# Unified Pre-training with Pseudo Texts for Text-To-Image Person Re-identification

Zhiyin Shao[1,2*], Xinyu Zhang[2*], Changxing Ding[1,3†], Jian Wang[2], Jingdong Wang[2]

[1]South China University of Technology, China [2]Baidu VIS, China [3]Pazhou Lab, China
eezyshao@mail.scut.edu.cn chxding@scut.edu.cn
{zhangxinyu14,wangjian33,wangjingdong}@baidu.com

## 1. More discussion

**Discussion on the format of the text descriptions in the symmetric contrastive loss (Eq. (1).**

As shown in the main paper, we utilize the MLM loss $L_{\text{mlm}}$ and the contrastive loss $L_{\text{con}}$ in the pre-training stage. In detail, we randomly mask 15% textual tokens and predict these masked patches with $L_{\text{mlm}}$. For the contrastive loss $L_{\text{con}}$, we simply use the intact text descriptions instead of the masked ones for the optimization.

Here, we product an exploratory experiment: *what if we use the masked text in the contrastive loss?* As illustrated in Table 1, compared with pre-trained model independently, our UniPT which employ the masked text descriptions in the contrastive loss achieves improvements by +1.35% and +1.13% in terms of Rank-1 accuracy on CUHK-PEDES and ICFG-PEDES. Meanwhile, using intact text descriptions in the contrastive loss achieves better performance. The underlying reason may be that the masked text descriptions may leave out some details intuitively, introducing more confusion to the training stage. Therefore, we use the intact text descriptions by default in the main paper.

## 2. Illustration of attributes and corresponding phrases.

As illustrated in Figure 1, we group total 14 attributes into two sets, *i.e.*, the required set containing 6 attributes and the optional set with 8 attributes. Each attribute has its own phrases. It is worth noting that, <upper clothes> is made up of <upper_wear_texture>, <upper_wear_color> and <upper_wear>, and <lower clothes> consists of <lower_wear_color> and <lower_wear>. The upper and lower clothes are the most distinctive features of a person, and we ensure the diversity of them by means of permutation and composition.

*Equal contribution. †Corresponding author.

| 2*Method | CUHK-PEDES | | ICFG-PEDES | |
|---|---|---|---|---|
| | Rank-1 | Rank-5 | Rank-1 | Rank-5 |
| Single | 65.30 | 83.19 | 57.60 | 74.53 |
| UniPT w/ masked text | 66.65 | 83.89 | 58.73 | 75.08 |
| UniPT | **66.83** | **84.16** | **59.08** | **75.92** |

Table 1. Comparison with the masked (2nd row) and intact (3rd) text descriptions in the symmetric contrastive loss (Eq. (1) in the main paper). By default, we use the intact text descriptions.

## 3. Qualitative evaluation

We conduct qualitative evaluation for our proposed UniPT. Figure 2 shows some examples of top-10 retrieved persons by three different model: (1) the model pre-trained by UniPT. (2) the model pre-trained by ImageNet [3] and [1], separately. (3) the model pre-trained by [2] and [1], separately. It is observed that our model with UniPT can find better images to match text descriptions.

## 4. More visualizations in LUPerson-T.

Figure 3 4 5 show more visualization examples in our proposed LUPerson-T. Each image has a pseudo text description on its right. *We will release the data and code*.

## References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[2] Hao Luo, Pichao Wang, Yi Xu, Feng Ding, Yanxin Zhou, Fan Wang, Hao Li, and Rong Jin. Self-supervised pre-training for transformer-based person re-identification. *arXiv preprint arXiv:2111.12084*, 2021.

[3] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, and Michael Bernstein. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015.

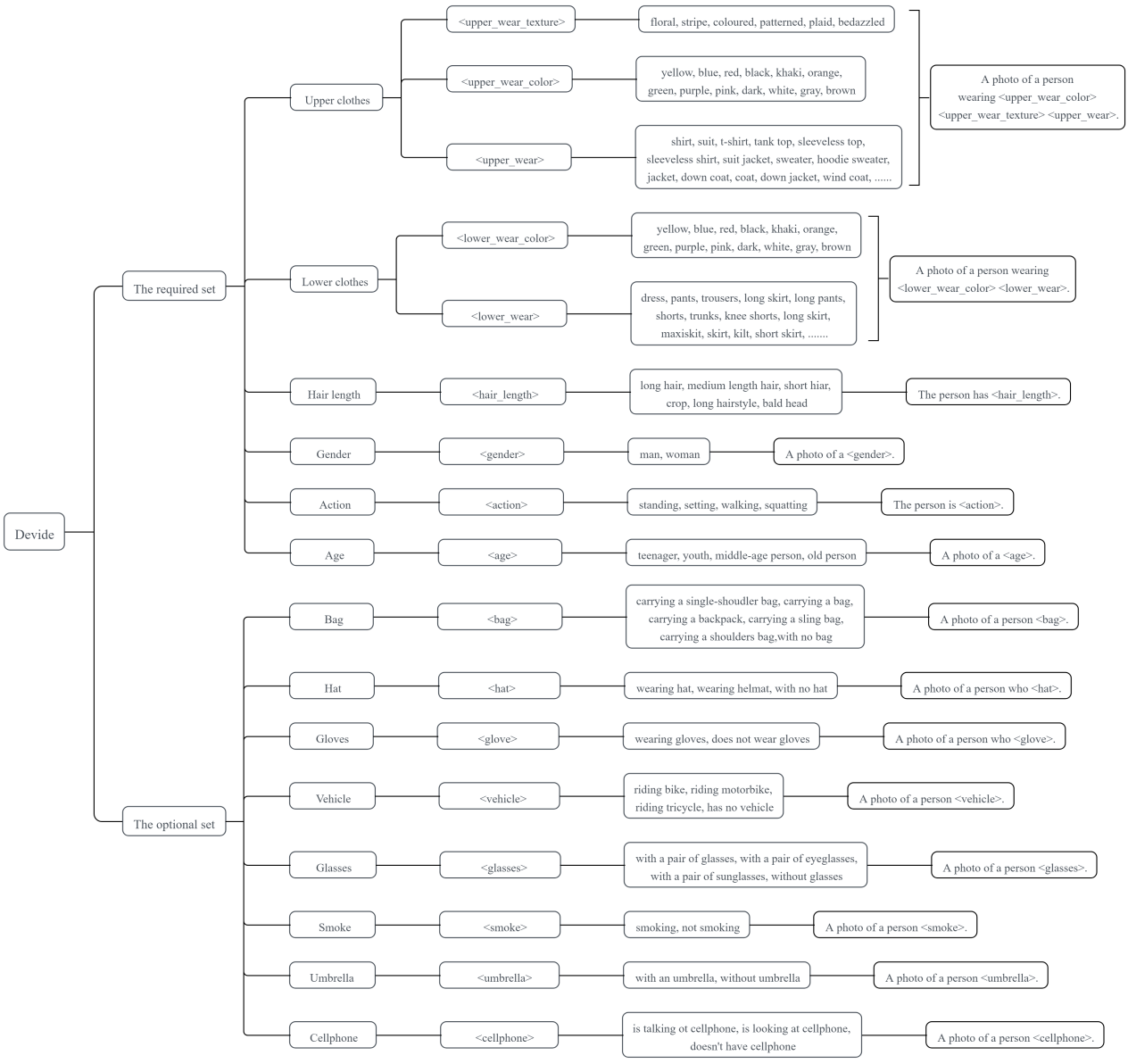


Figure 1. The tree diagram of attributes, phrases and the corresponding prompts.

A young man with short black hair is wearing a blue printed basketball jacket with grey sleeves. He is also wearing blue pants and grey running shoes with white soles. He is carrying a black tote bag in his hand.

(1)

(2)

(3)

A man carrying a black backpack is wearing a dark blue hooded coat and black jens with dark colored shoes with a white sole.

(1)

(2)

(3)

A man wearing a light brown color hoodie with blue jeans with blue and sneakers with white edging. He has black short hair, carries a grey color side bag.

(1)

(2)

(3)

Figure 2. Examples of top-10 person search results by three different re-trained model: (1) The visual and textual backbones are pre-trained by LUPerson-T. (2) The visual and textual backbones are pre-trained by ImageNet [3] and [1]. (3) The visual and textual backbones are pre-trained by [2] and [1].


Figure 3. Examples in LUPerson-T.


Figure 4. Examples in LUPerson-T.

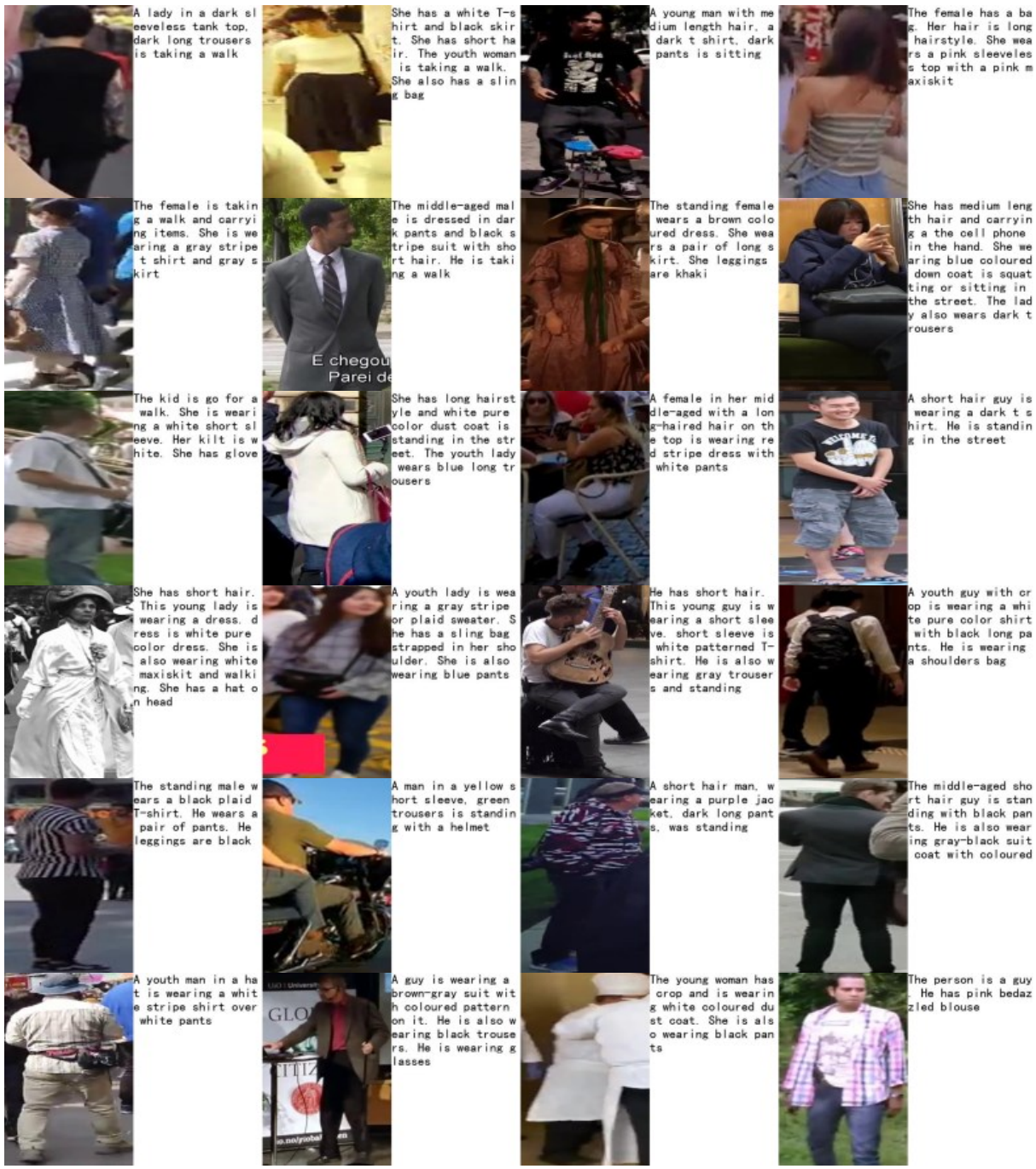


Figure 5. Examples in LUPerson-T.