Supplemental Material of VILLS: Video-Image Learning to Learn Semantics for Person Re-Identification

Siyuan Huang, Ram Prabhakar, Yuxiang Guo, Rama Chellappa, Cheng Peng
Johns Hopkins University
{shuan124, rprabha3, yguo87, rchella4, cpeng26}@jhu.edu

1. Acknowledgments

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via [2022-21102100005]. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U. S. Government. The US. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein. The authors thank Drs. Josh Gleason, Matt Meyn, Nathan Shnidman, and Soraya Stevens for helpful discussions.

2. Discussion of Potential Negative Societal Impact

The authors affirm that all datasets utilized in this paper originate from public sources or have been approved by the subjects themselves. This research adheres to ethical guidelines and does not raise privacy or safety concerns. The objective of this paper is to enhance advancements in smart city applications and autonomous driving technologies.

3. Fine-tuning and Inference

For downstream ReID tasks, we use only the teacher's shared encoder and the resampler, discarding the student, all heads in the teacher, and the local semantic extraction module. We then fine-tune the teacher using ReID task losses, including cross-entropy loss for classifying different identities and triplet loss [5] for clustering the same identity.

During inference, VILLS automatically extracts all features. For an input image, VILLS extracts coarse-grained and semantically consistent features. For an input video, VILLS also extracts temporal features. These features are then used for person matching, retrieval, or visualization based on the specific task.

4. Implementation Details

For the Unified Feature Learning and Adaptation module, we use a Vision Transformer (ViT) [2] as the shared encoder backbone. The resampler is a Perceiver Transformer [7], which consists of a small transformer layer with cross attention. All heads are constructed using Multi-Layer Perceptrons with Batch Normalization [6]. For the local semantic extraction (LSE) module, we use a pre-trained Mask R-CNN [4] from the COCO dataset [9], while the interactive segmentation model is based on the pre-trained Segment Anything Model [8].

During pre-training, we train ViT-S and ViT-B on 8$\times$ A40 GPUs for 100 epochs. The training process takes approximately 100 and 200 hours for ViT-S and ViT-B, respectively. We use video batch sizes of 32 (ViT-S) and 16 (ViT-B), and image batch sizes of 128 (ViT-S) and 64 (ViT-B). The video frame size is set to 64×44, with each video randomly sampling 8 frames. The image size is 256×128. The LSE-derived feature size is 128×64. The local areas are defined as head, upper body, and lower body. The balancing parameters λ_1, λ_2, λ_3, and λ_4 are set to 1.0, 1.0, 3.0, and 2.0, respectively.

For downstream ReID tasks, we follow standard settings for each task and dataset. In PRCC and LTCC, the input size is 384×192. In Market1501, the input size is 256×128. In PRID2011 and MARS, the input size is $8 \times 256 \times 128$. In BRIAR-2, BRIAR-3, and BRIAR-4, the input is 384×128 for images and $8 \times 384 \times 128$ for videos. Unless otherwise specififed, all main results are conducted using ViT-B, while ablation studies are conducted using ViT-S.

5. Visualization of Attention Maps

Fig. 1 compares attention maps between our method and others across different tasks. For image-based ReID, our coarse-grained spatial features outperform others. While existing methods lack a complete semantic concept of identity and focuses on peripheral parts, the proposed method clearly captures a complete identity with accurate focus. Moreover, our attention is semantically consistent, showing

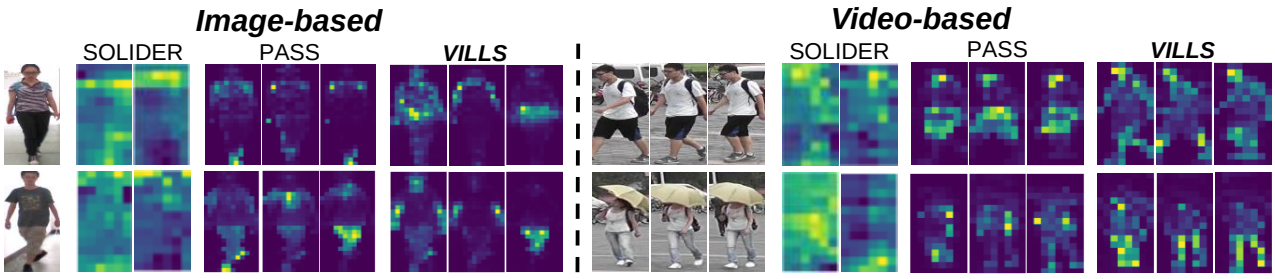

Figure 1. Visualization of attentions across different ReID methods. For SOLIDER, each attention map represents windowed attention. In image-based ReID, the first attention map for both PASS and VILLS is derived from coarse spatial features. The second and third attention maps for PASS are generated from local features, while for VILLS, they are generated from the LSE-derived features. In video-based ReID, all attention maps for PASS are derived from local features, whereas VILLS utilizes temporal features for all its attention maps. Notably, VILLS demonstrates semantically consistent attention patterns across both images and videos, highlighting its unified method to feature extraction.

clear focus on specific body parts (e.g., arms, thighs). These results demonstrate the effectiveness of our method, particularly the LSE module, in extracting semantically consistent spatial features.

In video-based ReID, most methods lack temporal features, resulting in incomplete attention that fails to connect across frames. In contrast, the attention behavior in VILLS shows clear motion patterns highly consistent with the original video. Furthermore, the attention consistently focuses on the most significant motion parts of the identity. These results highlight the effectiveness of UFLA module in successfully extracting temporal features. In summary, these visualizations showcase VILLS' ability to effectively and seamlessly extract both spatial and temporal features. These visualizations provide strong qualitative evidence for the effectiveness of VILLS in capturing semantically consistent and modality-appropriate features across various ReID tasks.

6. Test Accuracy Curves

Fig. 2 illustrates the test accuracy curves for our method and state-of-the-art (SOTA) methods on image-based ReID. Notably, our method achieves a rank-1 accuracy of 58.4% by epoch 3, outperforming other methods. In comparison, PASS (Zhu et al. 2022) reaches a rank-1 accuracy of 52.4% in epoch 11, while CAL [3] achieves 55.3% in epoch 39, demonstrating the effectiveness of our method.

A comparison between pre-training methods (PASS [11], SOLIDER [1], HAP [10], and VILLS) and methods specifically designed for this dataset (e.g., CAL [3]) reveals that pre-training methods converge faster. However, their performance falls short of dataset-specific methods. VILLS stands out by not only surpassing other pre-training methods but also outperforming dataset-specific methods. Moreover, VILLS converges in fewer epochs, highlighting both its effectiveness and efficiency.

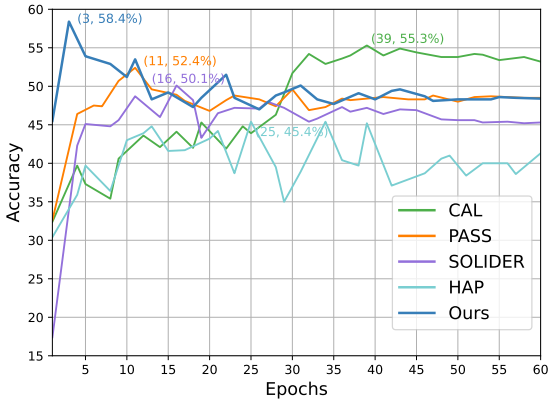

Figure 2. Test accuracy curves for various methods on the PRCC dataset. VILLS achieves the highest performance in the fewest epochs. This experiment was conducted using the image-only version of VILLS with ViT-B.

References

[1] Weihua Chen, Xianzhe Xu, Jian Jia, Hao Luo, Yaohua Wang, Fan Wang, Rong Jin, and Xiuyu Sun. Beyond appearance: a semantic controllable self-supervised learning framework for human-centric visual tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15050–15061, 2023. 2

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[3] 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

[4] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1

[5] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 1

[6] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015. 1

[7] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021. 1

[8] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 1

[9] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1

[10] Junkun Yuan, Xinyu Zhang, Hao Zhou, Jian Wang, Zhongwei Qiu, Zhiyin Shao, Shaofeng Zhang, Sifan Long, Kun Kuang, Kun Yao, et al. Hap: Structure-aware masked image modeling for human-centric perception. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[11] Kuan Zhu, Haiyun Guo, Tianyi Yan, Yousong Zhu, Jinqiao Wang, and Ming Tang. Pass: Part-aware self-supervised pre-training for person re-identification. In *European Conference on Computer Vision*, pages 198–214. Springer, 2022. 2