of leveraging cross-modal vision-language knowledge. We evaluate our framework against the basic single branch model, which consists of the pseudo cluster branch and the shared base, denoted as 'SB' in Tab. II, where 'SB' represents only the pseudo-cluster branch while 'DB' represents the dual-branch network with both branches. On MARS dataset, VLUReID, denoted as 'DB w/ STA', improves mAP by 2.9% and Rank-1 accuracy by 2.3%. Since our basic model has performed quite well on these two datasets, the

TABLE IV
RESULTS ON THE KNOWLEDGE DISTILLATION METHOD.

model	Rank-1	Rank-5	Rank-10	mAP
Basic model	77.1	91.3	94.6	73.5
VLUReID w/ CC.*	64.7	79.5	83.5	57.9
VLUReID w/o CC.	**77.8**	**92.0**	**95.0**	**74.4**

* 'CC.' represents concatenation.

TABLE V
ABLATION STUDY ON THE WEIGHTS OF SOFT SEMANTIC LOSS.

λ_{softSe}	Rank-1	Rank-5	Rank-10	mAP
0.1	87.3	93.9	95.7	83.8
0.3	86.5	94.4	96.0	83.2
0.5	**87.7**	95.0	96.7	84.6
0.7	87.5	**96.3**	**97.2**	**85.1**
0.9	87.2	94.9	96.4	84.1

improvement of this model shows the effectiveness of the incorporation of cross-modal vision-language information.

Effectiveness of the STA module. The pseudo-semantic branch and the pseudo-cluster branch have the same ResNet50 architecture in version 'DB w/ RN50'. The performances of both models excel in those of the basic single branch model, as seen in Tab. II, demonstrating the effectiveness of incorporating vision-language knowledge. The dual-branch network including the STA module further improves 2.1% in mAP and 1.2% in Rank-1 accuracy on MARS dataset, illustrating that the introduction of cross-modal knowledge and STA modules both boost performance.

Analysis of the frozen point for asymmetric training. On the DukeV dataset, the performances with various frozen points are shown in Tab. III. The pseudo-semantic branch and the pseudo-cluster branch are both optimized in the first K epochs to have the pseudo-semantic branch frozen then. Best mAP and Rank-1 accuracy are both obtained when K is set to 10. The empirical results show a constant downward trend across the measured range, indicating an inverse association between the frozen point and the associated performance. The results imply that mindlessly forcing the model to be aligned with pseudo-semantic labels would degrade performance when the model performs better in subsequent training stages.

Analysis of knowledge distillation method. The knowledge in pseudo-semantic labels is distilled into person re-ID model by training the shared base in our framework. Enhancing the features at high-level layers is another well-liked method. 'w/CC' in the Tab. IV represents that the output features of the pseudo-semantic branch are concatenated with those of the pseudo-cluster branch after unifying the channels, while 'w/o CC' means the shared base layers should be updated. For a fair comparison, performances at the 20^{th} epoch of each approach are compared, as the pseudo-semantic branch will stay frozen. Results in Tab. IV show that the method without concatenation performs well in both evaluation metrics. This may be due to the different levels of understanding in the prediction of pseudo-semantic labels and pseudo-cluster labels.

Analysis of the weights of soft semantic loss (λ_{softSe}). The

Fig. 3. Heat maps of the CLIP model when prompted by the color of the tops. Warmer colors represent higher saliency.

semantic loss is designed in response to the introduction of the pseudo-semantic branch. We conduct experiments on DukeV to analyze the influence of λ_{softSe}. The results in Tab. V are better than those of the basic model in Tab. IV, indicating that the value of λ_{softSe} does not affect the effectiveness of the pseudo-semantic branch. Furthermore, when λ_{softSe} is set to 0.7, the highest Rank-5, Rank-10, and mAP accuracy are achieved, indicating that in this branch, the soft version of the cross-entropy loss is more important.

The results are in line with the analysis because the purpose of using the temporal averaging model and soft loss is to improve training stability, performance, and generalization. However, when the proportion of the soft version further increases, the model's performance decreases. We can explain from an informational perspective. When the proportion of the soft version becomes too high, it prevents the model from learning knowledge from pseudo-semantic labels, thereby impacting the performance.

D. Qualitative Results

In this subsection, we mainly provide qualitative results of the VTA module to demonstrate the CLIP model's ability.

The purpose of our use of the CLIP model is to generate textual descriptions by associating the visual elements and textual descriptions. We prompt the CLIP model after combining the template, attribute, and its options, and then the heat maps on randomly sampled frames of videos are displayed in Fig. 3. Grad-CAM [20] creates a heat map that reveals the precise regions that the network focuses on for a given category, enabling us to assess if the model associates visual and textual elements effectively. Considering privacy concerns, the facial features of the people in the images are occluded.

The heat map results show that the CLIP model makes predictions based on the corresponding portions in frames. Since the prompts input into CLIP is about the color of the tops, ideally the region with high saliency should be the tops of the person. However, there may be distractors that influence the prediction of the models, for example, the backpack or the accompanies. It can be observed that CLIP successfully eliminates the impact of backpacks when predicting for the person in the first column, demonstrating its capacity to comprehend the relationship between visual elements and textual

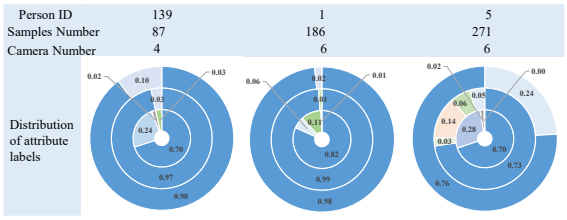


Fig. 4. Distribution of selected options of three attributes of three randomly sampled people. Best viewed in zoom and color.

concepts. It is also important to acknowledge that while this approach is generally effective, the accuracy of the predictions may be compromised in instances where the target object is significantly occluded. However, since we sample every frame of the video for prediction, the impact of occlusion can be mitigated to a certain degree.

We are interested in the distribution of generated pseudo-semantic labels for different videos featuring the same individual. Results for three randomly sampled persons are shown in Fig. 4. There are up to 271 videos for one person, which are captured by multiple different cameras. A doughnut chart is associated with a person and shows the distribution of chosen options for all three attributes across all of the person's videos. In a doughnut chart, the top colors, bottom colors, and image style are represented by circles that go from inside to outside. Within each circle, different colors are utilized to indicate different chosen options for the attributes. The percentile value and area ratio for each color both denote the proportion of this value's number of videos to all of this person's videos. Ideally, a circle should only have one color.

We can observe that every circle has a color that accounts for more than 70%. This implies that at training time, at least 70% of the samples will be correctly associated to guide the model to discover the connections between them. The same values for different videos, especially those captured from different cameras, provide appearance associations between them in training, which is hard to dig out when only the pseudo-cluster labels are used.

V. Conclusion

We demonstrate that cross-modal vision-language knowledge is valuable prior knowledge for person re-identification (re-ID) task. We propose the VLUReID framework for person re-ID that incorporates the cross-modal knowledge in vision-language models without relying on any manual annotations from datasets. The VLUReID framework includes a Vision-to-Text Association (VTA) module and a Dual-Branch Asymmetric Training (DBAT) module. We generate pseudo-semantic labels by prompting the vision-language model using the textual prompts designed for person re-ID tasks in the VTA module. In the DBAT module, the person re-ID model is improved by the cross-modal knowledge contained in pseudo-semantic labels through asymmetric network architecture and training strategy. We evaluate our VLUReID framework on two datasets of unsupervised video person re-ID, and the results demonstrate the effectiveness of cross-modal vision-language knowledge and the superiority of our framework.

References

[1] S. Li, L. Sun, and Q. Li, "Clip-reid: exploiting vision-language model for image re-identification without concrete text labels," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 1, 2023, pp. 1405–1413.

[2] H. Teng, T. He, Y. Guo, and G. Ding, "A high-accuracy unsupervised person re-identification method using auxiliary information mined from datasets," *arXiv preprint arXiv:2205.03124*, 2022.

[3] S. Yan, N. Dong, L. Zhang, and J. Tang, "Clip-driven fine-grained text-image person re-identification," *arXiv preprint arXiv:2210.10276*, 2022.

[4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.

[5] Y. Lin, C. Liu, Y. Chen, J. Hu, B. Yin, B. Yin, and Z. Wang, "Exploring part-informed visual-language learning for person re-identification," *arXiv preprint arXiv:2308.02738*, 2023.

[6] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian, "Mars: a video benchmark for large-scale person re-identification," *european conference on computer vision*, 2016.

[7] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Ouyang, and Y. Yang, "Exploit the unknown gradually: one-shot video-based person re-identification by stepwise learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5177–5186.

[8] Z. Chen, A. Li, and Y. Wang, "A temporal attentive approach for video-based pedestrian attribute recognition," in *Pattern Recognition and Computer Vision: Second Chinese Conference, PRCV 2019, Xi'an, China, November 8–11, 2019, Proceedings, Part II 2*. Springer, 2019, pp. 209–220.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *computer vision and pattern recognition*, 2015.

[10] H. Luo, W. Jiang, Y. Gu, F. Liu, X. Liao, S. Lai, and J. Gu, "A strong baseline and batch normalization neck for deep person re-identification," *IEEE Transactions on Multimedia*, 2019.

[11] Z. Chen, A. Li, S. Jiang, and Y. Wang, "Attribute-aware identity-hard triplet loss for video-based person re-identification," *arXiv preprint arXiv:2006.07597*, 2020.

[12] M. Li, X. Zhu, and S. Gong, "Unsupervised person re-identification by deep learning tracklet association," in *Proceedings of the European conference on computer vision (ECCV)*. Munich, Germany: IEEE, pp. 737–753.

[13] X. Wang, R. Panda, M. Liu, Y. Wang, and A. K. Roy-Chowdhury, "Exploiting global camera network constraints for unsupervised video person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 10, pp. 4020–4030, 2020.

[14] G. Wu, X. Zhu, and S. Gong, "Tracklet self-supervised learning for unsupervised person re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 12 362–12 369.

[15] X. Zang, G. Li, W. Gao, and X. Shu, "Exploiting robust unsupervised video person re-identification," *IET Image Processing*, vol. 16, no. 3, pp. 729–741, 2022.

[16] Y. Lin, L. Xie, Y. Wu, C. Yan, and Q. Tian, "Unsupervised person re-identification via softened similarity learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3390–3399.

[17] P. Xie, X. Xu, Z. Wang, and T. Yamasaki, "Unsupervised video person re-identification via noise and hard frame aware clustering," in *2021 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2021, pp. 1–6.

[18] P. Xie, X. Xu, Z. Wang, and T. Yamasaki, "Sampling and re-weighting: Towards diverse frame aware unsupervised video person re-identification," *IEEE Transactions on Multimedia*, vol. 24, pp. 4250–4261, 2022.

[19] H. Teng, T. He, Y. Guo, Z. Guo, and G. Ding, "A free lunch to person re-identification: learning from automatically generated noisy tracklets," *arXiv preprint arXiv:2204.00891*.

[20] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, 2016.