DeepChange: A Long-Term Person Re-Identification Benchmark with Clothes Change

Peng Xu
Tsinghua University
peng_xu@tsinghua.edu.cn

Xiatian Zhu
University of Surrey
xiatian.zhu@surrey.ac.uk

Abstract

Long-term re-id with clothes change is a challenging problem in surveillance AI. Currently, its major bottleneck is that this field is still missing a large realistic benchmark. In this work, we contribute a large, realistic long-term person re-identification benchmark, termed `DeepChange`*. Its unique characteristics are:* **(1)** *Realistic and rich personal appearance (*e.g.*, clothes and hair style) and variations: Highly diverse clothes change and styles, with varying reappearing gaps in time from minutes to seasons, different weather conditions (*e.g.*, sunny, cloudy, windy, rainy, snowy, extremely cold) and events (*e.g.*, working, leisure, daily activities).* **(2)** *Rich camera setups: Raw videos were recorded by* 17 *outdoor varying-resolution cameras operating in a real-world surveillance system.* **(3)** *The currently largest number of (*17*) cameras, (*1,121*) identities, and (*178,407*) bounding boxes, over the longest time span (*12 *months). We benchmark the representative supervised and unsupervised re-id methods on our dataset. In addition, we investigate multimodal fusion strategies for tackling the clothes change challenge. Extensive experiments show that our fusion models outperform a wide variety of state-of-the-art models on* `DeepChange`*. Our dataset and documents are available at* `https://github.com/PengBoXiangShang/deepchange`.

1. Introduction

Person re-identification (re-id) aims to match the person identity classes of bounding box images extracted from non-overlapping camera views [15, 62, 55]. Extensive re-id models were developed in the past decade [16, 38, 63, 60, 59, 53, 64, 19, 57, 28, 3, 4, 33, 43, 29, 27, 2, 30, 58, 11, 65, 56, 6, 67, 66], thanks to the availability of reasonably sized benchmarks [52, 61, 32, 49]. The majority of these methods consider *the short-term search scenario with a strong assumption that the appearance (*e.g.*, clothes, hair style) of each person is stationary*. We name this conventional setting as *short-term re-id* in the follows. Unfortunately, this assumption would be easily broken once the search time span is enlarged to long periods (such as days, weeks, or even months) as an average person often changes the outfit during different day time and across different weathers, daily activities and social events. As shown in Figure 1, a specific person was dressed the same only in a short time (*e.g.*, minutes or hours) but with different clothes/hairs and associations over a long time and across seasons and weathers. Relying heavily on the clothes appearance, the previous short-term re-id models are unsuitable and ineffective in dealing with unconstrained clothes changes over space and time.

Recently there have been a few studies for tackling the long-term re-id situations focusing on clothes change [54, 23, 39, 34, 48, 50, 51]. Since there are no appropriate datasets publicly available, to enable research these works introduced several small-scale long-term re-id datasets by using web celebrity images [24], synthesizing pseudo person identities [48, 34], collecting person images under simulated surveillance settings [54, 39, 48, 50]. While these dataset construction efforts are useful in initiating the research, it is obvious that a real, large-scale long-term re-id benchmark is missing and highly demanded. The impact and significance of such a benchmark in driving the research progress has been repeatedly demonstrated in the short-term re-id case [61, 32, 52, 51]. However, it is much more challenging to establish a large re-id dataset with clothes change as compared to the short-term counterparts. This is due to that: (1) More video data need to be collected and processed over long time periods; (2) Labeling person identity becomes much more difficult when a person is dressed up with different outfit from time to time. Regardless, we believe that it is worthwhile to overcome all these challenges.

To facilitate the research towards the applications of long-term person search in reality, we contribute the first large-scale person re-id dataset with native/natural appearance (mainly clothes) changes, termed as `DeepChange`. Different from existing datasets, `DeepChange` has several unique characteristics: (1) The raw videos are collected in

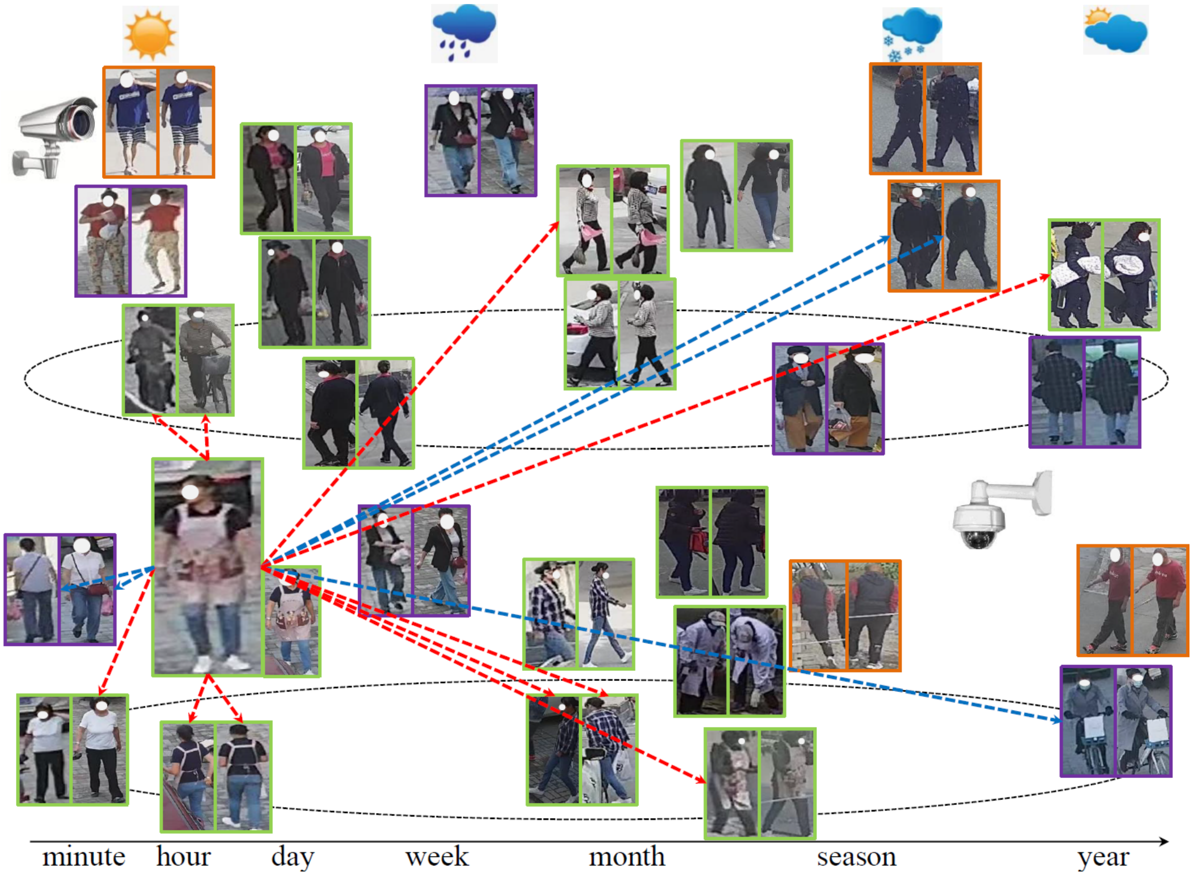


Figure 1. **Motivation**: An average person often changes the clothes over time and space, as shown by three random persons (color coded). Conventional re-id settings usually assume stationary clothes per person and are hence valid only for the short-term application scenarios. For long-term re-id cases, we must consider the unconstrained *clothes change* challenge. Note, for clarity only part of true (red arrow) and false (blue arrow) matches are plotted. Best viewed in color.

a real-world surveillance camera network deployed at a residential block where rich scenes and realistic background are presented over time and space. (2) The videos cover a period of 12 months with a wide variety of different weathers and contexts. To the best of our knowledge, this is the longest time duration among all re-id datasets, presenting natural/native personal appearance (*e.g*., clothes and hair style) variations with people from all walks of life. (3) Compared to existing long-term re-id datasets, it contains the largest number of (17) cameras, (1121) identities, and (178K) bounding boxes. Overall, `DeepChange` is the only realistic, largest long-term person re-id dataset, created using the real-world surveillance videos.

We make the following **contributions**:

(1) A large scale long-term person re-id dataset, called `DeepChange`, is introduced. Compared with existing alternative datasets, this dataset offers more realistic and more challenging person re-id tasks over long time with native appearance changes. With a much larger quantity of person images for model training, validation, and testing, `DeepChange` provides a more reliable and indicative test bed for future research.

(2) We conduct extensive experiments on the `DeepChange` dataset, including seminal CNNs [18, 22, 41, 45], Transformers [8, 47], and state-of-the-art long-term re-id models [24].

(3) To tackle the clothes change challenge, we investigate multimodal fusion strategies (*e.g*., gray images, edge maps [68], key points [1]) and achieve new state-of-the-art results on `DeepChange`.

2. DeepChange Benchmark

Venue and climate Our raw videos were collected from a real-world surveillance system for a wide (14 hectares) and dense block, with the middle temperate continental monsoon climate. This venue has various scenes, including crowded streets, shops, restaurants, construction sites, residential buildings, physical exercise areas, car parking, sparsely populated corner, *etc*. Thus it has two major advantages: (i) Identity Diversity: Persons cross a wide range

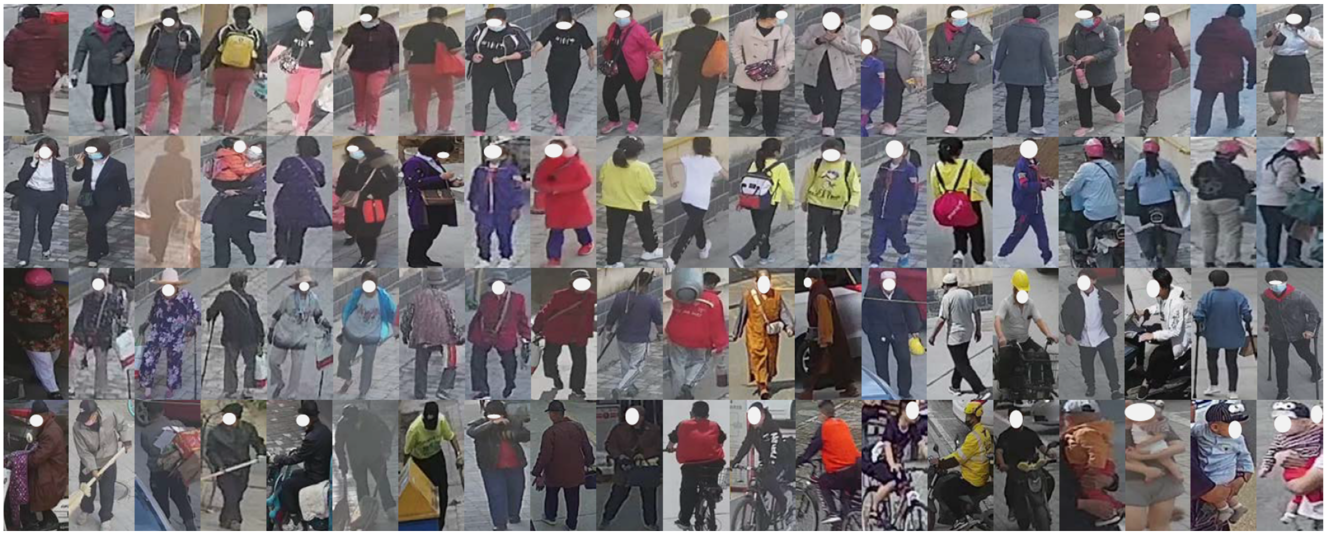


Figure 2. Image samples of random identities in DeepChange. Identities from top left to bottom right: an aunt (bbox#1-#19), an office lady (bbox#20-#27), a pupil (bbox#28-#36), a newspaper delivery (bbox#37-#41), an older aunt (bbox#42-#49), a worker (bbox#50-#51), a nun (bbox#52-#53), a Muslim man (bbox#54-#56), a chef (bbox#57-#58), a disabled person (bbox#59-#60), a dustman (bbox#61-#70), a middle school student (bbox#71-#74), a food delivery (bbox#75-#76), a baby (bbox#77-#80). Best viewed in color.

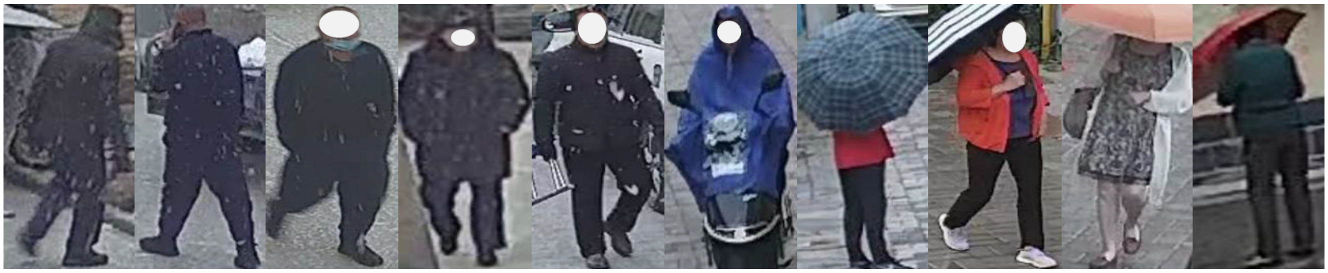


Figure 3. Samples collected in snow (bbox#1-#5) and rain (bbox#6-#10) weather from DeepChange. Best viewed in color.

of ages and professions, *e.g.*, lactating baby, very older person, office lady, elementary student, high school student, worker, deliveryman, religious. Some identity examples are illustrated in Figure 2. (ii) Weather Diversity: During our collecting, we have observed a temperature variation ranging from $-30°C$ (in winter) to $35°C$ (in summer). Therefore, we have collected persons appearing in various weathers, *e.g.*, sunny, cloudy, windy, rainy, snowy, extremely cold. Some image samples with snow are shown in Figure 3, where the snow can cause noticeable appearance changes on both clothes or background. Altogether, the identity and weather diversities will be embodied in drastic appearance changes, enabling realistic long-term re-id benchmarking with our DeepChange dataset. Figure 2 demonstrates some bounding boxes of an identity randomly selected from our dataset, where we can observe dramatic appearance changes across weathers, seasons, *etc*.

Security camera Our raw videos were recorded by 17 security cameras with a speed of 25 FPS (frames per second) and different resolutions (1920×1080 spear camera $\times 14$, 1280×960 spherical camera $\times 3$). These 17 cameras are part of a large-scale video surveillance system. In particular, these cameras are mounted on the exterior walls of buildings or on lamp posts, and their height is approximately 3 to 6 meters. These cameras are monitoring various views including crowded streets, construction sites, physical exercise areas, car parking, sparsely populated corner, *etc*. Therefore, these cameras provide diverse scenes, identities, behaviors, events, *etc*.

Video collecting Our collecting is a very long course over 12 months across two different calendar years. Every month, we randomly collected raw videos on 7 to 10 randomly selected days to cover as much weather as possible. On each selected day, we collected video from dawn to night to record comprehensive light variations. A huge amount of videos can provide highly diverse clothes changes and dressing styles, with the reappearing gap in time ranging from minutes, hours, and days to weeks, months, seasons, and years. The volunteers agreed to be collected and didn't mind being studied and released for

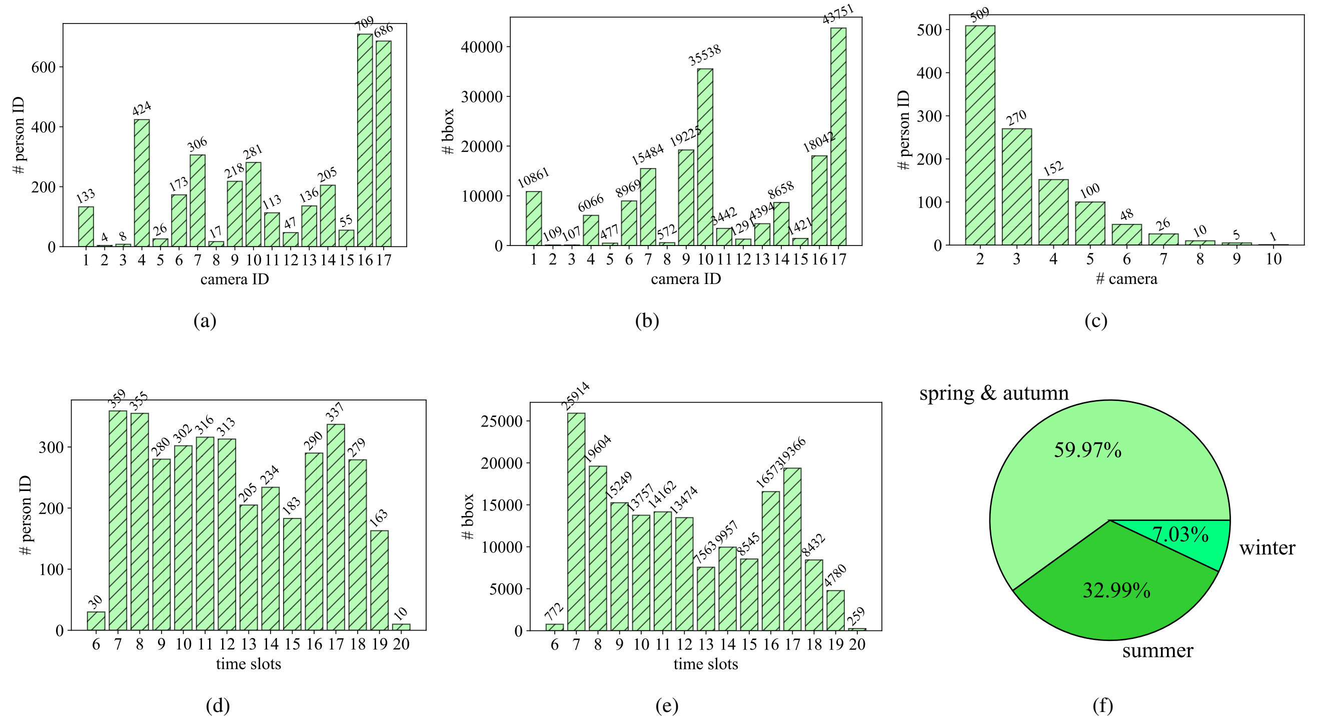


Figure 4. Statistical analysis of the DeepChange dataset. (a): Identity size captured by #1-#17 cameras, (b): Bounding box amounts across cameras, (c): Distribution of identity sizes captured by different camera numbers, (d): Identity size in different time slots, (e): Bounding box size in different time slots, (f): Bounding box ratios across seasons (clothes styles).

Table 1. Dataset splitting of DeepChange.

	Train set	Validation set	Test set
# Person	450	150	521
# Box	75,083	Probe: 4,976 Gallery: 17,865	Probe: 17,527 Gallery: 62,956

academic research, and some of them helped us with the labeling process. We promised to blur their faces for personal privacy and request the academic users to sign an agreement. The permission to use these videos was granted by the owner/authority for research purposes only.

Pedestrian detection After labeling, we used Faster RCNN [40] to detect bounding boxes. For each selected bounding box, we annotated person ID, camera ID, tracklet ID, and time stamp. Finally, we detected and blurred the face area for privacy protection.

Data statistics All person identities were captured by at least two cameras with most seen by 2 $\sim$ 6 cameras (as shown in Figure 4(c)). Figure 4(e) indicates that the labeled bounding boxes are distributed from 6 am to 9 pm. As illustrated in Figure 4(f), the bounding box ratios of persons wearing spring&autumn, summer, and winter clothes are 59.97%, 32.99%, and 7.03%, respectively. More detailed statistics can be found in Figure 4.

Data splitting We shuffled all our collected identities, and then orderly picked 450, 150, and 521 IDs for training, validation, and test, respectively. In validation and test sets, given a tracklet, we randomly chose $\sim$ 5 bounding boxes as queries/probes, and the remaining boxes were split into the gallery. Details were summarized in Table 1.

Diversity and challenge As aforementioned, this wide (14 hectares) and dense block provides various identities (as shown in Figure 2), and middle temperate continental monsoon climate causes diverse clothes changes (as demonstrated in Figure 2). Our long-term video collection makes full use of these characteristics of this venue and the climate. We observed that obvious hair-style changes often happened in long-term surveillance videos. In Figure 5, we present some random cases with simultaneous clothes and hair style changes. It is interesting to see that hair style changes should also be considered in long-term re-id, as this might lead to non-neglectable appearance alternation. [1]

2.1. Comparison with Existing Datasets

We compare our DeepChange with existing re-id datasets with and without clothes change. We only discuss the publicly available and non-synthetic datasets.

[1] More illustrations are provided in the appendix.

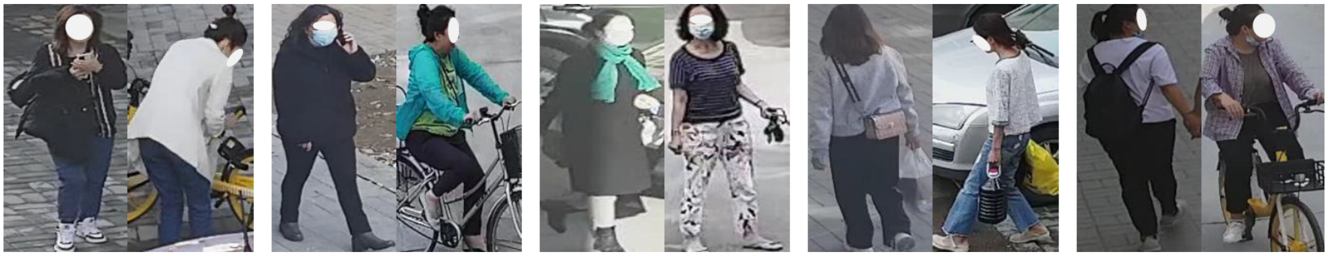


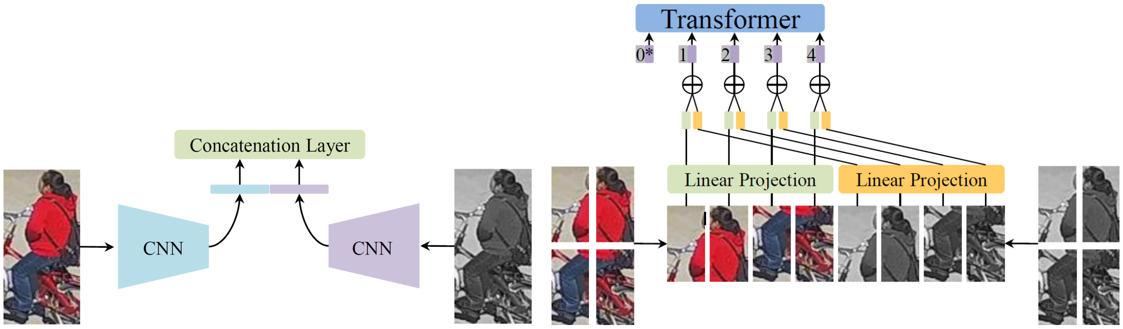
Figure 5. Sample pairs of the specific persons with simultaneous clothes and hair style changes in the DeepChange dataset. For each identity, only two cases are selected randomly. Best viewed in color.

Table 2. Comparison with existing long-term image-based re-id datasets with clothes change. 'Fas.': Faster RCNN [40], 'Mas.': Mask RCNN [17], 'ind.': indoor, 'out.': outdoor, 'CD': **cross-day**, 'CM': **cross-month**, 'CS': **cross-season**, 'CY': **cross-year**, 'spr.': **spring**, 'sum.': **summer**, 'aut.': **autumn**, 'win.': **winter**, 'sim.': simulated, 'sur.': surveillance, '-': unknown.

Dataset	# Person	# Box	# Cam	Source	Detector	Scene	Time Range					Cloth Style			
							course	CD	CM	CS	CY	spr.	sum.	aut.	win.
Real28 [48]	28	4.3K	4	sim.	Mas.	out., ind.	3 days	✓					✓		
NKUP [50]	107	9.7K	15	sim.	-	out., ind.	4 months	✓	✓	✓				✓	✓
LTCC [39]	152	17K	12	sim.	Mas.	ind.	2 months	✓	✓				✓		
PRCC [54]	221	33K	3	sim.	-	ind.	-	✓					✓		
DeepChange	**1,121**	**178K**	**17**	**sur.**	Fas.	out.	**12 months**	✓	✓	✓	✓	✓	✓	✓	✓

Table 3. Comparison with conventional short-term image-based Re-ID datasets without clothes change. ('Fas.': Faster RCNN [40], 'DPM': Deformable Part Model [10], 'ind.': indoor, 'out.': outdoor)

Dataset	DeepChange	MSMT17 [52]	Duke [64]	Market [61]	CUHK03 [32]	CUHK01 [31]	VIPeR [16]	PRID [20]	CAVIAR [5]
# Person	1,121	4,101	1,812	1,501	1,467	971	632	934	72
# Bbox	**178K**	126K	36K	32K	28K	3.8K	1.2K	1.1K	0.6K
# Camera	**17**	15	8	6	2	10	2	2	2
Detector	Fas.	Fas.	hand	DPM	DPM, hand	hand	hand	hand	hand
Scene	out.	out. & ind.	out.	out.	ind.	ind.	out.	out.	ind.



(a)

(b)

Figure 6. Two multimodal fusion strategies: (a) CNN-based late-fusion, (b) Transformer-based early-fusion. Both support more than two modalities.

As summarized in Table 2, compared with existing long-term datasets, DeepChange has the **largest** number of identities, cameras, and bounding boxes, with the **longest** time span. Besides, it is the only dataset offering four seasonal dressing styles.

As seen in Table 3, DeepChange still has the **largest** number of cameras and bounding boxes even in comparison to traditional short-term re-id datasets.

3. Multimodal Fusion for Tackling the Clothes Change Challenge

Conventional re-id methods mostly use the color information [15, 52, 62]. At presence of clothes change, this may be insufficient given that different clothes could vary arbitrarily in color. To tackle this challenge, we consider a simultaneous use of multiple modalities so that additional information irrelevant to the clothes' appearance can be leveraged concurrently. In particular, we consider three more modalities, grayscale images, edge maps, and skeleton key points, for providing the potentially useful knowledge about body contour and part proportion. For cost-effective multimodal fusion, we explore two strategies: (i) **CNN-based late-fusion** (Figure 6(a)): Multiple input modalities are separately encoded by a multi-branch network, followed by feature concatenation. (ii) **Transformer-based early-fusion** (Figure 6(b)): The patch tokens from multiple modalities are fused by weighted sum in a position-wise manner, followed by self-attention representation learning. These multimodal fusion strategies can be integrated into the feature encoding parts of other re-id models, to provide robust representation. Compared with other multimodal re-id methods, our baselines are more straightforward with lower system complexity.

4. Experiments

We evaluated common deep CNN models and state-of-the-art re-id methods on `DeepChange`.

Protocols and metrics In the traditional short-term person re-id, it is assumed that the appearance of a specific person does not change across time and space. However, this assumption often does not hold for the long-term setting. Hence, in our benchmark we allow the true matches coming from the same camera as the probe/query image but from a different time and trajectory. Following [61, 42, 52, 12, 9, 13], we used both Cumulated Matching Characteristics (CMC) and mean average precision (mAP) as accuracy metrics.

Implementation details For a fair comparison, all experiments were implemented in the same software and hardware platform. In training only random horizontal flipping was used for data augmentation. For reproducibility, we conducted a unified early stop strategy for all experiments. We empirically adopted mAP on evaluation set as early stop metric and set patience as 30 epochs, thus the checkpoints with the highest validation performance were chosen to report test performance. As all the backbones were initialized by the weights pretrained on ImageNet [7], we empirically used an initial learning rate of $1e^{-4}$, multiplied by a decay factor 0.1 every 20 epochs. ReIDCaps [24] and BNNeck re-id [36] use the customized loss functions, while the other models were trained by minimizing the softmax based cross-entropy loss. During evaluating and testing, we extracted the features of bounding boxes from the penultimate layer to conduct re-id matching. Edge Box [68] and Open Pose [1] toolboxes were used to extract edge maps and detect body keypoints (15 keypoints, $54D$ vector), respectively. For the multimodal methods, body keypoint vectors were passed into a module of two fully-connected layers with Batch Normalization [26] and ReLU [14] activations, while CNNs were used to encode RGB images, grayscale images, and edge maps.

Result analysis We report the results of unimodal and multimodal methods in Table 4 and Table 5, respectively. We consider five groups of methods: (1) #1 to #31: Common CNNs used in re-id; (2) #32 to #34: Latest CNNs; (3) #35 to #47: Several state-of-the-art short-term re-id methods; (4) #48 to #51: A state-of-the-art long-term re-id method ReIDCaps [24]; (5) #52 to #54: Vision Transformer (ViT) [8] and its variant DeiT [47]. We draw the following main observations:

(i) For the results with RGB input, deeper models are usually superior than shallower ones as expected. For example, we see a clear upward trend from #1 ResNet18 (rank@1: 34.45%, mAP: 08.44%) to #13 ResNet152 (rank@1: 39.84%, mAP: 11.49%). It is observed that ResNets and DenseNets outperform MobileNetv2 and Inceptionv3 by a clear margin. Lightweight networks tend to underperform, *e.g.*, ShuffleNetv2. In general, these observations are largely consistent with conventional short-term re-id results.

(ii) As re-id specific models, BNNeck re-id and ReIDCaps achieve good performances. Similarly, both benefit from deeper networks, as seen from #40 to #46, #49 to #48.

(iii) ViT obtains the best mAP score among all the unimodal methods (#52), indicating that self-attention representation is superior in dealing with clothes change challenge over CNN models.

(iv) With different modalities (#5 RGB, #6 grayscale, and #7 edge map) for ResNet50, it is observed that RGB is the best. Plausible reasons for lower performance with grayimages include: Although being more tolerant with clothes change, they offer less information than colorful ones. Besides, CNNs are pretrained on colorful ImageNet images. Note, we did not test keypoint modality in isolation, as it fails on 13.45% ($23K$ out of $178K$) of person images.

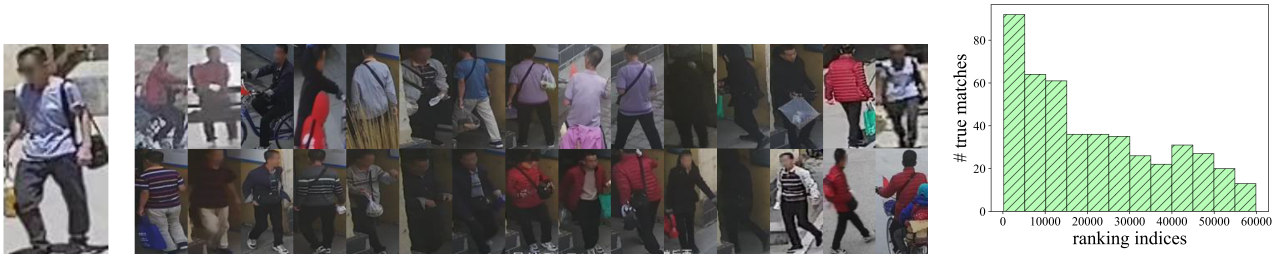
(v) From the multimodal results, we see that both edge maps and grayscale images are useful for performance by feature late-fusion (#5 vs. #56, #57, and #58). This validates our hypothesis that jointly using multiple modalities with specific appearance indeed provides extra information for tackling the clothes change challenge. It is expected that higher-quality edge maps would contribute more to the performance. Again, we find that using stronger networks (*e.g.*, #59, #60, #61) can further improve the results in the multimodal case. On the other hand, multimodal fusion also helps state-of-the-art re-id models (*e.g.*, #42 vs. #62). This indicates good complementary effect between multimodal

Table 4. Unimodal results on the test set of DeepChange. Both rank accuracy (%) and mAP (%) are reported. (‘R’: RGB, ‘G’: Grayscale, ‘E’: Edge map, ‘K’: Body key point, ‘Mod.’: Modalities)

Model	Input		Batch size	Rank				mAP
	Mod.	Resolution		@1	@5	@10	@20	
#1 ResNet18 [18]	R	256 × 128	256	34.45	46.01	51.72	58.26	08.44
#2 ResNet18 [18]	G	256 × 128	256	26.61	39.02	45.45	53.06	05.49
#3 ResNet34 [18]	R	256 × 128	256	35.21	47.37	53.61	60.03	09.49
#4 ResNet34 [18]	G	256 × 128	256	28.60	41.53	47.98	54.87	06.39
#5 ResNet50 [18]	R	256 × 128	192	36.62	49.88	55.46	61.92	09.62
#6 ResNet50 [18]	G	256 × 128	192	30.04	43.12	49.82	57.04	06.96
#7 ResNet50 [18]	E	256 × 128	192	16.05	28.51	35.59	43.28	03.17
#8 ResNext50_32x4d [18]	R	256 × 128	128	36.94	48.87	54.55	60.50	10.09
#9 Wide_ResNet50_2 [18]	R	256 × 128	128	33.68	45.57	51.45	57.72	08.71
#10 ResNet101 [18]	R	256 × 128	128	39.31	51.65	57.36	63.72	11.00
#11 ResNext101_32x8d [18]	R	256 × 128	64	40.93	52.68	58.05	64.21	11.07
#12 Wide_ResNet101_2 [18]	R	256 × 128	64	34.33	46.46	52.42	58.69	09.03
#13 ResNet152 [18]	R	256 × 128	96	39.84	52.51	58.35	64.75	11.49
#14 MobileNetv2 [41]	R	256 × 128	256	33.71	46.51	52.72	59.45	07.95
#15 MobileNetv3 Small [21]	R	256 × 128	768	28.73	40.87	47.24	54.12	06.09
#16 MobileNetv3 Large [21]	R	256 × 128	320	33.08	44.99	51.32	58.14	07.56
#17 GoogLeNet [44]	R	299 × 299	96	29.02	40.83	46.75	53.65	07.24
#18 Inceptionv3 [45]	R	299 × 299	96	35.02	47.71	53.91	60.64	08.85
#19 DenseNet121 [22]	R	256 × 128	128	38.26	50.27	55.91	62.40	09.12
#20 DenseNet121 [22]	G	256 × 128	128	30.64	43.02	49.41	56.44	06.14
#21 DenseNet121 [22]	E	256 × 128	128	18.21	30.40	36.86	44.04	02.88
#22 DenseNet121 [22]	R	224 × 224	64	39.92	51.76	57.21	62.98	09.99
#23 DenseNet161 [22]	R	256 × 128	64	45.92	56.72	61.79	67.41	12.30
#24 DenseNet169 [22]	R	256 × 128	96	43.40	54.80	60.11	65.90	11.25
#25 DenseNet201 [22]	R	256 × 128	64	44.98	56.13	61.32	66.98	11.71
#26 ShuffleNetv2_x0.5 [37]	R	256 × 128	1024	17.21	27.69	33.48	40.53	04.00
#27 ShuffleNetv2_x1.0 [37]	R	256 × 128	768	20.22	32.95	39.58	46.89	04.60
#28 SqueezeNet1_0 [25]	R	256 × 128	384	23.38	33.17	38.64	44.34	04.78
#29 SqueezeNet1_1 [25]	R	256 × 128	640	23.52	33.54	39.26	45.59	05.02
#30 MnasNet0_5 [46]	R	256 × 128	512	17.52	29.17	36.20	43.51	02.79
#31 MnasNet1_0 [46]	R	256 × 128	256	30.63	44.02	50.40	57.56	06.25
#32 SCNet50 [35]	R	256 × 128	160	35.07	47.73	54.32	60.73	09.53
#33 SCNet50_v1d [35]	R	256 × 128	128	37.46	50.20	56.13	62.36	09.42
#34 SCNet101 [35]	R	256 × 128	96	37.25	50.09	56.28	63.16	10.35
#35 OSNet ibn x1.0 [66]	R	256 × 128	96	42.75	54.91	60.82	66.80	10.97
#36 OSNet x1.0 [66]	R	256 × 128	96	39.65	52.22	58.32	64.23	10.34
#37 OSNet x0.75 [66]	R	256 × 128	160	39.96	51.89	57.69	63.85	09.92
#38 OSNet x0.5 [66]	R	256 × 128	256	38.09	51.08	56.79	63.27	09.59
#39 OSNet x0.25 [66]	R	256 × 128	512	34.94	47.70	54.10	60.77	08.62
#40 BNNeck re-id ResNet18 [36]	R	256 × 128	224	38.17	51.93	58.08	64.70	09.51
#41 BNNeck re-id ResNet34 [36]	R	256 × 128	128	40.06	53.49	59.55	66.25	10.52
#42 BNNeck re-id ResNet50 [36]	R	256 × 128	56	47.45	59.47	65.19	71.10	12.98
#43 BNNeck re-id ResNet50 [36]	G	256 × 128	56	40.02	54.04	60.19	67.09	09.43
#44 BNNeck re-id ResNet50 [36]	E	256 × 128	56	21.75	35.99	43.11	50.81	03.67
#45 BNNeck re-id ResNet101 [36]	R	256 × 128	40	48.10	60.70	66.10	72.06	13.72
#46 BNNeck re-id ResNet152 [36]	R	256 × 128	28	50.29	62.27	67.85	73.63	14.59
#47 BNNeck re-id DenseNet121 [36]	R	256 × 128	40	47.86	60.47	65.88	71.64	13.41
#48 ReIDCaps [24] (DenseNet121)	R	224 × 224	24	44.29	56.44	62.01	68.01	13.25
#49 ReIDCaps [24] (ResNet50)	R	224 × 224	32	39.49	52.28	58.68	64.99	11.33
#50 ReIDCaps [24] (no auxiliary)	R	224 × 224	24	35.41	46.66	52.09	58.13	09.25
#51 ReIDCaps [24] (no capsule)	R	224 × 224	80	39.38	51.86	57.82	64.44	11.16
#52 ViT B16 [8]	R	256 × 128	64	49.78	61.81	67.38	72.92	**14.98**
#53 ViT B16 [8]	G	256 × 128	64	38.52	51.85	58.32	65.12	10.63
#54 DeiT [47]	R	256 × 128	64	44.43	56.25	61.82	67.46	13.72

Table 5. Multimodal results on the test set of DeepChange. Both rank accuracy (%) and mAP (%) are reported. ('R': RGB, 'G': Grayscale, 'E': Edge map, 'K': Body key point, '2br': Two branches, '3br': Three branches, 'Mod.': Modalities, 'Dim.': Dimensions)

Model	Input		Batch size	Rank				mAP
	Mod.	Dim.		@1	@5	@10	@20	
#55 2br ResNet50	R, K	256 × 128	192	36.53	48.87	54.86	61.47	09.54
#56 2br ResNet50	R, E	256 × 128	96	40.26	52.91	59.11	65.47	10.43
#57 2br ResNet50	R, G	256 × 128	96	40.52	53.65	59.61	65.60	10.22
#58 3br ResNet50	R, G, E	256 × 128	64	41.67	54.28	60.04	66.37	11.03
#59 2br DenseNet121	R, E	256 × 128	64	44.55	56.40	62.03	67.85	11.21
#60 2br DenseNet121	R, G	256 × 128	64	44.80	56.79	62.48	68.06	11.36
#61 3br DenseNet121	R, G, E	256 × 128	32	45.36	57.36	62.91	69.29	11.73
#62 2br BNNeck re-id ResNet50 [36]	R, G	256 × 128	28	46.62	59.72	65.58	71.48	13.12
#63 ViT B16 [8]	1* R + 0.1 * G	256 × 128	32	47.83	59.29	64.59	70.23	15.13
#64 ViT B16 [8]	1* R + 0.3 * G	256 × 128	32	48.00	59.47	64.65	70.04	**15.19**



(a)

(b)

Figure 7. A qualitative analysis of person re-id with clothes change. (a) *Left*: A random query image; *Right*: 30 randomly selected true matches from the gallery. (b) The rank histogram of all the true matches.

fusion and model refinement. With the proposed fusion method on RGB and grayscale images, interestingly the standard ViT model achieves the best mAP scores, surpassing a wide variety of CNNs and CNN based re-id methods by a clear margin. This indicates that Transformers might be a stronger architecture for long-term person re-id, with promising investigation potentials for the future.

(vi) Unlike short-term person re-id, we see that in long-term setting mAP scores are significantly lower than rank@1 rates, a consistent finding as in [24, 54]. This is the most fundamental challenge with clothes change, as there is little correlation between different clothes a specific person would wear over space and time. It is also not an easy task for human beings. Due to high clothes diversity in our DeepChange dataset, this challenge becomes more severe. For example, given a random query image (Figure 7(a)), the true matches with varying clothes often form a long tail distribution in rank (Figure 7(b)). This contributes to a low mAP score.

5. Conclusions

In this paper, we have introduced the only realistic, large-scale long-term person re-id datasets with natural clothes change, named DeepChange. This aims for facilitating the research towards more realistic person re-id applications over space and time. Compared with existing alternative datasets, it is established uniquely using real-world video sources without any artificial simulation. Constructed with a huge amount of annotation efforts, DeepChange contains the largest number of cameras, identities, and bounding boxes, and covers the longest time period and native appearance change. We conduct extensive experiments using a wide variety of CNN and Transformer models and state-of-the-art re-id methods, offering a set of rich baselines for future research works. Further, we investigate multimodal fusion strategies for overcoming the clothes change challenge, and achieve new state-of-the-art results on our DeepChange dataset.

6. Future work

In the future efforts, we would try to extend this work mainly in following aspects: (i) continually collecting video to provide longer term test bed, *e.g.*, three years, (ii) annotating more person identities to enlarge training/validation/testing subsets and provide more complex appearance changes, (iii) accommodating more cameras to provide more view variety, (iv) creating a video-based long-term re-id dataset.

This work was partly supported by the National Natural Science Foundation of China under Grant 62276153.

References

[1] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *TPAMI*, 2019. 2, 6

[2] Xiaobin Chang, Timothy M Hospedales, and Tao Xiang. Multi-level factorisation net for person re-identification. In *CVPR*, 2018. 1

[3] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. Person re-identification by deep learning multi-scale representations. In *ICCVW*, 2017. 1

[4] Ying-Cong Chen, Xiatian Zhu, Wei-Shi Zheng, and Jian-Huang Lai. Person re-identification by camera correlation aware feature augmentation. *TPAMI*, 2017. 1

[5] Dong Seon Cheng, Marco Cristani, Michele Stoppa, Loris Bazzani, and Vittorio Murino. Custom pictorial structures for re-identification. In *BMVC*, 2011. 5

[6] Zhiyi Cheng, Qi Dong, Shaogang Gong, and Xiatian Zhu. Inter-task association critic for cross-resolution person re-identification. In *CVPR*, 2020. 1

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009. 6

[8] 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. 2, 6, 7, 8

[9] Chanho Eom and Bumsub Ham. Learning disentangled representation for robust person re-identification. In *NeurIPS*, 2019. 6

[10] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *TPAMI*, 2009. 5

[11] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. *arXiv preprint arXiv:2001.01526*, 2020. 1

[12] Yixiao Ge, Zhuowan Li, Haiyu Zhao, Guojun Yin, Shuai Yi, Xiaogang Wang, and Hongsheng Li. Fd-gan: Pose-guided feature distilling gan for robust person re-identification. In *NeurIPS*, 2018. 6

[13] Yixiao Ge, Feng Zhu, Dapeng Chen, Rui Zhao, and Hongsheng Li. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. In *NeurIPS*, 2020. 6

[14] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *AISTATS*, 2011. 6

[15] Shaogang Gong, Marco Cristani, Shuicheng Yan, and Chen Change Loy. *Person re-identification*. Springer, January 2014. 1, 5

[16] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, 2008. 1, 5

[17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 5

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2, 7

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

[20] Martin Hirzer, Csaba Beleznai, Peter M Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *SCIA*, 2011. 5

[21] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *ICCV*, 2019. 7

[22] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, 2017. 2, 7

[23] Yan Huang, Qiang Wu, Jingsong Xu, and Yi Zhong. Celebrities-reid: A benchmark for clothes variation in long-term person re-identification. In *IJCNN*, 2019. 1

[24] Yan Huang, Jingsong Xu, Qiang Wu, Yi Zhong, Peng Zhang, and Zhaoxiang Zhang. Beyond scalar neuron: Adopting vector-neuron capsules for long-term person re-identification. *TCSVT*, 2019. 1, 2, 6, 7, 8

[25] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016. 7

[26] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. 6

[27] Jiening Jiao, Wei-Shi Zheng, Ancong Wu, Xiatian Zhu, and Shaogang Gong. Deep low-resolution person re-identification. In *AAAI*, 2018. 1

[28] Xu Lan, Hanxiao Wang, Shaogang Gong, and Xiatian Zhu. Deep reinforcement learning attention selection for person re-identification. In *BMVC*, 2017. 1

[29] Minxian Li, Xiatian Zhu, and Shaogang Gong. Unsupervised person re-identification by deep learning tracklet association. In *ECCV*, 2018. 1

[30] Minxian Li, Xiatian Zhu, and Shaogang Gong. Unsupervised tracklet person re-identification. *TPAMI*, 2019. 1

[31] Wei Li, Rui Zhao, and Xiaogang Wang. Human reidentification with transferred metric learning. In *ACCV*, 2012. 5

[32] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014. 1, 5

[33] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *CVPR*, 2018. 1

[34] Yu-Jhe Li, Zhengyi Luo, Xinshuo Weng, and Kris M Kitani. Learning shape representations for clothing variations in person re-identification. *arXiv preprint arXiv:2003.07340*, 2020. 1

[35] Jiang-Jiang Liu, Qibin Hou, Ming-Ming Cheng, Changhu Wang, and Jiashi Feng. Improving convolutional networks with self-calibrated convolutions. In *CVPR*, 2020. 7

[36] Hao Luo, Wei Jiang, Youzhi Gu, Fuxu Liu, Xingyu Liao, Shenqi Lai, and Jianyang Gu. A strong baseline and batch normalization neck for deep person re-identification. *TMM*, 2020. 6, 7, 8

[37] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *ECCV*, 2018. 7

[38] Bryan James Prosser, Wei-Shi Zheng, Shaogang Gong, Tao Xiang, Q Mary, et al. Person re-identification by support vector ranking. In *BMVC*, 2010. 1

[39] Xuelin Qian, Wenxuan Wang, Li Zhang, Fangrui Zhu, Yanwei Fu, Tao Xiang, Yu-Gang Jiang, and Xiangyang Xue. Long-term cloth-changing person re-identification. In *ACCV*, 2020. 1, 5

[40] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 4, 5

[41] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, 2018. 2, 7

[42] Arulkumar Subramaniam, Moitreya Chatterjee, and Anurag Mittal. Deep neural networks with inexact matching for person re-identification. In *NeurIPS*, 2016. 6

[43] 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 *ECCV*, 2018. 1

[44] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, 2015. 7

[45] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016. 2, 7

[46] Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V Le. Mnasnet: Platform-aware neural architecture search for mobile. In *CVPR*, 2019. 7

[47] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021. 2, 6, 7

[48] Fangbin Wan, Yang Wu, Xuelin Qian, Yixiong Chen, and Yanwei Fu. When person re-identification meets changing clothes. In *CVPRW*, 2020. 1, 5

[49] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *CVPR*, 2018. 1

[50] Kai Wang, Zhi Ma, Shiyan Chen, Jinni Yang, Keke Zhou, and Tao Li. A benchmark for clothes variation in person re-identification. *IJIS*, 2020. 1, 5

[51] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by video ranking. In *ECCV*, 2014. 1

[52] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, 2018. 1, 5, 6

[53] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *CVPR*, 2016. 1

[54] Qize Yang, Ancong Wu, and Wei-Shi Zheng. Person re-identification by contour sketch under moderate clothing change. *TPAMI*, 2019. 1, 5, 8

[55] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *TPAMI*, 2021. 1

[56] Jiahang Yin, Ancong Wu, and Wei-Shi Zheng. Fine-grained person re-identification. *IJCV*, 2020. 1

[57] Hong-Xing Yu, Ancong Wu, and Wei-Shi Zheng. Cross-view asymmetric metric learning for unsupervised person re-identification. In *ICCV*, 2017. 1

[58] Hong-Xing Yu, Wei-Shi Zheng, Ancong Wu, Xiaowei Guo, Shaogang Gong, and Jian-Huang Lai. Unsupervised person re-identification by soft multilabel learning. In *CVPR*, 2019. 1

[59] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a discriminative null space for person re-identification. In *CVPR*, 2016. 1

[60] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Unsupervised salience learning for person re-identification. In *CVPR*, 2013. 1

[61] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015. 1, 5, 6

[62] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016. 1, 5

[63] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Reidentification by relative distance comparison. *TPAMI*, 2012. 1

[64] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, 2017. 1, 5

[65] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Learning to adapt invariance in memory for person re-identification. *TPAMI*, 2020. 1

[66] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Learning generalisable omni-scale representations for person re-identification. *TPAMI*, 2021. 1, 7

[67] Xiangping Zhu, Xiatian Zhu, Minxian Li, Pietro Morerio, Vittorio Murino, and Shaogang Gong. Intra-camera supervised person re-identification. *IJCV*, 2021. 1

[68] C. Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *ECCV*, 2014. 2, 6