Spatio-Temporal Representation Factorization for Video-based Person Re-Identification (Supplementary Material)

Abhishek Aich*[,2], Meng Zheng[1], Srikrishna Karanam[1], Terrence Chen[1],
Amit K. Roy-Chowdhury[2], and Ziyan Wu[1]
[1]United Imaging Intelligence, Cambridge MA, [2]University of California, Riverside CA
{aaich001@, amitrc@ece.}ucr.edu, {first.last}@united-imaging.com

CONTENTS

List of Tables

List of Figures

* This work was done during Abhishek Aich's internship with United Imaging Intelligence. Corresponding author: Srikrishna Karanam.

(a) DukeMTMC-VideoReID [15] (b) MARS [17] (c) iLIDS-VID [14]

Figure 1: **Sample tracklets from the (a) DukeMTMC-VideoReID, (b) MARS, (c) iLIDS-VID datasets.** Rows correspond to different persons. As seen from the tracklets, video-based person re-identification is a challenging problem due to occlusions, similar appearances in different identities, and misaligned frames. Best viewed in color.

1. Simplified Demonstration of STRF

We present a simplified demonstration of our proposed framework STRF. STRF is designed to extract four types of information from input feature maps. Intuitively, STRF learns:

I. What is static temporally or changing slowly in time (e.g., how people look, *TS*)
II. What is changing temporally or dynamic in time (e.g., how people move, *TD*)
III. What is coarsely observable spatially (e.g., global appearance/outline, *SC*)
IV. What is finely observable spatially (e.g., fine appearance details, *SF*)

Each "factor" above has its own contribution. For instance, *TD* can provide robust features when people can only be distinguished based on motion/dynamics (e.g., same dress code). Under frame misalignment, *TS* (with *SF*/*SC*) can provide person-specific features while suppressing background/occlusions. The questions, then, are

Q1. How are they learned?
Q2. How to "weight" the input feature map using them?

Factor-specific pooling functions $\mathcal{G}_{dk}(\cdot)$ help answer **Q1** above. Given input feature map $\boldsymbol{f}_{\ell}^{(i)} \in \mathbb{R}^{c_\ell \times f_\ell \times h_\ell \times w_\ell}$ (e.g., channels $c_\ell = 2048$, frames $f_\ell = 8$, height $h_\ell = 14$, width $w_\ell = 7$), each $\mathcal{G}_{dk}(\cdot)$ operates differently. For instance, $\mathcal{G}_{\boldsymbol{t}_S}(\cdot)$ of *TS* uses a $4 \times 1 \times 1$ kernel (with stride 4) to give intermediate feature map $\boldsymbol{f}_{\ell}^{(p)} \in \mathbb{R}^{c \times 2 \times h \times w}$, i.e., temporally pooling 8 into 2 feature maps to capture what changes slowly over time. On the other hand, *TD*'s $\mathcal{G}_{\boldsymbol{t}_T}(\cdot)$, with kernel $2 \times 1 \times 1$, gives $\boldsymbol{f}_{\ell}^{(p)} \in \mathbb{R}^{c \times 4 \times h \times w}$, i.e., more temporal feature maps (*i.e.* 4) since one needs more data points to capture what is changing dynamically (compared to *TS* above) in time. Similar argument holds for *SF*/*SC* spatially. Finally, the factor-specific attention map $\mathcal{M}_{dk}$ helps weight feature volumes appropriately using matrix multiplication in eq. (1) (`main paper`) towards our objective function (helping answer question **Q2** above). This shows why 4 pooling functions are necessary. As each FFM has unique $\mathcal{G}_{dk}(\cdot)$, they are different and help in focusing different aspects of information available in input feature maps.

2. Datasets Details

In this section, we provide more details for each of the three datasets, MARS [17], DukeMTMC-VideoReID [15], and iLIDS-VID [14], used in the paper. Sample video tracklets for each are shown in Figure 1.

- MARS [17]: MARS is a large-scale multi-camera (six views) dataset, comprising 17503 tracklets corresponding to 1261 identities, with an average number of 59 frames per tracklet. Of the 1261 identities, 625 identities are used for training and the rest for testing. Additionally, it has 3248 distractor tracklets to be used as part of the gallery. Each bounding box is detected and subsequently tracked using the DPM detection [5] and GMCP tracking [2] algorithms, respectively.

- DukeMTMC-VideoReID [15]: DukeMTMC-VideoReID is part of the DukeMTMC tracking dataset [12], comprising 1812 identities of which 702 are used for training, 702 for testing, and the rest 408 as distractors. In total, there are 2196 video tracklets for training and 2636 video tracklets for testing. Each frame in the video tracklet is sampled at an interval of 12 frames and has manually annotated bounding boxes.

- iLIDS-VID [14]: iLIDS-VID is a two-camera-view dataset comprising 600 video tracklets with 300 identities with an average of 73 frames per identity and manually annotated bounding boxes.

3. Implementation Details

Hyperparameters Details. We build our feature extractors by first inflating 2D-ResNet50 [7], pre-trained on ImageNet [3], with time dimension of all kernels set to 1 (See Figure 2(A) in main manuscript). The last stride of the model is set to 1 following [13, 16]. Then, we replace stage 2 and 3 with the proposed STRF-P3D residual blocks. We train our model with the Adam [9] optimizer with a weight decay of 0.0005 for 250 epochs. The initial learning rate is set to 0.0003, and is reduced by a factor of 10 times after every 50 epochs. For data augmentation, we use random erasing [18] and random horizontal flip following [1, 8]. As part of each training batch, we randomly sample 4 frames with a stride of 8 frames to form a clip for each tracklet. Each batch contains 8 persons with 4 video clips each. All the frames are resized to 256×128. The feature dimension is set to 2048 which is obtained after temporal pooling for both training and testing. We use PyTorch [10] for all our experiments. Training time is $\sim$10 hrs on 3 NVIDIA Tesla-V100 GPUs.

Testing Protocol. For fair comparisons, we follow exact testing protocols as in prior works [6, 8]. We split each video tracklet into several four-frame clips and extract their feature representations. The final feature representation is computed by averaging across all the clips. Finally, for retrieval, cosine distances are computed between query and gallery video features.

4. Additional Discussions on STRF

Location of STRF in Pseudo-3D [11] residual blocks. We observe in our preliminary experiments that STRF is more effective with the $3 \times 1 \times 1$ convolutional layer rather than the $1 \times 3 \times 3$ convolutional layer (see Figure 1). Hence, we place the STRF module with the $3 \times 1 \times 1$ convolutional layer as indicated in Figure 2(B) of the main manuscript. One explanation for this behavior of STRF can be attributed to the fact that time-degenerate convolutions are more effective in extracting rich information of temporal dimension which has shown to be more important for recognition in [1, 4]. Moreover, the temporal integrity is diminished with $1 \times 3 \times 3$ as each feature map in the volume is treated individually. Hence, after the proposed enhancement of the feature volume, the $3 \times 1 \times 1$ convolutional layer performs comparatively well.

Additional analysis of STRF on different stages of feature extractor. We present additional analysis of the impact of adding the proposed STRF module at various stages to the baseline model in Table 2. We can observe that the STRF module is effective at every stage to enhance the performance of the baseline model.

MODEL	STAGE	mAP (%)	R@1 (%)
Baseline		83.10	88.50
Baseline + STRF	2	85.40	89.70
	3	85.20	89.80
	3, 4	84.00	89.40
	2, 3, 4	85.30	90.10

Table 1: **Per-stage influence of STRF.** STRF is effective at various stages of STRF-P3DC on MARS [17].

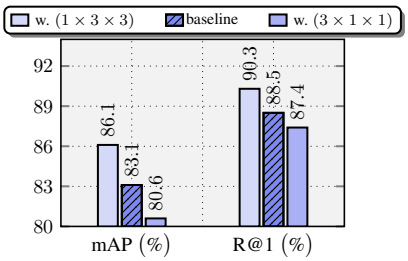


Figure 2: **Location of STRF.** Our STRF module performs the best with $3 \times 1 \times 1$ compared to $1 \times 3 \times 3$ as demonstrated here on MARS [17].

Additional analysis of STRF's four factorization components. We present additional analysis of the different combinations of factorization modules of STRF in Table 2.

Table 2: **Contribution of each factorization module.** Additional analysis of STRF's four factorization components with the P3DC baseline on MARS [17].

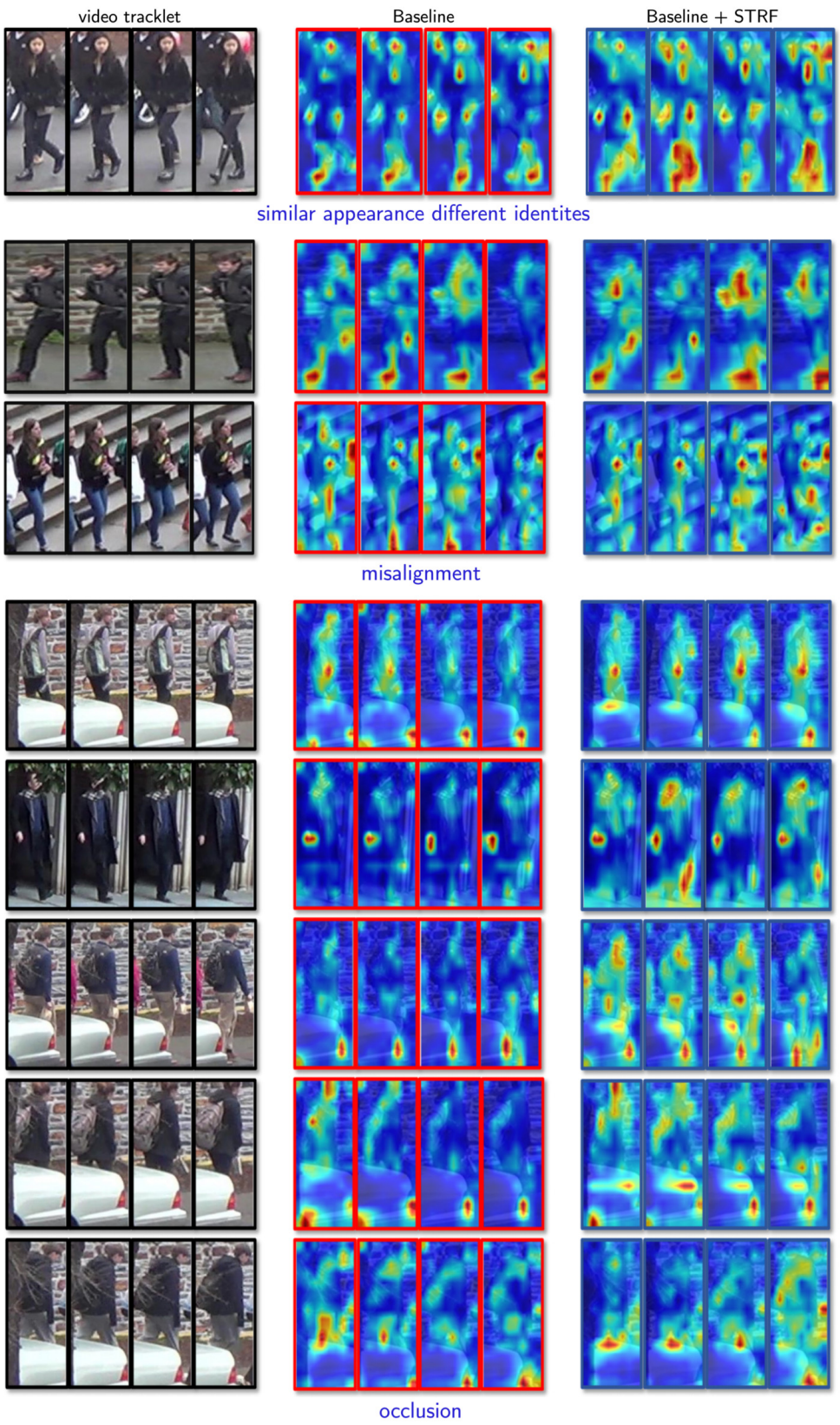
(s, τ)	(s, ς)	(t, τ)	(t, ς)	**mAP**(%)	**R@1**(%)
✓			✓	85.40	89.50
	✓	✓		85.30	89.60
✓	✓		✓	85.60	90.10
✓	✓	✓		85.70	90.20
✓		✓	✓	85.40	89.80
	✓	✓	✓	85.60	90.00

5. Attention Maps

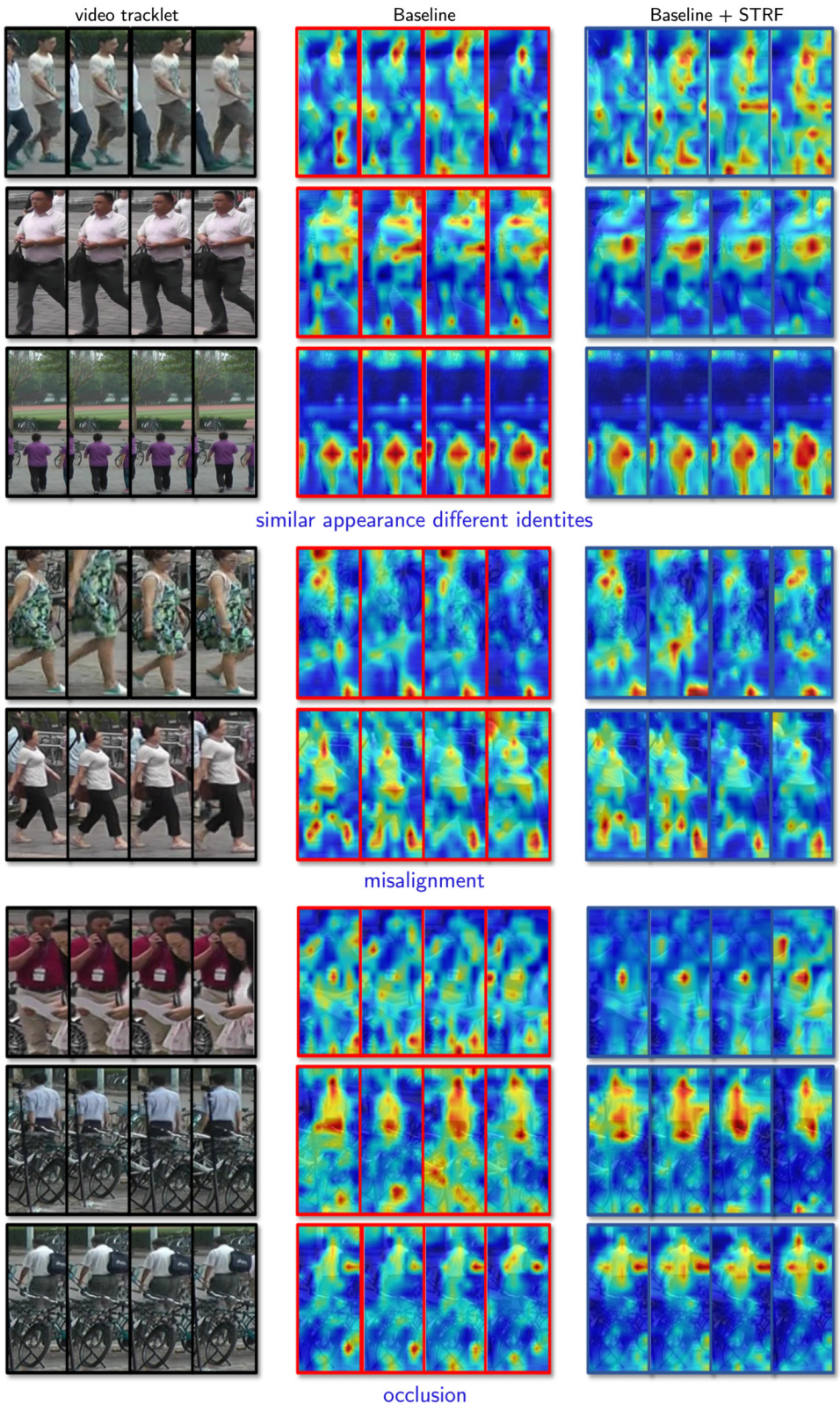
In this section, we present the efficacy of the proposed STRF module in challenging real-world scenarios of occlusion, frame misalignment, and different identities with similar appearance. From Figure 3(a) (DukeMTMC-VideoReID) and Figure 3(b) (MARS), it can be observed that STRF is able to locate the person of interest more precisely when employed with the baseline model. Note that these attention maps are obtained from stage `3` of the feature extractor as we add our proposed module here.

6. Qualitative Results

In this section, we present some cases where the baseline model was unable to find the right match of the query in the gallery (see Figure 4(a) for DukeMTMC-VideoReID and Figure 4(b) for MARS) in Rank-1 retrieval. It can be observed that our proposed module helps to enhance the ability of the baseline model to identify the person of interest in difficult examples.



(a) DukeMTMC-VideoReID [15]



(b) MARS [17]

Figure 3: **More attention maps visualization on DukeMTMC-VideoReID [15], MARS [17].** We present attention maps corresponding to different real-world challenges where our proposed module STRF enabled the baseline (P3DB for DukeMTMC-VideoReID [15], P3DC for MARS [17]) to correctly locate the person of interest in the video tracklet. Best viewed in color.

(a) DukeMTMC-VideoReID [15]

(b) MARS [17]

Figure 4: **Rank-1 (R@1) retrieval results in challenging scenarios on DukeMTMC-VideoReID [15], MARS [17].** We present R@1 retrieval cases where our proposed module STRF enabled the baseline (P3DB for DukeMTMC-VideoReID [15], P3DC for MARS [17]) to correctly identify the query in the gallery. Red bounding boxes indicates incorrect retrieval. Green bounding boxes indicates correct retrieval. Blue indicates labels. Best viewed in color.

References

[1] Guangyi Chen, Yongming Rao, Jiwen Lu, and Jie Zhou. Temporal Coherence or Temporal Motion: Which is More Critical for Video-based Person Re-identification? In *Proceedings of the European Conference of Computer Vision*, 2020. 3

[2] Afshin Dehghan, Shayan Modiri Assari, and Mubarak Shah. GMMCP tracker: Globally Optimal Generalized Maximum Multi-Clique Problem for Multiple Object Tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4091–4099, 2015. 2

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009. 3

[4] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. SlowFast Networks for Video Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6202–6211, 2019. 3

[5] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object Detection with Discriminatively Trained Part-based Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2009. 2

[6] Xinqian Gu, Hong Chang, Bingpeng Ma, Hongkai Zhang, and Xilin Chen. Appearance-Preserving 3D Convolution for Video-based Person Re-identification. In *Proceedings of the European Conference of Computer Vision*, 2020. 3

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 3

[8] Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Temporal Complementary Learning for Video Person Re-Identification. In *Proceedings of the European Conference of Computer Vision*, 2020. 3

[9] Diederik P Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*, 2014. 3

[10] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*, pages 8026–8037, 2019. 3

[11] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning Spatio-Temporal Representation with Pseudo-3D Residual Networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5533–5541, 2017. 3

[12] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance Measures and a Data set for Multi-Target, Multi-Camera Tracking. In *European Conference on Computer Vision*, pages 17–35. Springer, 2016. 2

[13] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond Part Models: Person Retrieval with Refined Part Pooling (and A Strong Convolutional Baseline). In *Proceedings of the European Conference of Computer Vision*, 2018. 3

[14] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person Re-Identification by Video Ranking. In *Proceedings of the European Conference on Computer Vision*, pages 688–703. Springer, 2014. 1, 2

[15] Yu Wu, Yutian Lin, Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Exploit the Unknown Gradually: One-Shot Video-based Person Re-Identification by Stepwise Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5177–5186, 2018. 1, 2, 5, 6, 7

[16] Jinrui Yang, Wei-Shi Zheng, Qize Yang, Ying-Cong Chen, and Qi Tian. Spatial-Temporal Graph Convolutional Network for Video-based Person Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3289–3299, 2020. 3

[17] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. MARS: A Video Benchmark for Large-Scale Person Re-Identification. In *Proceedings of the European Conference on Computer Vision*, pages 868–884, 2016. 1, 2, 3, 6, 7

[18] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random Erasing Data Augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13001–13008, 2020. 3