Dense Interaction Learning for Video-based Person Re-identification Supplementary Materials

Tianyu He[1], Xin Jin[2], Xu Shen[1], Jianqiang Huang[1], Zhibo Chen[2], and Xian-Sheng Hua[1]
[1]DAMO Academy, Alibaba Group
[2]University of Science and Technology of China
timhe.hty@alibaba-inc.com

Contents

1. Datasets Details

We evaluate our Dense Interaction Learning (DenseIL) on several commonly adopted video-based person re-ID benchmarks, including MARS [43], DukeMTMC-VideoReID (DukeV) [27, 36] and iLIDS-VID [34]. We give detailed statistics of three datasets as follows.

Datasets	MARS [43]	DukeV [27, 36]	iLIDS-VID [34]
# Identities	1,261	1,404	300
# Sequences	20,715	4,832	600
# Boxes	1,067,516	815,420	42,460
# Frames	58	168	73
# Cameras	6	8	2
# Detector	DPM	Hand	Hand

Table 1: The statistics of video-based person re-ID datasets.

MARS [43]. It is a large-scale video-based person re-identification (re-ID) benchmark dataset with 17,503 sequences of 1,261 identities and 3,248 distractor sequences. All sequences are captured by 6 cameras. There are 625 identities in the training set and 636 identities in the testing set. The bounding boxes are detected with DPM detector [7], and tracked using the GMMCP tracker [6]. It is one of the most challenging datasets due to the failure of detection or tracking.

DukeMTMC-VideoReID (DukeV) [27, 36]. This dataset is also a large-scale benchmark introduced for video-based person re-ID derived from the DukeMTMC dataset [27]. It comprises 4,832 tracklets of 1,404 identities and 408 distractor identities, where each pedestrian image are cropped from the videos for 12 frames every second. Each track contains 168 frames on average. The dataset is divided into 408, 702 and 702 identities for distraction, training and testing respectively. Detection ground truths are manually labeled.

iLIDS-VID [34]. It is created by observing pedestrians in two cameras. The outputs of two non-overlapping cameras

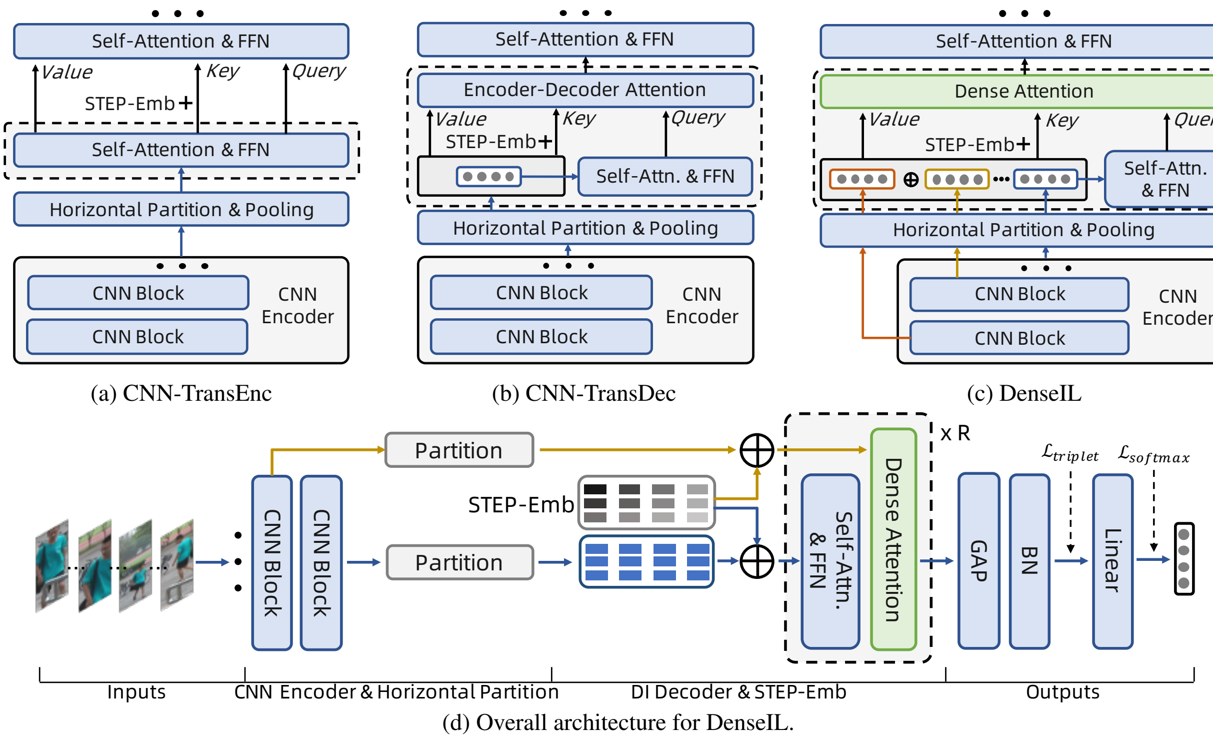


Figure 1: The proposed three model variants for the video-based person re-ID task. (a) The decoder only consists of self-attention (is equivalent to the encoder of vanilla Transformer). (b) The decoder contains both self-attention and encoder-decoder attention (is equivalent to the decoder of vanilla Transformer). (c) Our DI decoder involves self-attention and the proposed Dense Attention (The $\oplus$ denotes the concatenation operation). (d) The detailed architecture for our proposed DenseIL. All schemes are equipped with our proposed STEP-Emb. We omit the layer normalization for simplicity.

are captured at a crowded airport arrival hall. It comprises 600 image sequences of 300 identities with one pair of sequences from two cameras for each person. Each image sequence has a variable length ranging from 23 to 192 image frames, with an average number of 73 images. The bounding boxes are human annotated and the challenge is mainly due to the random occlusions.

In general, MARS and DukeV are large-scale video-based person re-ID benchmarks while iLIDS-VID is relatively small. Conducting experiments on all three datasets with different properties demonstrates a powerful generalization ability for various scenarios.

2. More Implementation Details

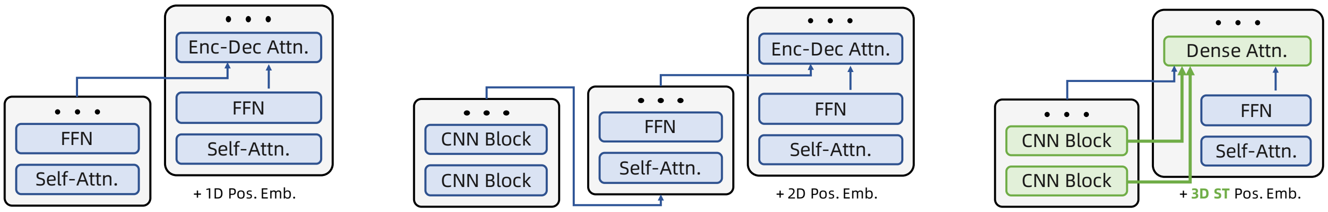
In the main body of the paper, we introduce three model variants for the overall architecture to dive deeply into the CNN-Attention hybrid structure. In this section, we give more details on implementation for the reproducibility, especially for our proposed DenseIL.

2.1. Overall Architecture

Figure 1a, 1b and 1c give detailed operating principle of various attention mechanisms, where the components contained in the dashed boxes can be regarded as basic building blocks to stack up. In Figure 1d, we demonstrate the whole data pipeline of the DenseIL. Each step is described in details in the following:

Inputs. We adopt restricted random sampling strategy [21, 22, 40] to randomly sample frames from equally divided 8 chunks for each video clip. The obtained 8 frames with RGB format are then preprocessed by resizing, random horizontal flips and random erasing for data augmentation before fed into CNN encoder. Note that, according to our experiments, we empirically find that frame-level random horizontal flips (randomly flip for each frame) and sequence-level random erasing (randomly erase the same region for the whole input sequence) achieve the highest performance, which is consistent with previous studies [22, 44, 41, 4]. We therefore apply such data augmentation strategy for all the settings in our experiments.

CNN Encoder & Horizontal Partition. The CNN encoder consists of several building blocks, each block can be an arbitrary CNN structure (*e.g.*, Res-Block [10], Dense-Block [16], *etc.*). For the spatial feature generated by each block, we perform `PPool`$(\cdot)$ on it and thus obtain a feature vector for each partition, as shown in Figure 1d. Note that,



(a) Transformer (b) DETR (c) DenseIL

Figure 2: The brief illustration of the proposed DenseIL compared with the vanilla Transformer architecture [33] and DETR architecture [2]. The components marked as green color denote the differences.

in our implementation, the CNN encoder is first pretrained and then fixed during training the DI decoder.

DI decoder & STEP-Emb. The DI decoder also consists of stacked building blocks (dashed boxes in Figure 1d). It takes partitioned spatial features from preceding CNN blocks as inputs. For each of them, we additionally equip the feature with the proposed STEP-Emb by summation, to explicitly inject information to indicate the absolute or relative position of inputs to the model.

Outputs. Basically, the outputs of the DI decoder share the same dimension with its input. Therefore, we perform spatial-temporal average pooling on the outputs of DI decoder to acquire a feature vector (descriptor) for each video clip. Following the common practice [25, 31, 14, 11, 9], the resulting feature vector is treated by a BatchNorm [17] layer and a linear classifier. We employ batch triplet loss [13] and cross-entropy loss for the features processed after BatchNorm and classifier respectively. In the inference, we use the features generated by BatchNorm layer to measure the cosine distance between two image pairs.

3. Comparison with Transformer-based Model

Attention has enjoyed rich success in tasks such as Neural Machine Translation [1, 33, 12, 37], of which Transformer [33] is the most success one. Inspired by this, someone starts to consider borrowing the entire Transformer architecture to jointly model vision-language representations [32, 24, 30, 26, 5] or exploit relations of the objects in image object detection [2, 46]. Among them, DETR [2] is proposed very recently and attracts lots of attention. DETR [2] is a end-to-end object detection framework that works by building both vanilla Transformer encoder and decoder *on the highest level of* CNN spatial features, as illustrated in Figure 2b. It shares the same high-level insight on leveraging Attention mechanism to model relationship between objects. However, the fine-grained information is still not fully exploited due to its cascaded architecture, while our DenseIL is able to pay attention to multi-scale fine-grained CNN representations by the proposed Dense Attention, as shown in Figure 2c.

4. Complete Performance Comparison

Due to the limited space, we only provide comparison with recently proposed state-of-the-art results in the main body of paper. Here, we give a full version of performance comparison in Table 2. From it, we can easily conclude that our DenseIL enables new top results in all datasets and metrics. In particular, our scheme increases over the existing best performance by 1.1% mAP in MARS dataset, 0.9% mAP in DukeMTMC-VideoReID dataset, 3.4% Rank-1 in iLIDS-VID dataset, demonstrating strong discriminative representation ability and great generalization ability.

5. More Qualitative Analysis

In this section, we provide more qualitative analysis on how DenseIL works. We illustrate the re-identification results of both baseline and our scheme in Figure 3. In each part of Figure 3, the left column is sampled frames of query sequence and the right five columns are the sampled frame of top-5 retrieved sequences in the gallery set. The item annotated with green box is correctly re-identified, and the red box denotes the wrong results.

We observe that, in the top-left, bottom-left and bottom-right cases, although there exists misalignment, movement and occlusion in the query respectively, our scheme is still able to match the person-of-interest accurately. While the baseline model misses the sequences of the same identity, especially in bottom-left case. Meanwhile, in the top-left, top-right and bottom-right cases, the baseline model re-identities the query incorrectly due to ignoring the fine-grained information between visually similar identities. For example, in the top-right case, the baseline model returns wrong results probably owing to the low light condition. In contrast, DenseIL captures the contour and the fine-grained characters on her back, yielding a satisfactory re-ID result.

Methods	Proc.	Backbone	MARS				DukeV				iLIDS-VID	
			mAP	R-1	R-5	R-20	mAP	R-1	R-5	R-10	R-1	R-5
CNN+XQDA [43]	ECCV16	CaffeNet	47.6	65.3	82.0	89.0	-	-	-	-	53.0	81.4
AMOC [23]	TCSVT17	AMOC	52.9	68.3	81.4	90.6	-	-	-	-	68.7	94.3
SeeForest [45]	CVPR17	CaffeNet	50.7	70.6	90.0	97.6	-	-	-	-	55.2	86.5
MSCAN [18]	CVPR17	MSCAN	56.1	71.8	86.6	93.1	-	-	-	-	-	-
QAN [18]	CVPR17	QAN	-	-	-	-	-	-	-	-	68.0	86.8
ASTPN [38]	ICCV17	ASTPN	-	44.0	70.0	81.0	-	-	-	-	62.0	86.0
MGCAM [29]	CVPR18	MSCAN	71.2	77.2	-	-	-	-	-	-	-	-
Snippet [3]	CVPR18	Res50	76.1	86.3	94.7	98.2	-	-	-	-	85.4	96.7
DuATM [28]	CVPR18	Dense121	67.7	81.2	92.5	-	64.6	81.8	90.2	-	-	-
STAN [21]	CVPR18	Res50	65.8	82.3	-	-	-	-	-	-	80.2	-
ETAP-Net [36]	CVPR18	Res50	67.4	80.8	92.1	96.1	78.3	83.6	94.6	97.6	-	-
STA [8]	AAAI19	Res50	80.8	86.3	95.7	-	94.9	96.2	99.3	99.6	-	-
M3D [20]	AAAI19	Res50-3D	74.1	84.4	93.8	97.7	-	-	-	-	74.1	94.3
ADFD [42]	CVPR19	Res50	78.2	87.0	95.4	98.7	-	-	-	-	86.3	97.4
VRSTC [15]	CVPR19	Res50	82.3	88.5	96.5	-	93.5	95.0	99.1	99.4	83.4	95.5
GLTR [19]	ICCV19	Res50	78.5	87.0	95.8	98.2	93.7	96.3	99.3	-	86.0	**98.0**
COSAM [31]	ICCV19	SE-Res50	79.9	84.9	95.5	97.9	94.1	95.4	99.3	-	79.6	95.3
STE-NVAN [22]	BMVC19	Res50-NL	81.2	88.9	-	-	93.5	95.2	-	-	-	-
MG-RAFA [41]	CVPR20	Res50	85.9	88.8	97.0	98.5	-	-	-	-	88.6	**98.0**
MGH [39]	CVPR20	Res50-NL	85.8	90.0	96.7	98.5	-	-	-	-	85.6	97.1
STGCN [40]	CVPR20	Res50	83.7	90.0	96.4	98.3	95.7	97.3	99.3	-	-	-
TCLNet [14]	ECCV20	Res50-TCL	85.1	89.8	-	-	96.2	96.9	-	-	86.6	-
AP3D [9]	ECCV20	AP3D	85.1	90.1	-	-	95.6	96.3	-	-	86.7	-
AFA [4]	ECCV20	Res50	82.9	90.2	96.6	-	95.4	97.2	99.4	99.7	88.5	96.8
Ours	-	Res50	**87.0**	**90.8**	**97.1**	**98.8**	**97.1**	**97.6**	**99.7**	**99.9**	**92.0**	**98.0**

Table 2: Comparison with state-of-the-art results. NL means the backbone is integrated with Non-Local block [35].

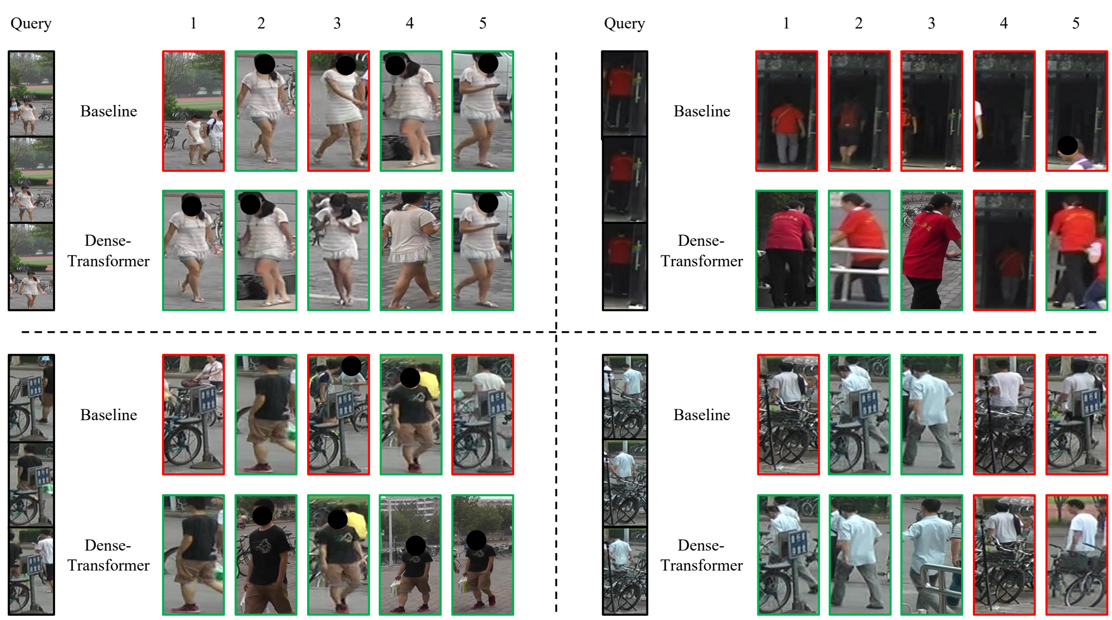


Figure 3: Visualization of the re-identification results of both baseline and our scheme. The left column of each part is sampled frames of query sequence and the right five columns are the sampled frame of top-5 retrieved sequences in the gallery set, where the item annotated with green box is correctly re-identified, and the red box denotes the wrong results.

References

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*, 2015. 3

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, 2020. 3

[3] Dapeng Chen, Hongsheng Li, Tong Xiao, Shuai Yi, and Xiaogang Wang. Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1169–1178, 2018. 4

[4] 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 *European Conference on Computer Vision*, 2020. 2, 4

[5] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10578–10587, 2020. 3

[6] 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 Conference on Computer Vision and Pattern Recognition*, pages 4091–4099, 2015. 1

[7] 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. 1

[8] Yang Fu, Xiaoyang Wang, Yunchao Wei, and Thomas Huang. Sta: Spatial-temporal attention for large-scale video-based person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8287–8294, 2019. 4

[9] Xinqian Gu, Hong Chang, Bingpeng Ma, Hongkai Zhang, and Xilin Chen. Appearance-preserving 3d convolution for video-based person re-identification. In *European Conference on Computer Vision*. Springer, 2020. 3, 4

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[11] Tianyu He, Xu Shen, Jianqiang Huang, Zhibo Chen, and Xian-Sheng Hua. Partial person re-identification with part-part correspondence learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9105–9115, 2021. 3

[12] Tianyu He, Xu Tan, Yingce Xia, Di He, Tao Qin, Zhibo Chen, and Tie-Yan Liu. Layer-wise coordination between encoder and decoder for neural machine translation. In *Advances in Neural Information Processing Systems*, pages 7944–7954, 2018. 3

[13] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 3

[14] Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Temporal complementary learning for video person re-identification. 2020. 3, 4

[15] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. Vrstc: Occlusion-free video person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7183–7192, 2019. 4

[16] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 2

[17] 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. 3

[18] Dangwei Li, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 384–393, 2017. 4

[19] Jianing Li, Jingdong Wang, Qi Tian, Wen Gao, and Shiliang Zhang. Global-local temporal representations for video person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3958–3967, 2019. 4

[20] Jianing Li, Shiliang Zhang, and Tiejun Huang. Multi-scale 3d convolution network for video based person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8618–8625, 2019. 4

[21] Shuang Li, Slawomir Bak, Peter Carr, and Xiaogang Wang. Diversity regularized spatiotemporal attention for video-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 369–378, 2018. 2, 4

[22] Chih-Ting Liu, Chih-Wei Wu, Yu-Chiang Frank Wang, and Shao-Yi Chien. Spatially and temporally efficient non-local attention network for video-based person re-identification. In *BMVC*, 2019. 2, 4

[23] Hao Liu, Zequn Jie, Karlekar Jayashree, Meibin Qi, Jianguo Jiang, Shuicheng Yan, and Jiashi Feng. Video-based person re-identification with accumulative motion context. *IEEE transactions on circuits and systems for video technology*, 28(10):2788–2802, 2017. 4

[24] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23, 2019. 3

[25] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 3

[26] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency,

and Ehsan Hoque. Integrating multimodal information in large pretrained transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2359–2369, 2020. 3

[27] 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. 1

[28] Jianlou Si, Honggang Zhang, Chun-Guang Li, Jason Kuen, Xiangfei Kong, Alex C Kot, and Gang Wang. Dual attention matching network for context-aware feature sequence based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5363–5372, 2018. 4

[29] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Mask-guided contrastive attention model for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1179–1188, 2018. 4

[30] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*, 2019. 3

[31] Arulkumar Subramaniam, Athira Nambiar, and Anurag Mittal. Co-segmentation inspired attention networks for video-based person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 562–572, 2019. 3, 4

[32] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7464–7473, 2019. 3

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 3

[34] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by video ranking. In *European conference on computer vision*, pages 688–703. Springer, 2014. 1

[35] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 4

[36] 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 Conference on Computer Vision and Pattern Recognition*, pages 5177–5186, 2018. 1, 4

[37] Yingce Xia, Tianyu He, Xu Tan, Fei Tian, Di He, and Tao Qin. Tied transformers: Neural machine translation with shared encoder and decoder. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5466–5473, 2019. 3

[38] Shuangjie Xu, Yu Cheng, Kang Gu, Yang Yang, Shiyu Chang, and Pan Zhou. Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In *Proceedings of the IEEE international conference on computer vision*, pages 4733–4742, 2017. 4

[39] Yichao Yan, Jie Qin, Jiaxin Chen, Li Liu, Fan Zhu, Ying Tai, and Ling Shao. Learning multi-granular hypergraphs for video-based person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2899–2908, 2020. 4

[40] 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. 2, 4

[41] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Multi-granularity reference-aided attentive feature aggregation for video-based person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10407–10416, 2020. 2, 4

[42] Yiru Zhao, Xu Shen, Zhongming Jin, Hongtao Lu, and Xiansheng Hua. Attribute-driven feature disentangling and temporal aggregation for video person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4913–4922, 2019. 4

[43] 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 *European Conference on Computer Vision*, pages 868–884. Springer, 2016. 1, 4

[44] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *AAAI*, pages 13001–13008, 2020. 2

[45] Zhen Zhou, Yan Huang, Wei Wang, Liang Wang, and Tieniu Tan. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4747–4756, 2017. 4

[46] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 3