Surveillance Face Recognition Challenge

Zhiyi Cheng · Xiatian Zhu · Shaogang Gong

Abstract Face recognition (FR) is one of the most extensively investigated problems in computer vision. Significant progress in FR has been made due to the recent introduction of the larger scale FR challenges, particularly with *constrained* social media web images, e.g. high-resolution photos of celebrity faces taken by professional photo-journalists. However, the more challenging FR in *unconstrained and low-resolution* surveillance images remains largely under-studied. To facilitate more studies on developing FR models that are effective and robust for low-resolution surveillance facial images, we introduce a new Surveillance Face Recognition Challenge, which we call the *QMUL-SurvFace* benchmark. This new benchmark is the *largest* and more importantly the only *true* surveillance FR benchmark to our best knowledge, where low-resolution images are not synthesised by artificial down-sampling of native high-resolution images[1]. This challenge contains 463,507 face images of 15,573 distinct identities captured in real-world uncooperative surveillance scenes over wide space and time. As a consequence, it presents an extremely challenging FR benchmark due to severely low resolution, motion blur, uncontrolled poses, varying occlusion, poor illumination, and background clutters, all inherent to true surveillance facial images. We benchmark the FR performance on this challenge using five representative deep learning face recognition models (DeepID2, CentreFace, VggFace, FaceNet, and SphereFace), in comparison to existing benchmarks. We show that the current state of the arts are still *far from being satisfactory* to tackle the under-investigated surveillance FR problem in practical forensic scenarios. In particular, we discovered there is a dramatic gap in FR performance between existing celebrity photoshot benchmarks, e.g. the popular MegaFace, and this newly introduced surveillance imagery QMUL-SurvFace benchmark. For example, the CentreFace model can only yield 25.8% in Rank-1 rate on this new QMUL-SurvFace benchmark, whilst it was reported a significantly higher performance at 65.2% on the MegaFace benchmark in a closed-set setting. Face recognition is generally more difficult in an open-set setting which is typical for surveillance scenarios, owing to a large number of non-target people (distractors) appearing open spaced scenes. This is evidently so that on the new Surveillance FR Challenge, the top-performing CentreFace deep learning FR model on the MegaFace benchmark can now only achieve 13.2% *success rate* (at Rank-20) at a 10% *false alarm rate*. In this study, we also investigate the inevitable low-resolution problem inherent in surveillance facial imagery data by testing a combination of modern image super-resolution and face recognition models on both web and surveillance face images. Our evaluations suggest that contemporary image super-resolution methods are largely ineffective in improving surveillance FR performance, even though such methods have been shown to be effective in recovering high frequency details in web image data. Finally, we discuss some open research problems that need be addressed in order to overcome this Surveillance FR Challenge, which is evidently under-studied in the current literature.

Keywords Face Recognition · Surveillance Facial Imagery · Low-Resolution · Super-Resolution · Deep Learning · Large Scale · Open-Set · Closed-Set

Zhiyi Cheng and Shaogang Gong
School of Electrical Engineering and Computer Science, Queen Mary University of London, London, UK.
E-mail: z.cheng, s.gong@qmul.ac.uk

Xiatian Zhu
Vision Semantics Ltd., London, UK.
E-mail: eddy@visionsemantics.com

[1] The QMUL-SurvFace challenge is publicly available at `https://qmul-survface.github.io/`.

1 Introduction

Face recognition (FR) is a well established research problem in computer vision with the objective to recognise the human identities by their facial images (Zhao et al, 2003). There are other visual recognition based biometrics approaches to human identification, for example, whole-body person re-identification (Gong et al, 2014), iris recognition (Phillips et al, 2010), gait recognition (Sarkar et al, 2005), and fingerprint recognition (Maltoni et al, 2009). In this context, FR is considered as one of the more convenient and non-intrusive means for a wide range of identification applications from law enforcement and information security to business, entertainment and e-commerce (Wechsler et al, 2012). This is largely due to that face appearance is more reliable and stable than clothing appearance, provided that facial images are visible in sufficient resolution.

The FR problem has been extensively studied over the past few decades since 1970s, and it has become widely adopted in main-stream applications within the last five years. In 2016 alone, there were over 77,300 publications including patents (Google Scholar). In particular, the last few years have witnessed FR performance on high-resolution good quality web images reaching a level arguably better than the humans, e.g. 99.83% for 1:1 face verification on the LFW challenge and 91.76% for 1:N face identification on the million-scale MegaFace challenge. Increasingly, FR commercial products have become more mature and appeared increasingly in our daily life, e.g. web photo-album, online e-payment, and smart-phone e-banking, with such a trend in large scale deployment accelerating. One driving force behind this success in FR is the synergistic and rapid advances in deep learning neural network techniques, large scale facial imagery benchmarks, and powerful computing devices, not only enjoyed by the academic research community but also advocated at such speeds by the industry and commerce. One may argue then: "*The state-of-the-art FR algorithms, especially with the help of deep learning on large scale data, have reached a sufficient level of maturity and application readiness, evidenced by almost saturated performances on large scale public benchmark challenges*[2]*. Therefore, the FR problem should be considered as "solved" and the remaining efforts lie mainly in system production engineering.*"

However, as shown in this study, current FR models scale poorly to natively noisy and low-resolution images (not synthesised by down-sampling), that are typified by facial data captured in *unconstrained wide-field surveillance videos and images*, perhaps one of the most important FR application fields in practice. Specifically, we show that FR in the typical surveillance images is far away from being satisfactory especially at a large scale. Unlike recognising high-resolution web photoshot images of limited noise, the problem of surveillance face recognition remains extremely challenging and is unsolved. This should not be a surprise because surveillance visual data are characterised typically by low-resolution imagery with complex noise, e.g. motion blur, subject to poor imaging conditions giving rise to unconstrained pose, expression, occlusion, lighting and background clutter (Figure 1).

In the literature, the *surveillance face recognition* problem is significantly under-studied compared to FR on web photoshots. Whilst *web face recognition* is popular and commercially attractive due to the proliferation of social media and e-commerce on smart-phones, solving the surveillance face recognition challenge is critical for public safety and law enforcement applications. A major reason for the lack of development of robust and scalable FR models suitable for surveillance scenes is the lack of large scale true surveillance facial imagery benchmark data for both model learning and testing, unlike the widely available high-resolution web photoshot FR challenges that have provided the necessary data to enable FR benefiting from deep learning (Table 1). For example, there are 4,753,320 web face images from 672,057 face identities in the MegaFace2 challenge[3] (Nech and K-S, 2017). This is made possible by relatively easy (and cheap) collection and labelling of large scale facial images in the public domain from the Internet. On the contrary, it is prohibitively more expensive and less feasible to construct large scale native surveillance facial imagery data as a benchmark for wider studies, due both to highly restricted data access and very hard and expensive data labelling cost. Currently, the largest surveillance FR challenge benchmark is the UnConstrained College Students (UCCS) dataset[4] (Günther et al, 2017), which contains 100,000 face images from 1,732 face identities, at a significantly smaller scale than the MegaFace celebrities photoshot dataset. Moreover, critically the UCCS dataset images were captured by a high-resolution camera at a single location. In this study, we show that (1) the state-of-the-art deep learning based FR models trained on large scale high-resolution benchmark datasets such as the MegaFace generalise poorly to true surveillance face recognition tasks, and (2) the FR performance test on artificially synthesised low-resolution images by down-sampling does not reflect the true challenge of surveil-

[2] The FR performance differences among top-performing models are very marginal and negligible (see Figure 2).

[3] `http://megaface.cs.washington.edu/`

[4] `http://vast.uccs.edu/Opensetface/`

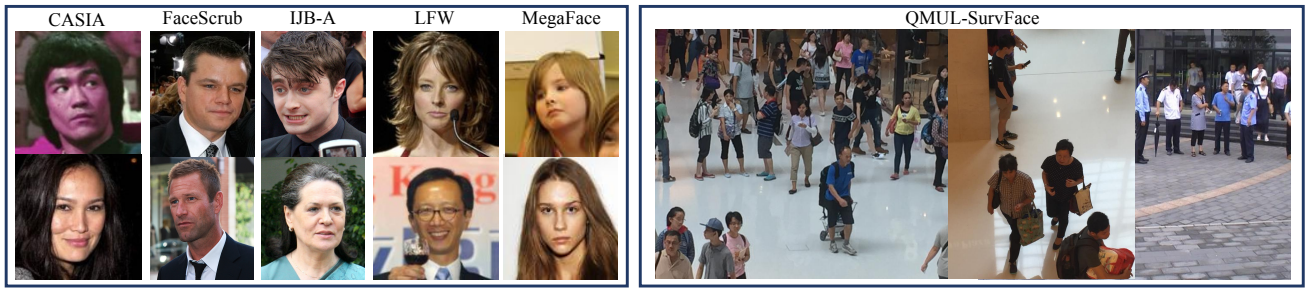


Fig. 1: Example comparisons of (Left) **web face images** from five renowned standard face recognition challenges and (Right) **native surveillance face images** from typical public surveillance scenes in real-world applications.

lance FR tasks when only native low-resolution facial images are available in model deployment.

More specifically, this study considers a more realistic and large scale *Surveillance Face Recognition Challenge* than what is currently available in the public domain, that is, recognising a person's identity by natively low-resolution surveillance facial images taken from unconstrained public scenes. We make three contributions: **(I)** We provide a dataset of large scale face identities with *native* surveillance facial imagery data for model development and evaluation in the community. More specifically, we introduce the *QMUL-SurvFace* challenge, which contains 463,507 face images of 15,573 face identities. To our best knowledge, this is the largest sized dataset for surveillance face recognition challenge, with *native* low-resolution surveillance facial images captured by unconstrained wide-field cameras from distances. This benchmark dataset is constructed by data-mining 17 public domain person re-identification datasets (Table 4) using a deep learning face detection model, so to assemble a large pool of labelled surveillance face images in an affordable manner. A unique feature of this new surveillance face recognition benchmark dataset is the provision of cross-location (cross camera views) identity label annotations, available from their corresponding person re-identification datasets. This cross-view labelling information can be useful for open-world (vs. existing closed-world) face recognition testing. **(II)** We benchmark five representative deep learning FR models (Liu et al, 2017, Parkhi et al, 2015, Schroff et al, 2015, Sun et al, 2014a, Wen et al, 2016) for both 1:N face identification and 1:1 verification tasks. In contrast to existing FR challenges typically considering a *closed-set* face identification, we particularly evaluated these algorithms for performing a more realistic *open-set* surveillance face recognition task. That is, the closed-set test assumes the existence of every probe face ID in the gallery so a true-match always exists for each probe, whilst the open-set test does not therefore a true-match may not exist, conforming to the more realistic large scale surveillance face recognition task. **(III)** We investigate the effectivenss of existing FR models on native low-resolution surveillance imagery data by exploiting simultaneously image super-resolution (Dong et al, 2014, 2016, Kim et al, 2016a, Lai et al, 2017, Tai et al, 2017) and the five representative deep learning face recognition models. We studied different combinations of super-resolution and face recognition models including both independent and joint model training schemes. Moreover, we further compared model performances on the existing MegaFace and UCCS benchmarks to gain better understanding of the unique characteristics of the proposed new QMUL-SurvFace challenge. We finally provide discussions on future research directions towards solving the Surveillance Face Recognition Challenge.

2 Related Work

In this section, we review and discuss the popular FR challenges (Section 2.1) and representative FR methods (Section 2.2) in the literature. In Section 2.2, we focus on the models closely related to the surveillance FR challenge including more recent deep learning algorithms. More general and extensive reviews can be found in many other surveys (Abate et al, 2007, Adini et al, 1997, Chang et al, 2005, Chellappa et al, 1995, Daugman, 1997, Ersotelos and Dong, 2008, Kong et al, 2005, Prado et al, 2016, Samal and Iyengar, 1992, Tan et al, 2006, Wang et al, 2014b, Zhao et al, 2003, Zou et al, 2007) and books (Gong et al, 2000, Li and Jain, 2011, Wechsler, 2009, Wechsler et al, 1998, 2012, Zhao and Chellappa, 2011, Zhou et al, 2006).

2.1 Face Recognition Challenges

An overview of representative FR challenges and benchmarks are summarised in Table 1. Specifically, early

Table 1: The statistics of representative publicly available face recognition benchmarks. Celeb: Celebrity.

Challenge	Year	IDs	Images	Videos	Subject	Surveillance?
Yale (Belhumeur et al, 1997)	1997	15	165	0	Cooperative	No
QMUL-MultiView (Gong et al, 1998)	1998	25	4,450	5	Cooperative	No
XM2VTS (Messer et al, 1999)	1999	295	0	1,180	Cooperative	No
Yale B (Georghiades et al, 2001)	2001	10	5,760	0	Cooperative	No
CMU PIE (Sim et al, 2002)	2002	68	41,368	0	Cooperative	No
Multi-PIE (Gross et al, 2010)	2010	337	750,000	0	Cooperative	No
Morph (Ricanek and Tesafaye, 2006)	2006	13,618	55,134	0	Celeb (Web)	No
LFW (Huang et al, 2007)	2007	5,749	13,233	0	Celeb (Web)	No
YouTube Wolf et al (2011)	2011	1,595	0	3,425	Celeb (Web)	No
WDRef (Chen et al, 2012)	2012	2,995	99,773	0	Celeb (Web)	No
FaceScrub (Ng and Winkler, 2014)	2014	530	100,000	0	Celeb (Web)	No
CASIA (Yi et al, 2014)	2014	10,575	494,414	0	Celeb (Web)	No
CelebFaces (Sun et al, 2014b)	2014	10,177	202,599	0	Celeb (Web)	No
IJB-A (Klare et al, 2015)	2015	500	5,712	2,085	Celeb (Web)	No
VGGFace (Parkhi et al, 2015)	2015	2,622	2.6M	0	Celeb (Web)	No
UMDFaces (Bansal et al, 2016)	2016	8,277	367,888	0	Celeb (Web)	No
CFP (Sengupta et al, 2016)	2016	500	7,000	0	Celeb (Web)	No
UMDFaces (Bansal et al, 2016)	2016	8,277	367,888	0	Celeb (Web)	No
MS-Celeb-1M (Guo et al, 2016)	2016	99,892	8,456,240	0	Celeb (Web)	No
UMDFaces-Videos (Bansal et al, 2017)	2017	3,107	0	22,075	Celeb (Web)	No
IJB-B (Whitelam et al, 2017)	2017	1,845	11,754	7,011	Celeb (Web)	No
VGGFace2 (Cao et al, 2017b)	2017	9,131	3.31M	0	Celeb (Web)	No
MegaFace2 (Nech and K-S, 2017)	2017	672,057	4,753,320	0	Non-Celeb (Web)	No
FERET (Phillips et al, 2000)	1996	1,199	14,126	0	Cooperative	No
FRGC (Phillips et al, 2005)	2004	466+	50,000+	0	Cooperative	No
PaSC (Beveridge et al, 2013)	2013	293	9,376	2,802	Cooperative	No
FRVT(Visa) (Grother et al, 2017)	2017	O(10^5)	O(10^5)	0	Cooperative	No
FRVT(Mugshot) (Grother et al, 2017)	2017	O(10^5)	O(10^6)	0	Cooperative	No
FRVT(Selfie) (Grother et al, 2017)	2017	<500	<500	0	Cooperative	No
FRVT(Webcam) (Grother et al, 2017)	2017	<1,500	<1,500	0	Cooperative	No
FRVT(Wild) (Grother et al, 2017)	2017	O(10^3)	O(10^5)	0	Uncooperative	No
FRVT(Child Exp) (Grother et al, 2017)	2017	O(10^3)	O(10^4)	0	Uncooperative	No
SCface (Grgic et al, 2011)	2011	130	4,160	0	Cooperative	Yes
UCCS (Günther et al, 2017)	2017	1,732	14,016+	0	Uncooperative	Yes
QMUL-SurvFace	2018	15,573	463,507	0	Uncooperative	Yes

challenges focus on *small-scale constrained* face recognition scenarios with limited number of images and identity classes available (Belhumeur et al, 1997, Georghiades et al, 2001, Gross et al, 2010, Messer et al, 1999, Phillips et al, 2010, Samaria and Harter, 1994, Sim et al, 2002). They provide neither sufficient appearance variation and diversity for robust model training, nor practically solid testing benchmarks. In 2007, the seminal Labeled Faces in the Wild (LFW) challenge (Huang et al, 2007) was proposed and started to shift the community towards recognising unconstrained celebrity faces by providing in-the-wild web face images and a standard performance evaluation protocol. The LFW challenge has contributed significantly to a spurring of interest and progress in the FR researches. This trend towards large scale benchmark datasets has been amplified by the creation of even larger scale FR benchmarks such as CASIA (Yi et al, 2014), CelebFaces (Sun et al, 2014b), VGGFace (Parkhi et al, 2015), MS-Celeb-1M (Guo et al, 2016), MegaFace (K-S et al, 2016) and MegaFace2 (Nech and K-S, 2017). Thus far, it seems that the availability of large scale training and test data benchmark for web photoshot images has been mostly addressed.

Due to such large scale benchmark challenges, face recognition accuracy in good quality (high-resolution) images has reached an unprecedented level by leveraging deep learning. For example, the FR performance has reached 99.83% (1:1 pair verification) on the LFW challenge and 91.76% (1:N identification with 1,000,000 distractors in the gallery) on the MegaFace challenge. However, such dramatic progresses on the web FR benchmarks do not generalise well to *native* low-resolution surveillance facial imagery data captured in arbitrarily unconstrained views from a distance (see our evaluations in Section 5.1). This is because: (1) Existing FR challenges such as LFW have varying degrees of data selection bias (near-frontal pose, less motion blur, good illumination and so forth); and (2) Contemporary deep learning algorithms are often domain-specific (i.e. only generalise well to face images similar to the training data in terms of appearance characteristics such as high

quality web face images under good lighting conditions from professional cameras and photographers). On the other hand, there is usually a huge appearance change in facial images between a web photoshot view and a surveillance view in-the-wild (Figure 1).

Research on surveillance face recognition has made little progress since the early days in 1996 when the well-known FERET challenge was introduced (Phillips et al, 2000). It is under-studied by large, with a very few benchmark challenges available. One of the major obstacles is the difficulty of establishing a large scale surveillance FR challenge due to the high cost and limited feasibility in collecting surveillance facial imagery data whilst also providing exhaustive facial identity annotation. Even in the FERET dataset, only simulated (framed) surveillance face images were collected in most cases with carefully controlled imaging settings, therefore it provided a much better facial image quality than those from truely native surveillance face data.

A notable recent study introduced the UCCS face challenge (Günther et al, 2017), which is currently the largest surveillance FR benchmark in the public domain. The facial images in the UCCS challenge were captured from a long-range distance without subjects' cooperation (unconstrained). The faces in these images are of various poses, blurriness and occlusion (Figure 9(b)). This benchmark represents a relatively more realistic surveillance FR scenario in comparison to the FERET dataset. However, the UCCS images were captured in high-resolution from a single view[5], therefore providing significantly more facial details with less viewing angle variations than what is available from typical surveillance facial images. Moreover, UCCS is relatively small in size, particularly in term of the face identity class numbers (1,732), therefore statistically limited for evaluating surveillance face recognition challenge (Section 5.1). In this study, we address the limitations of the UCCS benchmark by constructing a larger scale natively low-resolution surveillance face recognition challenge, the QMUL-SurvFace benchmark. It consists of 463,507 real-world surveillance face images (natively low-resolution) of 15,573 different face identity classes captured from a diverse source of public spaces, as described in Section 3.

2.2 Face Recognition Methods

We provide a brief review and discussion on the vast number of existing FR algorithms in the literature, including hand-crafted FR, deep learning FR, and low-resolution FR methods. The low-resolution FR methods are selected particularly for representing the state-of-the-art methods in handling poor image quality from surveillance face images. In this context, we also discuss image super-resolution (hallucination) techniques for enhancing image fidelity and improving face recognition model performance on low-resolution imagery data.

(I) Hand Crafted Feature Based Methods Most early FR methods rely on hand-crafted features (e.g. Color Histogram, LBP, SIFT, Gabor) and matching model learning algorithms (e.g. discriminative margin mining, subspace learning, dictionary based sparse coding, Bayesian modelling) (Ahonen et al, 2006, Belhumeur et al, 1997, Cao et al, 2013, Chen et al, 2012, 2013, Liu and Wechsler, 2002, Tan and Triggs, 2010, Turk and Pentland, 1991, Wolf and Levy, 2013, Wright et al, 2009, Zhang et al, 2007). These methods are often inefficient subject to heavy computational cost of high-dimensional feature representations and complex image preprocessing. Moreover, they also suffer from sub-optimal recognition generalisation particularly in a large sized data when there are significant facial appearance variations. This is jointly due to weak representation power (human domain knowledge used in hand-crafting features is limited and incomplete) and the lack of end-to-end interaction learning between feature extraction and model optimisation.

(II) Deep Learning Based Methods In the past five years, FR models based on deep learning, particularly deep convolutional neural networks (CNNs) (K-S et al, 2016, Klare et al, 2015, Liu et al, 2017, Masi et al, 2016, Parkhi et al, 2015, Schroff et al, 2015, Sun et al, 2014c, Taigman et al, 2014, Wen et al, 2016), have achieved remarkable success (Table 2). This paradigm benefits from superior network architectures (He et al, 2016, Krizhevsky et al, 2012, Simonyan and Zisserman, 2015, Szegedy et al, 2015) and discriminative learning optimisation algorithms (Schroff et al, 2015, Sun et al, 2014c, Wen et al, 2016). Deep FR methods naturally address the above limitations of hand-crafted features based alternatives by jointly learning the face representation and matching model end-to-end directly from raw training images. To achieve good performance, a large collection of labelled face images is usually necessary to optimise the large number (millions) of parameters in deep models. This requirement can be commonly met by using millions of web face images collected and labelled (filtered) from the Internet sources at a relatively lower cost. Consequently, modern FR models are often trained, evaluated and deployed on web face datasets (Table 1 and Table 2).

While the web FR performance achieves an unprecedented level, it is remains unclear how well the same

[5] A single Canon 7D camera equipped with a Sigma 800mm F5.6 EX APO DG HSM lens.

Table 2: Face verification performance of state-of-the-art FR methods on the LFW challenge. "*": Results from the official challenge leaderboard at `http://vis-www.cs.umass.edu/lfw/results.html`. M: Million.

Feature Representation	Method	Accuracy (%)	Year	Training IDs	Training Images
Hand Crafted	Joint-Bayes (Chen et al, 2012)	92.42	2012	2,995	99,773
	HD-LBP (Chen et al, 2013)	95.17	2013	2,995	99,773
	TL Joint-Bayes (Cao et al, 2013)	96.33	2013	2,995	99,773
	GaussianFace (Lu and Tang, 2015)	98.52	2015	16,598	845,000
Deep Learning	DeepFace (Taigman et al, 2014)	97.35	2014	4,030	4.18M
	DeepID (Sun et al, 2014c)	97.45	2014	5,436	87,628
	LfS (Yi et al, 2014)	97.73	2014	10,575	494,414
	Fusion (Taigman et al, 2015)	98.37	2015	250,000	7.5M
	VggFace (Parkhi et al, 2015)	98.95	2015	2,622	2.6M
	BaiduFace (Liu et al, 2015a)	99.13	2015	18,000	1.2M
	DeepID2 (Sun et al, 2014a)	99.15	2014	10,177	202,599
	CentreFace (Wen et al, 2016)	99.28	2016	17,189	0.7M
	SphereFace (Liu et al, 2017)	99.42	2017	10,575	494,414
	DeepID2+ (Sun et al, 2015b)	99.47	2015	12,000	290,000
	DeepID3 (Sun et al, 2015a)	99.53	2015	12,000	290,000
	FaceNet (Schroff et al, 2015)	99.63	2015	8M	200M
	TencentYouTu* (Tencent, 2017)	99.80	2017	20,000	2M
	EasenElectron* (Electron, 2017)	99.83	2017	59,000	3.1M

Table 3: The performance summary of state-of-the-art image super-resolution methods on five popular benchmarks. While none of these benchmarks are designed for the FR problem, these evaluations represent a generic capacity of contemporary super-resolution techniques in synthesising the high frequency details particularly for low-resolution web-style images. Metric: Peak Signal-to-Noise Ratio (PSNR), higher is better.

Model	Unscalling Times	Set5	Set14	B100	URGAN	MANGA
Bicubic	2	33.65	30.34	29.56	26.88	30.84
SRCNN (Dong et al, 2014)	2	36.65	32.29	31.36	29.52	35.72
FSRCNN (Dong et al, 2016)	2	36.99	32.73	31.51	29.87	36.62
VDSR (Kim et al, 2016a)	2	37.53	33.03	31.90	30.76	-
DRCN (Kim et al, 2016b)	2	37.63	32.98	31.85	30.76	37.57
LapSRN (Lai et al, 2017)	2	37.52	33.08	31.80	30.41	37.27
DRRN (Tai et al, 2017)	2	37.74	33.23	32.05	31.23	-
Bicubic	4	28.42	26.10	25.96	23.15	24.92
SRCNN (Dong et al, 2014)	4	30.49	27.61	26.91	24.53	27.66
FSRCNN (Dong et al, 2016)	4	30.71	27.70	26.97	24.61	27.89
VDSR (Kim et al, 2016a)	4	31.35	28.01	27.29	25.18	-
DRCN (Kim et al, 2016b)	4	31.53	28.04	27.24	25.14	28.97
LapSRN (Lai et al, 2017)	4	31.54	28.19	27.32	25.21	29.09
DRRN (Tai et al, 2017)	4	31.68	28.21	27.38	25.44	-
Bicubic	8	24.39	23.19	23.67	20.74	21.47
SRCNN (Dong et al, 2014)	8	25.33	23.85	24.13	21.29	22.37
FSRCNN (Dong et al, 2016)	8	25.41	23.93	24.21	21.32	22.39
LapSRN (Lai et al, 2017)	8	26.14	24.44	24.54	21.81	23.39

state-of-the-art methods can generalise to the poor quality surveillance face data. Intuitively, more computing challenges are involved in the surveillance FR scenario owing to three reasons: **(1)** Surveillance face images contain much less fine-grained details with much poorer quality and lower resolution in comparison to the web data (Figure 1), thus hindering severely the FR process from insufficient appearance information required by discriminative matching. **(2)** Deep models are highly domain-specific and likely yield great performance degradation in the cross-domain deployments. This is particularly so when the domain gap between the training and testing data is wide, such as the high quality web faces and low quality surveillance faces. In such cases, transfer learning of FR models is also non-trivial (Pan and Yang, 2010). The modelling challenge can be further increased by the scarcity of labelled surveillance face data and the data imbalance between two domains. **(3)** Instead of the common closed-set search considered in most existing methods, FR in the surveillance scenario is intrinsically an open-set search process where the probe face identity (the vast search space) is not necessarily present in the gallery (enrolled target people representing the operational interest). This property brings about a significant matching challenge by additionally requiring the FR system

to accurately reject non-target people (i.e. distractors) whilst simultaneously not missing the alarms of target identities in the operations. Intuitively, the open-set FR represents a more challenging task than the closed-set counterpart since the distractors can be associated with arbitrary variety and inherently constitute a vast majority proportion in surveillance searches.

(III) Low-Resolution Face Recognition One unique FR challenge in the real-world surveillance applications is the low-resolution problem (Wang et al, 2014b). Low-resolution means the lacking of fine-grained appearance information in the input facial images which inevitably gives rise to a higher matching difficulty. In addition, when performing FR on image pairs with large resolution difference, the resolution discrepancy problem will emerge. Generally, existing low-resolution FR methods fall into two strategies: **(1)** image super-resolution (Bilgazyev et al, 2011, Fookes et al, 2012, Gunturk et al, 2003, Hennings-Yeomans et al, 2008, Jia and Gong, 2005, Jiang et al, 2016, Li et al, 2008, Liu et al, 2005, Wang et al, 2016b, Xu et al, 2013, Zou and Yuen, 2012), and **(2)** resolution-invariant learning (Abiantun et al, 2006, Ahonen et al, 2008, Biswas et al, 2010, 2012, Choi et al, 2008, 2009, He et al, 2005, Lei et al, 2011, Li et al, 2009, 2010, Ren et al, 2012, Shekhar et al, 2011, Wang and Miao, 2008, Zhou et al, 2011).

In the first strategy, two model optimisation criteria are involved: pixel-level visual fidelity and face identity discrimination. Most existing methods separate the two optimisation processes in training with the focus on enhancing visual appearance details other than from the FR discrimination perspective (Gunturk et al, 2003, Wang and Tang, 2003). More recently, there are a few attempts (Hennings-Yeomans et al, 2008, Wang et al, 2016b, Zou and Yuen, 2012) on linking the two learning sub-tasks in a unified framework for more discriminative super-resolution.

In the second strategy, existing methods are designed to extract resolution-invariant features (Ahonen et al, 2008, Choi et al, 2009, Lei et al, 2011, Wang and Miao, 2008) or learning a cross-resolution structure transformation (Choi et al, 2008, He et al, 2005, Ren et al, 2012, Shekhar et al, 2011, Wong et al, 2010). As deep FR models are inherently data driven methods, they can also be conceptually categorised into this strategy as long as suitable training data is available to model optimisation.

Generally, all existing low-resolution FR methods share a number of limitations: (1) Considering small scale and/or downsampled artificial low-resolution face images under the closed-set protocol, therefore unable to accurately reflect the genuine surveillance FR challenge at scale. (2) Mostly relying on hand-crafted features and linear or shallow model structures, therefore subject to suboptimal generalisation. (3) Often requiring coupled low-resolution and high-resolution image pairs from the same domain for model training, which however may be insufficient or even unavailable in the real surveillance scenarios.

Image Super-Resolution In parallel to FR, recently image super-resolution techniques have also been rapidly developed thanks to the powerful modelling capacity of deep models (especially the family of CNNs) in regressing the pixel-wise loss between the reconstructed and ground-truth high-resolution images (Dong et al, 2014, 2016, Kim et al, 2016a,b, Lai et al, 2017, Ledig et al, 2016, Tai et al, 2017, Yang et al, 2014, Yu and Porikli, 2016). A performance summary of six state-of-the-art deep super-resolution models is given on five popular benchmarks in Table 3. Mostly, the FR and image super-resolution researches advance independently, with both assuming the availability of large scale high-resolution training data such as good-quality web images. In surveillance, high-resolution images are typically unavailable, which in turn resorts these existing methods to the less effective model transfer learning strategy. When the training and test data distributions are very different (typical in reality for the surveillance FR since it is rather difficult to collect pseudo image data with visual quality and distribution sufficiently close to the genuine surveillance data), this will become an extremely challenging image super-resolution problem with an extra need for new domain generalisation.

As an important domain-specific super-resolution, face hallucination is dedicated to the fidelity restoration of facial appearance by particularly exploiting face specific information such as facial part structure prior (Baker and Kanade, 2000, Cao et al, 2017a, Chakrabarti et al, 2007, Jia and Gong, 2008, Jin and Bouganis, 2015, Liu et al, 2007, Wang and Tang, 2005, Yu and Porikli, 2017, Zhu et al, 2016). A classic approach to hallucination is transferring the high-frequency details and structure information from exemplar high-resolution images based on the global and/or local cross-resolution relationship. This mapping relationship is typically learned from aligned low- and high-resolution image pairs. Moreover, existing methods often require noise-free input images and assume stringent part detection and dense correspondence alignment; Otherwise, overwhelming artifacts can be easily introduced in hallucination. These requirements significantly restrict their usability to the low-resolution surveillance face images due to the presence of uncontrolled noise and poor image quality, and the absence of coupled high-resolution images.

Table 4: Person re-identification datasets utilised in building the *QMUL-SurvFace* challenge.

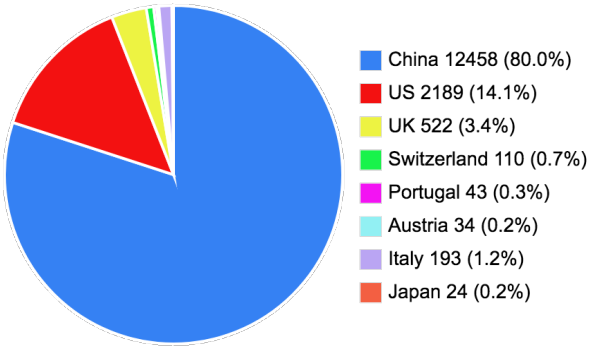
Person Re-Identification Dataset	IDs	Detected IDs	Bodies	Detected Faces	Nation
Shinpuhkan (Kawanishi et al, 2014)	24	24	22,504	6,883	Japan
WARD (Martinel and Micheloni, 2012)	30	11	1,436	390	Italy
RAiD (Das et al, 2014)	43	43	6,920	3,724	US
CAVIAR4ReID (Cheng et al, 2011)	50	43	1,221	141	Portugal
SARC3D (Baltieri et al, 2011b)	50	49	200	107	Italy
ETHZ (Schwartz and Davis, 2009)	148	110	8,580	2,681	Switzerland
3DPeS (Baltieri et al, 2011a)	192	133	1,012	366	Italy
QMUL-GRID (Loy et al, 2009)	250	242	1,275	287	UK
iLIDS-VID (Wang et al, 2014a)	300	280	43,800	14,181	UK
SDU-VID (Liu et al, 2015b)	300	300	79,058	67,988	China
PRID 450S (Roth et al, 2014)	450	34	900	34	Austria
VIPeR (Gray and Tao, 2008)	632	456	1,264	532	US
CUHK03 (Li et al, 2014)	1,467	1,380	28,192	7,911	China
Market-1501 (Zheng et al, 2015)	1,501	1,429	25,261	9,734	China
Duke4ReID (Gou et al, 2017)	1,852	1,690	46,261	17,575	US
CUHK-SYSU (Xiao et al, 2016)	8,351	6,694	22,724	12,526	China
LPW (Song et al, 2017)	4,584	2,655	590,547	318,447	China
Total	20,224	15,573	881,065	463,507	Multiple

3 The QMUL-SurvFace Recognition Challenge

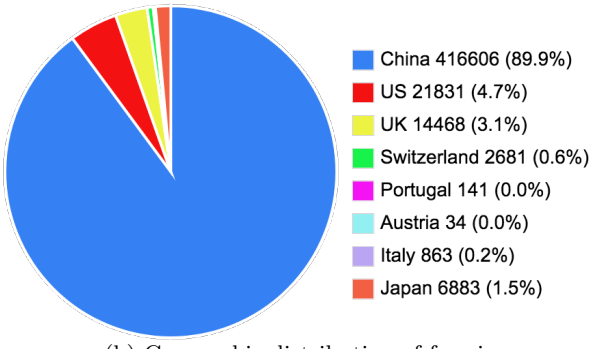
3.1 A Native Low-Res Surveillance Face Dataset

To our best knowledge, there is still no large scale genuine surveillance face recognition challenge in the public domain. To enable the research of this problem, we particularly construct a new large scale benchmark (challenge) by automatically extracting faces of the uncooperative general public from real-world surveillance images. We call this challenge ***QMUL-SurvFace***. Unlike most existing FR challenges utilising either high-quality web image sources or simulated surveillance images captured in controlled and/or limited forensic conditions and therefore *unsuitable* for evaluating and reflecting the realistic surveillance face recognition performance, we explore the real-world native surveillance image data from a combination of 17 person re-identification benchmarks. It is important to note that all these person re-identification datasets were collected in different real-world surveillance scenarios across diverse sites and multiple countries (Table 4).

Dataset Statistics The *QMUL-SurvFace* challenge contains 463,507 low-resolution face images from 15,573 person identities with other uncontrolled appearance variations in pose, illumination, motion blur, occlusion and background clutter (Figure 3). Among all, there are 10,638 (68.3%) people each associated with two or more detected face images. To our best knowledge, this is the largest native surveillance face dataset constructed on genuine surveillance image and video data sources in the public domain (Table 1).



(a) Geographic distribution of face identities.



(b) Geographic distribution of face images.

Fig. 2: Geographic distributions of (a) face identities and (b) face images in the *QMUL-SurvFace* challenge.

Face Image Collection To enable this benchmark represent a scalable evaluation scenario, these faces were extracted automatically by deploying the TinyFace detector (Hu and Ramanan, 2016) (one of the best thus-

far models with publicly available codes[6]) on person re-identification surveillance images (Figure 5). In real-world, manually labelling faces is non-scalable for processing the huge amount of surveillance image and video data. Note that not all faces in the person re-identification images can be successfully identified due to the imperfect generalisation of the detector, poor image quality, and extreme head poses. The average face detection recall is 77.0% (15,573 out of 20,224) in terms of identity and 52.6% (463,507 out of 881,065) in terms of image. Face detection statistics across all person re-identification datasets are summarised in Table 4.

Face Image Cleaning and Annotation To make an accurate FR challenge, the surveillance face dataset is further manually cleaned by filtering out the false detections. We also drop all the movie and TV program (non-surveillance) image data in the CUHK-SYSU dataset through a careful manual labelling process. These annotations were labelled by two independent annotators and further cross-checked by each other. For face annotation, we utilise the person class labels available in all these re-identification datasets. We assume that all these datasets have no identity overlap. This is rational since they were independently created over different time and surveillance venues, that is, persons appearing in more than one such person re-identification datasets should be extremely rare statistically.

Face Characteristics In contrast to existing FR challenges, the *QMUL-SurvFace* is uniquely characterised by very low resolution faces as typical in open surveillance scenarios (Figure 6) – one major source that makes the surveillance FR tasks very challenging (Wang et al, 2014b). The face spatial resolution ranges from 6/5 to 124/106 pixels in height/width, with the average as 24/20. Across all identities, face images exhibit a power-law distribution in frequency ranging from 1 to 558 (Figure 7). Besides, the people in this new benchmark have a wide variation in geographic origin with the majority from China[7] (Figure 2). Given the absence of fine-grained facial appearance, surveillance FR may be relatively less sensitive to face nationality in comparison to the high-resolution web FR setting.

[6] `http://mmlab.ie.cuhk.edu.hk/projects/WIDERFace/`

[7] In computing the nationality statistics, we simply assume that all people from the same person re-identification dataset share the same corresponding nationality label, since the per-identity nationality labels are not available in the source datasets.

Table 5: Benchmark data partition of the *QMUL-SurvFace* challenge. Numbers in parentheses: per-identity image size range.

Split	All	Training	Test
IDs	15,573	5,319	10,254
Images	463,507 (1∼558)	220,890 (2∼558)	242,617 (1∼482)

3.2 Evaluation Protocols

FR methods are typically evaluated under two applications (e.g. verification and identification) and two protocols (e.g. closed-set and open-set).

Data Partition To benchmark the evaluation protocol, we firstly need to split the entire *QMUL-SurvFace* data into the training and test sets. Specifically, we divide the 10,638 persons each with ≥ 2 face images into two halves: one half (5,319) as the training data, the other half (5,319) plus the remaining 4,935 single-shot identities (in total 10,254) as the test data See the summary in Table 5. We benchmark only *one* training-test data split rather than multiple ones since the dataset is already sufficiently large in terms of identity classes and face images to support statistically stable surveillance FR evaluations.

This single data partition universally applies to both identification and verification although their respective ways of utilising the test data are different due to distinct operational means. The face images of all training people are available for building FR models with the aim at learning the fine-grained variations in facial appearance. Additional imagery from other sources such as web images may be utilised for data augmentation in model training under the condition that no images of any test people are included for ensuring the complete non-overlapping between training and test data at identity level. Test data may be exploited differently under distinct operational protocols (see details below).

Next, we present the approach for benchmarking the *face identification* and *face verification* tasks on the *QMUL-SurvFace* challenge.

Face Verification The verification protocol measures the comparing performance of face pairs. Most FR methods evaluated on the popular LFW challenge adopt this protocol. Specifically, one presents a face image to a FR system with a claimed identity represented by an enrolled face. The system accepts the claim if their matching similarity score is greater than a threshold t, or rejects otherwise (Phillips et al, 2010). Similar to existing benchmarks (Huang et al, 2007, Klare et al, 2015), the protocol specifies the sets of *matched* and *unmatched* pairs that FR methods should perform in evaluation. For each test identity, we generate one matched pair,

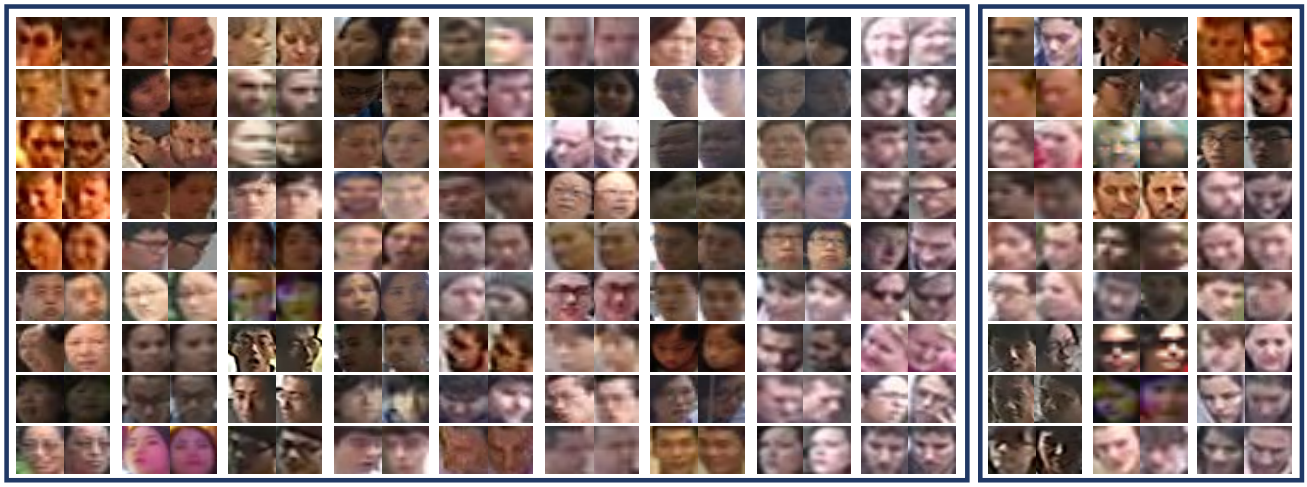


Fig. 3: Examples of face image pairs from the *QMUL-SurvFace* challenge. **Left box**: matched pairs. **Right box**: unmatched pairs.

Table 6: Benchmark face verification and identification protocols in the *QMUL-SurvFace* challenge. TAR: True Accept Rate; FAR: False Accept Rate; ROC: Receiver Operating Characteristic; FPIR: False Positive Identification Rate; TPIR: True Positive Identification Rate.

1:1 Face Verification Protocol			
Matched Pairs	5,319	Unmatched Pairs	5,319
Metrics	TAR@FAR, ROC		
1:N Face Identification Protocol			
Scenario	Open-Set		
Partition	Probe	Gallery	
IDs	10,254	3,000	
Images	182,323	60,294	
Metrics	TPIR@FPIR, ROC		

i.e. a total of 5,319 pairs (Table 6). We generate the same number (5,319) of unmatched pairs by randomly sampling between a face and nonmated ones. For performance measurement, each of these pairs is to be evaluated by computing a matching similarity score.

In the verification process, two types of error can occur: (1) A false accept – a distractor claims an identity of interest; (2) A false reject – the system mistakenly declines the identity of interest. As such, we define the False Accept Rate (FAR) as the fraction of unmatched pairs with the corresponding score s above threshold t

$$\mathrm{FAR}(t) = \frac{|\{s \geq t, \text{ where } s \in U\}|}{|U|} \tag{1}$$

where U denotes the set of unmatched pairs. In contrast, the False Rejection Rate (FRR) represents the fraction of matched pairs with matching score s below a threshold t:

$$\mathrm{FRR}(t) = \frac{|\{s < t, \text{ where } s \in M\}|}{|M|} \tag{2}$$

where M is the set of matched pairs. For understanding convenience, we further define the True Accept Rate (TAR), the complement of FRR, as

$$\mathrm{TAR}(t) = 1 - \mathrm{FRR}(t). \tag{3}$$

For face verification evaluation in this QMUL-SurvFace challenge, we utilise the paired TAR@FAR measure.

We also utilise the receiver operating characteristic (ROC) analysis measurement by varying the threshold t and generating a TAR-vs-FAR ROC curve. The overall accuracy performance can be measured by the area under the ROC curve, which is shorten as AUC. See the top half of Table 6 for the verification protocol summary.

Face Identification In forensic and surveillance applications however, it is the face identification that is of more interest (Best-Rowden et al, 2014, Ortiz and Becker, 2014). Face identification is arguably a more intricate and non-trivial problem, since a probe image must be compared against all gallery identities in the identification search process (K-S et al, 2016, Wang et al, 2016a).

Most existing FR methods in the literature consider the *closed-set* scenario by assuming that each probe subject is present in the gallery. We construct the evaluation setup for the closed-set scenario for this QMUL-SurvFace challenge through the following process. For each of the 5,319 multi-shot test identities, we randomly sample the corresponding images into the either probe or gallery set. Conceptually, the gallery set represents

Fig. 4: A glimpse of native surveillance face images from the *QMUL-SurvFace* challenge.

Fig. 5: An illustration of face detections in native surveillance person images. **Left box**: Auto-detected faces. **Right box**: Failure cases, i.e. the detector fails to identify the face due to low-resolution, motion blur, extreme pose, poor illumination and background clutters.

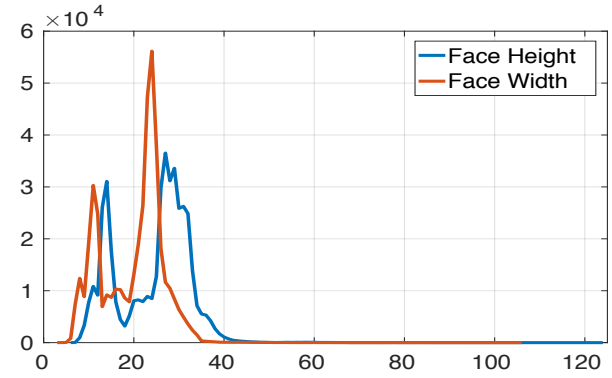


Fig. 6: The face image scale (width and height) distributions in the *QMUL-SurvFace* challenge.

Fig. 7: Face image frequency over all identities in the *QMUL-SurvFace* challenge.

imagery involved in an operational database, e.g. access control system's repository. For any unique person, we generate a single identity-specific face template from either one or multiple images enrolled into the gallery (Klare et al, 2015). This also makes the ranking list concise and more efficient for post-rank manual validation if available, e.g. without the case that a single identity takes multiple rank positions. Whilst the probe set represents imagery utilised to query a face identification system.

For performance evaluation in the *closed-set* identification, we select the widely used measure: the *Cumulative Matching Characteristic* (CMC) curve (Klare et al, 2015). The CMC curve reports the fraction of searches returning the mate (true match) at rank r or better, with the rank-1 rate as the most common summary indicator of an algorithm's efficacy. It is a non-threshold rank based metric. Formally, the CMC at rank r is defined as:

$$\mathrm{CMC}(r) = \sum_{i=1}^{r} \frac{N_{\mathrm{mate}}(i)}{N} \tag{4}$$

where $N_{\mathrm{mate}}(i)$ denotes the number of probe images with the mate ranked at position i, and N the total probe number.

In realistic surveillance applications, however, most faces captured by CCTV cameras are not of any gallery person and therefore should be detected as unknown, leading to the *open-set* face recognition protocol (Grother and Ngan, 2014, Liao et al, 2014). This is often referred to the *watch-list identification* (forensic search) scenario where only persons of interest are enrolled into the gallery, typically each identity with several different images such as the FBIs most wanted list[8]. To allow for the *open-set* surveillance FR test, we construct a watch list identification protocol where only face identities of interest are enrolled in the gallery. specifically, we create the following probe and gallery sets: (1) Out of the 5,319 multi-shot test identities, we randomly select 3,000 and sample half face images for each selected identity into the gallery set, i.e. the watch list. (2) All the remaining images including these single-shot imagery are used to form the probe set. As such, the majority of probe people are *unknown* (not enrolled gallery identities), which more accurately reflects the open space forensic search nature in practice.

For the *open-set* FR performance evaluation, we must quantify two error types (Grother and Ngan, 2014). The first type is *false alarm* – a face image from an unknown person (i.e. nonmate search) is incorrectly associated with one or more enrollees' data. This error can be quantified by the *False Positive Identification Rate* (FPIR):

$$\mathrm{FPIR}(t) = \frac{N_{\mathrm{nm}}^{\mathrm{m}}}{N_{\mathrm{nm}}} \tag{5}$$

which measures the proportion of nonmate searches $N_{\mathrm{nm}}^{\mathrm{m}}$ (i.e. no mate faces in the gallery) that produce one or more enrolled candidates at or above a threshold t (i.e. false alarm), among a total of N_{nm} nonmate searches attempted.

The second type of error is *miss* – a search of an enrolled target persons data (i.e. mate search) does not return the correct identity. We quantify the miss error by the *False Negative Identification Rate* (FNIR):

$$\mathrm{FNIR}(t, r) = \frac{N_{\mathrm{m}}^{\mathrm{nm}}}{N_{\mathrm{m}}} \tag{6}$$

which is the proportion of mate searches $N_{\mathrm{m}}^{\mathrm{nm}}$ (i.e. with mate faces present in the gallery) with enrolled mate

[8] `www.fbi.gov/wanted`

found outside top r ranks or matching similarity score below the threshold t, among N_{m} mate searches. By default, we set $r = 20$ (i.e. FNIR$(t, 20)$) which assumes a small workload by a human reviewer employed to review the candidates returned from an identification search. In practice, a more intuitive measure may be the "hit rate" or *True Positive Identification Rate* (TPIR):

$$\mathrm{TPIR}(t, r) = 1 - \mathrm{FNIR}(t, r) \tag{7}$$

which is the complement of FNIR offering a positive statement of how often mated searches are succeeded. In our QMUL-SurvFace challenge, we therefore adopt the TPIR@FPIR measure as the open-set face identification performance metrics. TPIR-vs-FPIR can similarly generate an ROC curve as above, the AUC of which stands for an overall measurement (See the bottom half of Table 6).

Link of open-set and closed-set The aforementioned performance metrics of closed-set and open-set FR are not completely independent but correlated. Specifically, the CMC(r) (Eqn. (4)) is can be regarded as a special case of TPIR(t, r) (Eqn. (7)) with ignored similarity scores by relaxing the threshold requirement t as:

$$\mathrm{CMC}(r) = \mathrm{TPIR}(t, r) \tag{8}$$

This metrics linkage is useful in enabling performance comparisons between closed-set and open-set.

Considerations In the literature, existing FR challenges adopt the closed-set evaluation protocol including the renowned MegaFace challenge (K-S et al, 2016). While being able to evaluate the FR model generalisation capability in the perspective of large scale search (e.g. 1 million gallery images in MegaFace), it does not fully confirm to the surveillance FR operation in law enforcement and other security applications. For surveillance FR settings, the human operators often have a list of target people in the mission with their face images enrolled in a working system (i.e. gallery). The fundamental FR task is hence to search the targets in wide and extensive public spaces over time against the gallery face images. This is intrinsically an open-set face matching problem. As a consequence, we mainly adopt the open-set identification evaluation protocol in our QMUL-SurvFace challenge (Table 6). In our evaluations, however, we still consider some closed-set FR experiments for enabling like-for-like comparisons with existing benchmarks.

4 Benchmark Evaluation

In the following, we first describe the deep learning FR methods used in our benchmarking evaluations. We then present the image super-resolution models useful for enhancing the poor surveillance face resolution.

4.1 Face Recognition Models

We present the formulation details of the FR models used in our surveillance FR challenge evaluation. We select 5 state-of-the-art deep learning FR methods which have previously shown a good performance on existing challenges (Table 2): DeepID2 (Sun et al, 2014a), CentreFace (Wen et al, 2016), FaceNet (Schroff et al, 2015), VggFace (Parkhi et al, 2015), and SphereFace (Liu et al, 2017). All these methods are based on the CNN model architecture with each characterised by different loss function and network designs. In model formulation, they aim at a common target of learning a discriminative face representation space that increases the inter-identity variations whilst decreases the intra-identity variations. Both types of variations are intrinsically complex and highly nonlinear considering that faces of the same identity may appear very differently under varying conditions whereas faces of some different identities however may look alike among a large population. Once any of these deep FR models is properly trained by the standard stochastic gradient descent (SGD) algorithm (Bottou, 2010, Rumelhart et al, 1988), we can deploy it as a discriminative feature extractor and perform the FR task with the general matching metrics such as the L_2 distance. We describe the selected FR models below.

The **DeepID2** model (Sun et al, 2014a) is characterised by simultaneously learning both face identification and face verification supervision signals in the training. Identification is to classify a face image into a large number of identity classes by the softmax cross-entropy loss (Krizhevsky et al, 2012). Formally, we predict the posterior probability $\tilde{y}_i$ of a face image $\boldsymbol{I}_i$ over the groundtruth identity label y_i among a total n_{id} distinct training people:

$$p(\tilde{y}_i = y_i|\boldsymbol{I}_i) = \frac{\exp(\boldsymbol{w}_{y_i}^\top \boldsymbol{x}_i)}{\sum_{k=1}^{|n_{\mathrm{id}}|} \exp(\boldsymbol{w}_k^\top \boldsymbol{x}_i)} \tag{9}$$

where $\boldsymbol{x}_i$ refers to the DeepID2 feature vector of $\boldsymbol{I}_i$, and $\boldsymbol{W}_k$ the prediction function parameter of the training identity class k. The identification training loss is computed as:

$$l_{\mathrm{id}} = -\log\Big(p(\tilde{y}_i = y_i|\boldsymbol{I}_i)\Big) \tag{10}$$

On the other hand, the verification signal encourages the DeepID2 features extracted from the same-identity faces to be similar, i.e. reducing the intra-person variations. This is achieved by enforcing the pairwise contrastive loss (Hadsell et al, 2006):

$$l_{\text{ve}} = \begin{cases} \frac{1}{2}\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2 & \text{if same ID,} \\ \frac{1}{2}\max\left(0, m - \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2\right)^2 & \text{otherwise.} \end{cases} \tag{11}$$

where m represents the discriminative identity class margin. The final DeepID2 model loss function is a weighted summation of the above two as:

$$\mathcal{L}_{\text{DeepID2}} = l_{\text{id}} + \lambda_{\text{bln}} l_{\text{ve}} \tag{12}$$

where λ_{bln} represents the balancing hyperparameter. A customised 5-layers CNN architecture is used in the DeepID2.

The **CentreFace** model (Wen et al, 2016) also adopts the softmax cross-entropy loss function (Eqn. (10)) as the DeepID2 for learning inter-class discrimination. However, it seeks for the intra-identity compactness in a class-wise manner by enforcing a representation constraint that any face image features should be close to the corresponding identity centre as possible. Learning this class compactness is accomplished by a simple centre loss function defined as:

$$l_{\text{centre}} = \frac{1}{2}\|\boldsymbol{x}_i - \boldsymbol{c}_{y_i}\|_2^2 \tag{13}$$

where y_i denotes the identity class of face images $\boldsymbol{x}_i$ and $\boldsymbol{c}_{y_i}$ the up-to-date feature centre of the class y_i. As such, all face images of the same identity are constrained to group together and therefore the intra-person variations can be effectively suppressed. The final loss function is similarly integrated with the identification supervision as:

$$\mathcal{L}_{\text{CentreFace}} = l_{\text{id}} + \lambda_{\text{bln}} l_{\text{centre}} \tag{14}$$

Since the feature space is dynamic in the course of model training, all class centres are progressively undated on-the-fly so as to preserve the compatibility with the SGD optimisation algorithm. The CentreFace model is implemented in a 28-layers ResNet architecture (He et al, 2016).

The **FaceNet** model (Schroff et al, 2015) utilises the triplet loss function to learn a binary-class (i.e. positive pairs *versus* negative pairs) discriminative feature embedding, different from both DeepID2 and CentreFace. The triplet loss (Liu et al, 2009) aims to obtain a discrimination margin between each face pair from one person to all other faces with the formal definition as:

$$l_{\text{tri}} = \max\left\{0,\ \alpha - \|\boldsymbol{x}_a - \boldsymbol{x}_n\|_2^2 + \|\boldsymbol{x}_a - \boldsymbol{x}_p\|_2^2\right\}, \quad \textit{subject to: } (\boldsymbol{x}_a, \boldsymbol{x}_p, \boldsymbol{x}_n) \in \mathcal{T} \tag{15}$$

where $\mathcal{T}$ denotes the set of triplets generated based on the identity labels, and α is a pre-fixed margin for separating the positive $(\boldsymbol{x}_a,\ \boldsymbol{x}_p)$ and negative $(\boldsymbol{x}_a,\ \boldsymbol{x}_n)$ training pairs. By doing so, the face images for one training identity are constrained to populate on an isolated manifold against other identities by a certain distance and therefore providing the discrimination capability. For fast convergence, it is critical to deploy more triplets that violate the triplet constraint (Eqn. (15)) for model update other than satisfied ones. To achieve this in a scalable fashion, we select hard positives and negatives within a mini-batch without the need for computing pairwise distances across the entire training set. In our FaceNet implementation, the Inception-ResNet CNN architecture (Szegedy et al, 2017) is utilised as a stronger replacement of the originally adopted ZF CNN (Zeiler and Fergus, 2014).

The **VggFace** model (Parkhi et al, 2015) considers both the identification and triplet training schemes in a sequential manner. Specifically, we firstly train the VggFace model by the softmax cross-entropy loss (Eqn. (10)). We then learn the feature embedding by the triplet loss (Eqn. (15)) where only the last full-connected layer is updated to implement the discriminative projection whilst all other layers are frozen. A similar hard sample mining strategy is applied in the second process for more efficient model optimisation. For the CNN design, the VggFace adopts the state-of-the-art 16-layers VGG16 architecture (Simonyan and Zisserman, 2015).

The **SphereFace** model (Liu et al, 2017) exploits a newly designed angular margin based angular softmax loss function. This differs from the triplet loss (Eqn. (15)) utilising an Euclidean distance margin by performing the feature discrimination learning in a hypersphere manifold. The motivation is that, multi-class deep features learned by the identification loss exhibit an intrinsic angular distribution. Formally, the angular softmax loss is formulated as:

$$\begin{aligned} l_{\text{ang}} &= -\log\Big(\frac{e^{\|\boldsymbol{x}_i\|\psi(\theta_{y_i,i})}}{e^{\|\boldsymbol{x}_i\|\psi(\theta_{y_i,i})} + \sum_{j\neq y_i} e^{\|\boldsymbol{x}_i\|\psi(\theta_{j,i})}}\Big), \\ &\textit{where } \psi(\theta_{y_i}, i) = (-1)^k \cos(m\theta_{y_i}, i) - 2k, \\ &\textit{subject to: } \theta_{y_i,i} \in [\frac{k\pi}{m}, \frac{(k+1)\pi}{m}],\ \ k \in [0, m-1] \end{aligned} \tag{16}$$

where $\theta_{j,i}$ represents the angle between the normalised identification weight $\boldsymbol{W}_j$ ($\|\boldsymbol{W}_j\| = 1$) of j-th class and the training sample $\boldsymbol{x}_i$, m ($m \geq 2$) the preset angular margin, and y_i the groundtruth class of $\boldsymbol{x}_i$. Specifically, this design manipulates the angular decision boundaries between different classes and enforces the constraint $\cos(m\theta_{y_i}) > \cos(\theta_j)$ for any $j \neq y_i$. When $m \geq 2$ and

$\theta_{y_i} \in [0, \frac{\pi}{m}]$, this inequation $\cos(\theta_{y_i}) > \cos(m\theta_{y_i})$ holds. Hence, $\cos(m\theta_{y_i})$ represents a lower bounder of $\cos(\theta_{y_i})$ and larger m leads to a wider angular inter-class margin. Similar to the CentreFace, the 28-layers ResNet CNN is adopted in the SphereFace implementation.

4.2 Image Super-Resolution Models

We present the formulation details of the image super-resolution methods for synthesising the high-frequency observations missing in the native low-resolution surveillance face images. We choose 5 state of the arts (Table 3) in our tests: SRCNN (Dong et al, 2014), FSRCNN (Dong et al, 2016), LapSRN (Lai et al, 2017), VDSR (Kim et al, 2016a), and DRRN (Tai et al, 2017). Similar to the selected FR methods above, these super-resolution models also exploit deep CNN architectures and can be optimised by the SGD algorithm in an end-to-end manner. By design, a super-resolution model aims to learn a highly non-linear mapping function between low-resolution and high-resolution images at the pixel level. This requires the availability of groundtruth low-resolution and high-resolution training pairs. Once a super-resolution model is optimised, we can deploy it to restore poor resolution surveillance faces before performing the FR task. Next, we describe the key designs of the selected super-resolution models.

The **SRCNN** model (Dong et al, 2014) is one of the first deep methods achieving remarkable success in image super-resolution. Its design is largely motivated by the pipeline of earlier sparse-coding based super-resolution methods (Kim and Kwon, 2010, Yang et al, 2010). By taking the end-to-end advantage of deep learning techniques, the SRCNN formulates originally separated components in the sparse-coding alternative methods into an unified deep learning framework and consequently realises a better mapping function optimisation. For training the SRCNN, the standard mean squared error (MSE) is adopted as the loss function:

$$l_{\text{mse}} = \|f(\boldsymbol{I}^{\text{lr}}; \boldsymbol{\theta}) - \boldsymbol{I}^{\text{hr}}\|_2^2 \tag{17}$$

where $\boldsymbol{I}^{\text{lr}}$ and $\boldsymbol{I}^{\text{hr}}$ denotes a pair of low-resolution and high-resolution training images, and the function $f()$ represents the to-be-learned super-resolution function with the parameters denoted by $\boldsymbol{\theta}$. Note that, this model takes the bicubic interpolated low-resolution images as input in that it is a fully convolutional network without the upscaling capability.

The **FSRCNN** model (Dong et al, 2016) is an accelerated and more accurate variant of SRCNN (Dong et al, 2014). This is achieved by taking the original low-resolution images as model input, designing a deeper hourglass (shrinking-then-expanding) shaped non-linear mapping module, and adopting a deconvolutional layer as the replacement of bicubic interpolation for unscaling the low-resolution input. The MSR loss function (Eqn. (17)) is used for model training.

The **VDSR** model (Kim et al, 2016a) improves over the SRCNN (Dong et al, 2014) by increasing the network depth from 3 to 20 convolutional layers. The rational before deeper cascaded network structure is to exploit richer contextual information over large image regions (e.g. 41×41 in pixel) for enhancing high frequency detail recovery. For effectively training this deep model, the residual learning scheme is adopted. That is, the model is optimised to learn the residual image between the input interpolated low-resolution image and the groundtruth high-resolution image. The VDSR model training is supervised by the MSE loss (Eqn. (17)) between the reconstructed image (i.e. the sum of model input and output due to the residual learning nature) and the groundtruth.

The **DRRN** model (Tai et al, 2017) further constructs a even deeper 52-layers network by jointly exploiting residual and recursive learning in a unified framework, going beyond VDSR. In particular, except for the global residual learning between the input and output as VDSR, this method also exploits local residual learning via short-distance identity branches to mitigate the information loss across all the layers. This leads to a multi-path structured network module. Inspired by (Tai et al, 2017), all modules share the parameters and input so that multiple recursions can be performed in an iterative fashion without increasing the model parameter size. The same MSE loss function (Eqn. (17)) is used to supervise the model training.

The **LapSRN** model (Lai et al, 2017) consists in a multi-levels of cascaded sub-networks that is designed to progressively predict high-resolution reconstructions in a coarse-to-fine fashion. This scheme is therefore contrary to the four one-step reconstruction models as above. Similar to VDSR and DRRN, the residual learning scheme is exploited in conjunction with the upscaling mapping function for reducing the model optimisation difficulty whilst enjoying more discriminative learning in a deeper network. To supervise the model training, it adopts the Charbonnier penalty function (Bruhn et al, 2005):

$$l_{\text{cpf}} = \sqrt{\|f(\boldsymbol{I}^{\text{lr}}; \boldsymbol{\theta}) - \boldsymbol{I}^{\text{hr}}\|_2^2 + \varepsilon^2} \tag{18}$$

where ε (e.g. set to 10^{-3}) is a pre-given noise constant. The additional benefit over the MSE is the potential to suppress outliers in training data. Note that, each level is supervised concurrently with a separate loss signal against the corresponding groundtruth high-resolution

images. This multi-loss structure resembles the benefits of deeply-supervised models (Lee et al, 2015, Xie and Tu, 2015).

Note that, using the MSE loss function for model optimisation intrinsically favours the Peak Single-to-Noise Ratio (PSNR) metric, a popular quantitative super-resolution performance measurement, whilst also suiting the deep learning methods since it is derivable. However, this cannot guarantees the human perceptual quality, e.g. face identity discrimination. One main phenomenon is that the synthesised high-resolution images by an MSE supervised model are very likely to be overly-smoothed and blurred. We will evaluate the effects of these super-resolution models for low-resolution surveillance FR tasks below (Section 5.2).

5 Experimental Results

In this section, we present and discuss the experimental results of the surveillance FR evaluation. The performance is compared among different FR methods firstly on the native low-resolution surveillance faces (Section 5.1) and then on super-resolved faces (Section 5.2).

5.1 Low-Resolution Surveillance Face Recognition

We evaluated the FR performance on the *native* low-resolution and poor quality QMUL-SurvFace images. Note that, apart from the low-resolution issue, there are also other uncontrolled covariates and noises such as illumination variations, expression, occlusions, background clutter, compression artifacts, all of which can cause observation ambiguity and inference uncertainty to varying degrees (Figure 3).

Model Training and Test For training any selected FR model, we adopted three strategies in terms of what face data is used for model training optimisation: **(1)** Only using the low quality QMUL-SurvFace training set (220,890 images from 5,319 identities). **(2)** Only using the larger scale high quality CASIA web face images (494,414 images from 10,575 identities). We do not choose the larger-sized MegaFace2 (Nech and K-S, 2017) as the model training dataset due to two reasons. The *first* is that the CASIA has provided sufficient training face images for achieving descent web FR performance. For example, the CASIA trained DeepID2, CentreFace, FaceNet, VggFace, SphereFace models can already reach 95.0%, 98.9%, 96.0%, 97.3%, 98.7% face verification results on the LFW challenge (Huang et al, 2007) in our experiments. The *second* is that, the huge MegaFace2 contains an over large number of face identities which in turn causes additional model training difficulty in practice. For instance, the need itself for fitting an over-large and thus over-sparse class distribution is non-trivial (Nech and K-S, 2017). **(3)** Firstly using the CASIA web faces to pre-train a FR model, then performing domain adaptation by fine-tuning the model on QMUL-SurvFace training faces. Once any FR model was trained with any of above strategies, we deployed it to perform the FR tasks with the generic Euclidean distance metric.

In both model training and deployment, we rescaled all the face images by the *bicubic* interpolation to the desired input size of any FR model. For easing conceptual understanding, such interpolated images are still considered as of "low-resolution" since the underlying resolution is mostly unchanged even though it may have a larger spatial size (Figure 14).

Evaluation Settings We considered both face verification and face identification. By default, we utilised the more realistic open-set evaluation protocol for face identification, unless stated otherwise. For open-set face identification evaluation (Section 5.1.1), we used TPIR (Eqn. (6)) at varying FPIR rates (Eqn. (5)). The true match ranked in top-r (i.e. $r = 20$ in Eqn. (7)) is considered as a successful identification. For face verification evaluation (Section 5.1.2), we used the TAR (Eqn. (3)) and FAR (Eqn. (1)).

Implementation Details The codes for CentreFace (Wen et al, 2016), VggFace (Parkhi et al, 2015), and SphereFace (Liu et al, 2017) were obtained from the original authors. Since the original code of FaceNet (Schroff et al, 2015) was not released, we utilised a high-quality TensorFlow reimplementation available online[9]. We reimplemented the DeepID2 model (Sun et al, 2014a). Throughout the following experiments, we either adopted the parameter setting as suggested by the authors when available, or carefully tuned them by grid search.

5.1.1 Face Identification Evaluation

We conducted six sets of face identification evaluation: **(I)** Benchmarking the face verification performance on the QMUL-SurvFace dataset; **(II)** Comparing the identification performance between open-set and closed-set; **(III)** Comparing the identification performance between surveillance faces and web faces; **(IV)** Evaluating the effect of surveillance face image quality; **(V)** Evaluating the effect of test data scalability; **(VI)** Qualitative evaluation.

(I) Benchmarking Results on QMUL-SurvFace We benchmarked the face verification results on the

[9] `https://github.com/davidsandberg/facenet`

Table 7: Face identification evaluation on the *QMUL-SurvFace* challenge. Protocol: Open-Set. Metrics: TPIR20@FPIR ($r = 20$) and AUC. "-": No results available due to failure of model convergence.

Training Source	QMUL-SurvFace				CASIA (Liu et al, 2015c)				CASIA + QMUL-SurvFace			
Metrics	TPIR20@FPIR			AUC	TPIR20@FPIR			AUC	TPIR20@FPIR			AUC
	0.3	0.2	0.1		0.3	0.2	0.1		0.3	0.2	0.1	
DeepID2	-	-	-	-	0.040	0.021	0.008	0.079	0.120	0.075	0.028	0.187
CentreFace	**0.236**	**0.173**	**0.101**	**0.323**	0.057	0.044	0.023	0.076	**0.245**	**0.196**	**0.132**	**0.337**
FaceNet	0.087	0.059	0.032	0.163	0.040	0.030	0.008	0.064	0.120	0.075	0.039	0.186
VggFace	-	-	-	-	**0.065**	**0.048**	**0.025**	**0.096**	0.030	0.017	0.006	0.067
SphereFace	-	-	-	-	0.059	0.042	0.022	0.090	0.158	0.118	0.068	0.215

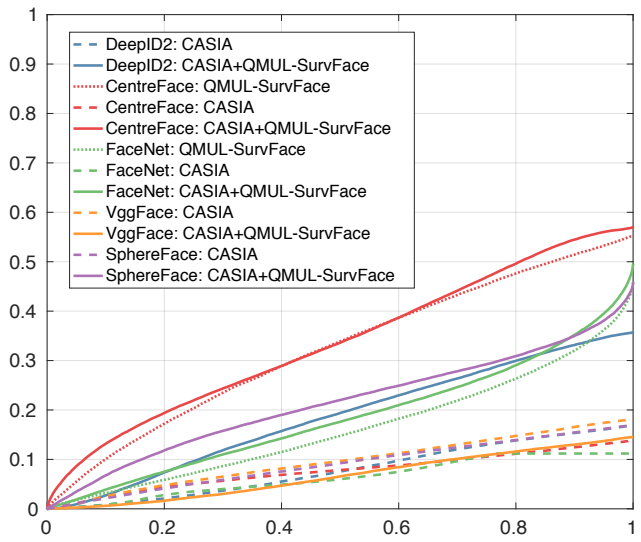


Fig. 8: Face identification evaluation on the *QMUL-SurvFace* challenge. Protocol: Open-Set. Metrics: TPIR20@FPIR ($r = 20$).

QMUL-SurvFace. From Table 7 and Figure 8, We have these main observations:

(1) Not all FR models (only CentreFace and FaceNet) can converge when directly training on the QMUL-SurvFace images from random model parameter initialisation. In contrast, all models can be successfully trained on the CASIA data. It is true that, the CASIA (494,414 images of 10,575 identities) is larger than the QMUL-SurvFace training set (220,890 images of 5,319 identities), i.e. 2.24× as many images and 1.99× as many identities. However, we believe that the scale is not a major issue concerning the model convergence, since the QMUL-SurvFace training set shall be sufficiently large for general deep learning as validated by the success training of CentreFace and FaceNet. We consider the more plausible reason to be some extremely high challenges imposed to the learning algorithms from such poor quality surveillance face images with significant lacking of fine-grained appearance details. That is, the FR learning models may tend to run into a difficulty of extracting discriminative representations from enormously ambiguous observations, particularly starting with arbitrary parameter initialisation. Additionally, only from these model performances, one can recognise a clear and significant difference between native surveillance FR and web FR, and importantly the unusual challenges involved in the former task.

(2) The lowest FR accuracies are generally yielded by the models trained with only the CASIA web face images. Among these, the VggFace model achieves the overall performance. We note again that the CASIA provides more data than the QMUL-SurvFace training set. Despite that, the results are still not surprised, as there exists a wide domain gap between CASIA and QMUL-SurvFace (Figure 9), and simply using more training data of visually less-similar auxiliary domain helps little. For example, both CentreFace and FaceNet achieve respectively much better FR performance when trained on the QMUL-SurvFace training set. This shows that, the domain gap problem really matters.

(3) However, when using the CASIA faces for model pre-training and then the QMUL-SurvFace images for model fine-tuning, better performances can be obtained by most FR algorithms. This indicates a positive effect of web face information as rich prior knowledge in boosting the surveillance FR performance. By this scheme, both the arbitrary model initialisation and domain gap problems are addressed to some degree. Nevertheless, one exception is the VggFace algorithm which turns out to degrade after the fine-tuning on the surveillance images (the target domain data). A possible reason is due to the more stringent high-resolution face assumption – requiring a 224×224 input scale *versus* 112×96 by other models – and therefore introducing larger amount of artefacts and noises in the process of upscaling the native low-resolution surveillance faces. This requirement is likely to introduce negative effects larger in amount than the positive benefits from model fine-tuning.

(4) Among all these algorithms, the CentreFace, trained with both CASIA and QMUL-SurvFace images, generates the best face matching performance, e.g. achieving 0.337 AUC. This indicates the superiority of the

idea of explicitly controlling intra-class variation over other competing techniques in the challenging surveillance FR tasks, an extended observation complementary to the finding in web FR setting as provided by Wen et al (2016).

(II) Open-Set *versus* Closed-Set We particularly compared the FR difficulty in the open-set and closed-set protocol on the same QMUL-SurvFace challenge, i.e. a comparison setting of cross-protocol. Specifically, the closed-set task in this evaluation corresponds to that all the probe persons are target people without any distractors involved. For this evaluation, we selected the top-2 FR models CentreFace (Wen et al, 2016) and SphereFace (Liu et al, 2017). We adopted the third training strategy to optimise the FR models using both CASIA web faces and QMUL-SurvFace native surveillance faces in a sequential manner.

Table 8 shows that, ***the closed-set FR is easier than the open-set counterpart***, e.g. 0.569 Rank20 *versus* 0.245 $\sim$ 0.132 TPIR20@FPIR0.3 $\sim$ 0.1 by the CentreFace model, and 0.460 Rank20 *versus* 0.158 $\sim$ 0.068 TPIR20@FPIR0.3$\sim$0.1 by the SphereFace model. This means that, once constrained by the false alarm rate through a decision threshold (t), the FR success performance (in TPIR) will drop dramatically particularly given strict FPIR requirements. In other words, when considering the attacks of distractor faces, an inherent deployment characteristic in the open surveillance scenario, the FR task is made much challenging.

(III) Surveillance Faces *versus* Web Faces We further compared the FR difficulty in the native surveillance images from the QMUL-SurvFace against in the web images from the MegaFace. For this evaluation, we used the common *closed-set* protocol (CMC metric) as the MegaFace to make a more conventional comparison. Specifically, we chose the CentreFace, the best model on QMUL-SurvFace, for an example. It is evident that FR in the low quality QMUL-SurvFace images (0.258 Rank1, Table 8) is much harder than in the high quality MegaFace images (0.652 Rank1 (Wen et al, 2016)), i.e. 60.4% (1 - 0.258/0.652) performance drop. We note that, the model needs to handle additionally 1 million distractors in the MegaFace gallery, which brings extra difficulty. In contrast, no distractors are involved in the gallery of QMUL-SurvFace. This indicates that the surveillance FR task can be even harder than what is shown by the above numerical comparisons.

(IV) The Quality of Surveillance Face Images We evaluated the effect of surveillance image quality in the FR task. To this purpose, we carried out a face identification performance comparison on UCCS (Günther et al, 2017) and QMUL-SurvFace. The UCCS face images have much better quality than the QUML-SurvFace ones (Figure 9), as they were captured by a selected high-resolution camera (unrepresentative to real-world surveillance scenarios with much low imaging quality).

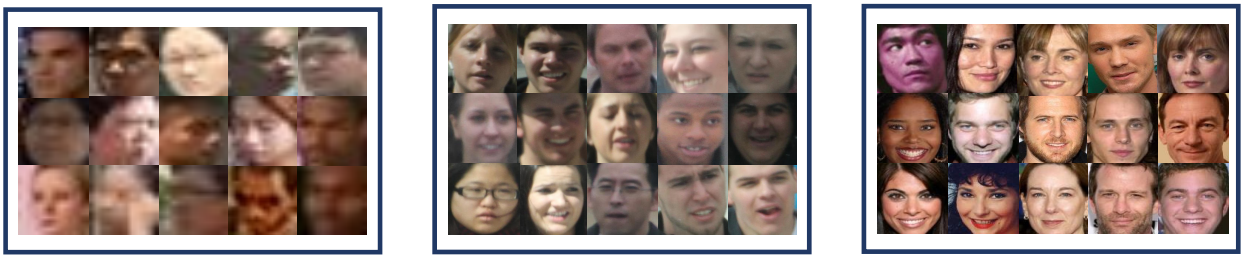
Table 8: Open-Set versus Closed-Set in face identification on the *QMUL-SurvFace* challenge. Protocols: Open-Set and Closed-Set. Metrics: TPIR20@FPIR ($r = 20$) for open-set, CMC for closed-set.

Metrics	TPIR20@FPIR (Open-Set)		
	0.3	0.2	0.1
CentreFace	0.245	0.196	0.132
SphereFace	0.158	0.118	0.068
Metrics	CMC (Closed-Set)		
	Rank1	Rank10	Rank20
CentreFace	0.258	0.497	0.569
SphereFace	0.221	0.406	0.460

Evaluation setting For the UCCS, we have access to the released face images from 1,090 identities out of 1,732 with the remaining held only by the challenge organisers. We randomly performed a 545/545 training/test identity split, corresponding to a 6,948/7,068 image split. To make a like-for-like comparison, we particularly built a same-ID QMUL-SurvFace training/test counterpart by randomly sampling the full training/test data: 7,018/6,919 images from 545/545 training/test identities. We call this set ***QMUL-SurvFace(1090ID)***. We utilised the more realistic *open-set* identification evaluation. For each of the two datasets, we therefore constructed a test setting by randomly selecting 100 test identities for the gallery set and using all 545 for the probe set. To train the FR models, we applied the third strategy using the CASIA web faces for pre-training and the UCCS or QMUL-SurvFace surveillance faces for fine-tuning.

Results Table 9 show that the QMUL-SurvFace images present more identification challenges to all evaluated FR models. For example, by CentreFace, the AUC decreases from 0.959 to 0.599, and the TPIR20@FPIR0.1 from 0.914 to 0.370. Varying-degree performance drops are observed in other models. This confirms the expectation that the image quality is an important factor in the face identification performance. Critically, this indicates that the UCCS challenge is a less accurate benchmark in reflecting the surveillance FR difficulty in the real-world operations.

(V) The Test Scalability of Surveillance FR We also examined the impact of face image scalability in surveillance FR deployment. We performed this by comparing QMUL-SurvFace (*10,254 probe and 3,000 gallery identities*) with the smaller QMUL-SurvFace(1090ID) (*545 probe and 100 gallery identities*).



(a) QMUL-SurvFace (b) UCCS (c) CASIA

Fig. 9: Quality comparison of example faces from (a) QMUL-SurvFace, (b) UCCS, and (c) CASIA.

Table 9: Evaluating the effect of surveillance image quality in face identification on UCCS and *QMUL-SurvFace(1090ID)*. Protocol: Open-Set. Metrics: TPIR20@FPIR and AUC.

Challenge	UCCS (Günther et al, 2017)				QMUL-SurvFace(1090ID)			
Metrics	TPIR20@FPIR			AUC	TPIR20@FPIR			AUC
	0.3	0.2	0.1		0.3	0.2	0.1	
DeepID2 (Sun et al, 2014a)	0.690	0.600	0.492	0.739	0.518	0.402	0.249	0.599
CentreFace (Wen et al, 2016)	**0.952**	**0.939**	**0.914**	**0.959**	0.526	0.456	0.370	0.599
FaceNet (Schroff et al, 2015)	0.768	0.738	0.666	0.817	**0.566**	**0.506**	**0.413**	**0.620**
VggFace (Parkhi et al, 2015)	0.740	0.611	0.445	0.770	0.445	0.357	0.225	0.538
SphereFace (Liu et al, 2017)	0.671	0.619	0.519	0.700	0.492	0.403	0.304	0.559

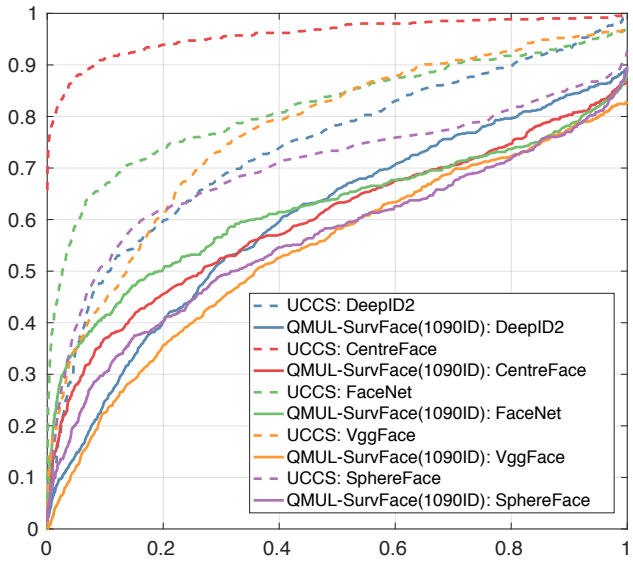


Fig. 10: Evaluating the effect of image quality in face identification on UCCS and *QMUL-SurvFace(1090ID)*. Protocol: Open-Set. Metrics: TPIR20@FPIR.

Table 10: Face verification evaluation on *QMUL-SurvFace*. Metric: TAR@FAR and AUC.

Metrics	TAR@FAR				AUC
	0.3	0.1	0.01	0.001	
DeepID2	0.804	0.634	0.366	0.229	0.846
CentreFace	**0.954**	**0.858**	**0.624**	**0.407**	**0.949**
FaceNet	0.945	0.808	0.482	0.208	0.937
VggFace	0.647	0.424	0.150	0.044	0.742
SphereFace	0.789	0.648	0.390	0.225	0.835

Table 11: Face verification evaluation on *QMUL-SurvFace*. Metric: Mean Accuracy.

Metric	Mean Accuracy
DeepID2 (Sun et al, 2014a)	76.90
CentreFace (Wen et al, 2016)	**87.90**
FaceNet (Schroff et al, 2015)	85.67
VggFace (Parkhi et al, 2015)	67.59
SphereFace (Liu et al, 2017)	77.46

The comparison between Figure 8 (Table 7) and Figure 10 (Table 9) shows that clearly higher FR performances can be obtained on the smaller test data for all five FR models. For example, the AUC score of the CentreFace model increases from 0.337 on QMUL-SurvFace to 0.599 on QMUL-SurvFace(1090ID), i.e. a 77.7% (0.599 / 0.337 - 1) relative boost. The relative increase in the TPIR20@FPIR0.1 is even higher at 180.3% (0.370 / 0.132 - 1). This evidence suggests that a large scale testing benchmark shall be necessary and crucial in order to more accurately reflect the inherent search difficulty of real-world surveillance FR.

(VI) Qualitative Evaluation To provide a visually intuitive evaluation for the surveillance FR challenge, we show 8 face identification examples by the best FR model CentreFace on the QMUL-SurvFace benchmark in Figure 11. Among them, the first six tasks succeed to find the true match in top-20 while the remaining last two fail.

Fig. 11: Qualitative evaluations of face identification on the *QMUL-SurvFace* challenge. Each row represents a face identification task, with the leftmost image as the probe followed by top-20 ranks of gallery images. The true match faces are indicated by red bounding boxes. These results are generated by the best face identification model CentreFace.

Table 12: Face verification evaluation on UCCS and *QMUL-SurvFace(1090ID)*. Metric: TAR@FAR and AUC.

Challenge	UCCS (Günther et al, 2017)					QMUL-SurvFace(1090ID)				
Metrics	TAR@FAR				AUC	TAR@FAR				AUC
	0.3	0.1	0.01	0.001		0.3	0.1	0.01	0.001	
DeepID2 (Sun et al, 2014a)	0.931	0.824	0.585	0.379	0.934	0.797	0.592	0.294	0.161	0.839
CentreFace (Wen et al, 2016)	**0.995**	**0.969**	**0.871**	**0.733**	**0.989**	**0.858**	**0.678**	**0.322**	**0.182**	**0.878**
FaceNet (Schroff et al, 2015)	0.962	0.897	0.670	0.361	0.961	0.820	0.609	0.283	0.143	0.851
VggFace (Parkhi et al, 2015)	0.965	0.899	0.728	0.509	0.934	0.693	0.467	0.188	0.037	0.767
SphereFace (Liu et al, 2017)	0.932	0.827	0.546	0.214	0.933	0.741	0.539	0.209	0.067	0.797

(a) Ground-truth: matched pairs. (b) Ground-truth: unmatched pairs.

Fig. 12: Qualitative evaluations of face verification on the *QMUL-SurvFace* challenge: (a) 12 matched pairs and (b) 12 unmatched pairs. The model prediction for each face pair is shown underneath with *cross mark* as unmatched and *tick mark* as matched. The failure predictions are indicated in red colour. These results are generated by the best face verification model *CentreFace*.

Table 13: Evaluating the effect of image super-resolution in surveillance face identification on the *QMUL-SurvFace* challenge. Protocol: Open-Set. Metric: TPIR20@FPIR and AUC. SR: Super-Resolution. Jnt Train: Joint Training; Ind Train: Independent Training. The best performance by each FR method across all SR models is indicated in **purple** colour. The relatively better performance between the Independent (Ind) and Joint (Jnt) training strategies using the same FR and SR models is highlighted in **bold**.

Metrics		TPIR20@FPIR			AUC	TPIR20@FPIR			AUC	TPIR20@FPIR			AUC
		0.3	0.2	0.1		0.3	0.2	0.1		0.3	0.2	0.1	
SR Model \ FR Model		DeepID2 (Sun et al, 2014a)				CentreFace (Wen et al, 2016)				FaceNet (Schroff et al, 2015)			
No SR		**0.120**	**0.075**	**0.028**	**0.187**	**0.245**	**0.196**	**0.132**	**0.337**	**0.120**	**0.075**	**0.039**	**0.186**
SRCNN	Ind Train	0.037	0.021	0.008	0.078	0.063	0.051	0.029	0.081	0.060	0.051	0.028	0.070
	Jnt Train	**0.045**	**0.030**	**0.008**	**0.075**	**0.230**	**0.166**	**0.097**	**0.329**	-	-	-	-
FSRCNN	Ind Train	0.035	0.020	0.008	0.078	0.061	0.047	0.026	0.080	0.058	0.049	0.028	0.070
	Jnt Train	**0.074**	**0.039**	**0.012**	**0.162**	**0.208**	**0.148**	**0.083**	**0.310**	-	-	-	-
VDSR	Ind Train	0.037	0.020	0.008	0.078	0.060	0.047	0.026	0.078	0.061	0.050	0.027	0.071
	Jnt Train	**0.087**	**0.048**	**0.018**	**0.180**	**0.244**	**0.179**	**0.103**	**0.337**	-	-	-	-
DRRN	Ind Train	0.036	0.020	0.008	0.078	0.058	0.045	0.024	0.078	0.060	0.049	0.027	0.071
	Jnt Train	-	-	-	-	-	-	-	-	-	-	-	-
LapSRN	Ind Train	0.036	0.020	0.008	0.078	0.059	0.047	0.026	0.078	0.060	0.049	0.027	0.070
	Jnt Train	-	-	-	-	-	-	-	-	-	-	-	-
SR Model \ FR Model		VggFace (Parkhi et al, 2015)				SphereFace (Liu et al, 2017)							
No SR		0.030	0.017	0.006	0.067	**0.158**	**0.118**	**0.068**	**0.215**				
SRCNN	Ind Train	**0.064**	**0.046**	**0.025**	**0.098**	0.054	0.038	0.020	0.086				
	Jnt Train	-	-	-	-	-	-	-	-				
FSRCNN	Ind Train	0.061	0.043	0.023	0.094	0.048	0.032	0.016	0.083				
	Jnt Train	-	-	-	-	-	-	-	-				
VDSR	Ind Train	0.062	0.044	0.022	0.095	0.054	0.036	0.019	0.087				
	Jnt Train	-	-	-	-	-	-	-	-				
DRRN	Ind Train	0.061	0.042	0.022	0.094	0.053	0.036	0.018	0.086				
	Jnt Train	-	-	-	-	-	-	-	-				
LapSRN	Ind Train	0.062	0.043	0.023	0.095	0.053	0.036	0.018	0.086				
	Jnt Train	-	-	-	-	-	-	-	-				

5.1.2 Face Verification Evaluation

Following the face identification above, we further evaluated the surveillance face verification (1:1) performance on both QMUL-SurvFace and UCCS challenges. For training the FR models, we adopted the two-stage fine-tuning based strategy. We conducted four sets of evaluations: **(I)** Benchmarking the face verification performance on QMUL-SurvFace; **(II)** Evaluating the effect of surveillance face image quality; **(III)** Comparing the verification performance between surveillance faces and web faces; **(IV)** Qualitative evaluations.

(I) Benchmarking Results on QMUL-SurvFace We benchmarked the face verification results on the QMUL-SurvFace. In Table 10, we similarly observed that the CentreFace model is still the best performer as in Table 7. A clear difference is that the FaceNet achieves the second rank by beating both SphereFace and DeepID2. This somehow suggests the task difference between face identification and verification, with the relative superiority of a FR model dependent on the target task. All FR models perform very poorly at the lower FAR (e.g. 0.001). This means that the face verification task in surveillance images also remains a very difficult task.

(II) Surveillance Faces *versus* Web Faces For this comparison, we selected the most popular LFW web FR challenge (Huang et al, 2007). To allow for a direct comparison with the published results (Table 2), we additionally evaluated the QMUL-SurvFace by the *mean accuracy* metric, i.e. the proportion of correctly verified pairs among all, which is benchmarked in the LFW challenge. We observed from Table 11 that the best performance on the QMUL-SurvFace is 87.90% obtained by the CentreFace model, hence considerably lower than the best result 99.83% on LFW. This confirms again the higher challenge of surveillance FR as compared to web FR, consistent with the results of face identification above.

(III) The Quality of Surveillance Face Images We evaluated the effect of surveillance image quality in face verification by comparing again QMUL-SurvFace and UCCS. To this end, we similarly generated 5,319 positive and 5,319 negative test pairs for UCCS (Table

6). For a like-for-like comparison between UCCS and QMUL-SurvFace, we used the training data (7,922 images from 545 identities) of the QMUL-SurvFace(1090ID) set for FR model optimisation. We observed in Table 12 that: For any of these five FR methods, the performance on the UCCS is clearly higher than that on the QMUL-SurvFace. For example, the CentreFace obtains 0.989 on the UCCS *versus* 0.878 on the QMUL-SurvFace in AUC. This is largely consistent with the face identification results presented in Table 9 and Figure 10.

(IV) Qualitative Evaluation For giving some visual evaluation, we show 24 face verification examples by the strongest face model CentreFace at FAR=0.1 on the QMUL-SurvFace in Figure 12.

5.2 Super-Resolution in Surveillance Face Recognition

Following the face recognition evaluation on the original low resolution images, we evaluated the FR performance on the *super-resolved* surveillance face images. The purpose is to examine the effect of contemporary image super-resolution or hallucination deep learning techniques in addressing the low-resolution challenge in surveillance FR. In the followings, we only evaluated the native and more challenging surveillance faces in the QMUL-SurvFace challenge, since the UCCS images are already of much higher resolutions (Figure 9).

Model Training and Test For enhancing the FR model with image super-resolution, we took two approaches in deep learning paradigm:

(1) ***Independent Training***: We firstly trained the FR models by pre-training on the CASIA (Liu et al, 2015c) and fine-tuning on the QMUL-SurvFace. This is the same as in Section 5.1. We then *independently* trained the image super-resolution models with the CASIA face data alone, since no high-resolution QMUL-SurvFce faces are available. For this, we need to additionally generate the low-resolution training faces by down-sampling the high-resolution CASIA faces to construct the training pairs required by the super-resolution models.

At test time, we deployed the learned super-resolution model to first restore the low-resolution surveillance face images before performing the FR matching with deep features and the Euclidean distance metric.

(2) ***Joint Training***: Training the super-resolution and FR models *jointly* in a single hybrid deep model to maximise their mutual compatibility. To achieve this, we united the super-resolution model with the FR model by connecting the output of the former with the input of the latter and simultaneously trained the united model. More specifically, we firstly performed the joint learning using only the CASIA data including both low-resolution and high-resolution faces. We then fine-tuned the FR network with the QMUL-SurvFace training data to mitigate the domain gap problem. That is, all the super-resolution layers are frozen in the fine-tuning process.

However, we found in our experiments that this joint training is not always feasible for any FR and super-resolution model combination. This is due to additional challenges arising, such as too-large model size and higher training difficulty to converge the model. In our experiments, we succeeded in jointly training six hybrid models among three super-resolution models (SRCNN (Dong et al, 2014), FSRCNN (Dong et al, 2016), VDSR (Kim et al, 2016a)) and two FR models (DeepID2 (Sun et al, 2014a) and CentreFace (Wen et al, 2016)).

At test time, the unified model can be directly deployed on the native low-resolution surveillance images to extract deep features for performing the FR tasks using the Euclidean distance.

Evaluation Settings We utilised the same performance metrics for face verification (TPIR Eqn. (6) and FPIR Eqn. (5)) and face identification (TAR Eqn. (3) and FAR Eqn. (1)).

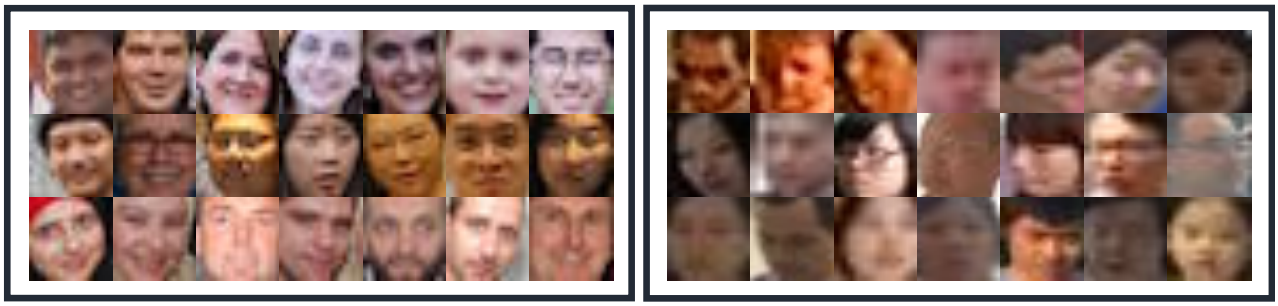
Implementation Details For super-resolution, we performed the common $4\times$ upscaling restoration on the low-resolution input faces. We used the codes from the original authors for all image super-resolution models (Dong et al, 2014, 2016, Kim et al, 2016a, Lai et al, 2017, Tai et al, 2017). In model training, we either followed the parameter setting as suggested by the authors if available, or carefully tuned them throughout all the experiments below.

5.2.1 Face Identification Evaluation

In this section, we conducted two sets of evaluation: **(I)** Evaluating the effect of image super-resolution on surveillance face identification; **(II)** Comparing the effect of image super-resolution on surveillance faces and web faces.

(I) Effect of Image Super-Resolution We tested the effect of image super-resolution for surveillance FR on the QMUL-SurvFace. Table 13 reports the face identification performance on the super-resolved QMUL-SurvFace test data. We have these observations:

(1) It is surprisingly found that, exploiting existing image super-resolution algorithms to enhance the face image quality mostly does not bring any positive but *negative* effect to the surveillance FR performance, either by independent or joint model training. This suggests that applying existing SR models cannot solve the low-resolution problem in the native surveillance FR. The



(a) Simulated low-resolution MegaFace2 images. (b) Genuine low-resolution QMUL-SurvFace.

Fig. 13: A visual perception comparison between (a) simulated low-resolution MegaFace2 images and (b) genuine low-resolution QMUL-SurvFace images.

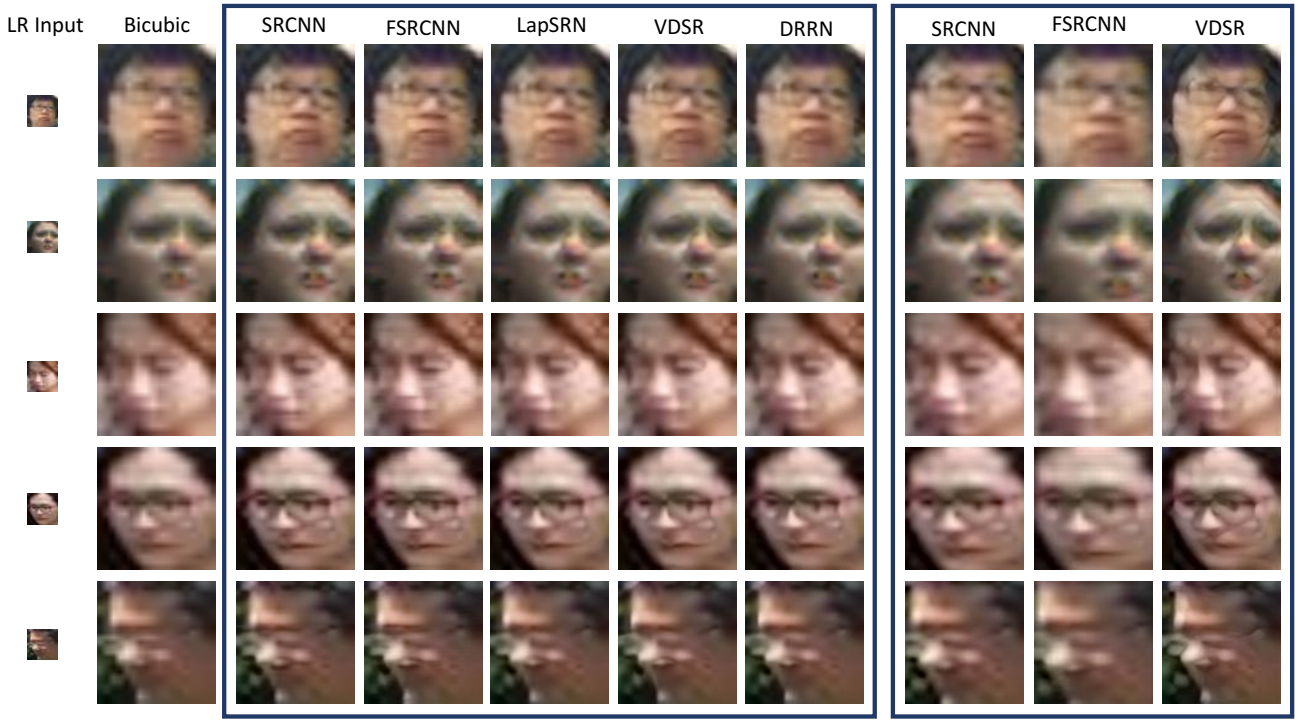


Fig. 14: Examples of super-resolved faces on the low-resolution (LR) QMUL-SurvFace images. In **left** box: the high-resolution images generated by independently trained super-resolution models. In **right** box: the high-resolution images generated by jointly trained super-resolution models. The *CentreFace* model is exploited in the joint training scheme along with these super-resolution models.

plausible reasons are threefold. The *first* is inherent to the super-resolution model design that the MSE (mean squared error) supervision (Eqn. (17)) is not a high-level perceptual quality measurement, but a low-level pixel-wise metric. The *second* is that, the super-resolution models were trained by the CASIA *non-surveillance* web faces which appear very different from the QMUL-SurvFace images (Figure 1), and therefore probably leading to the notorious domain drift problem (Pan and Yang, 2010). The *third* is the introduction of additional artefacts and noises which may adversely affect the FR performance (Figure 14). The exception case (same as in Section 5.1) is the VggFace model which requires originally higher resolution input faces, and therefore is more likely to benefit due to less noises introduced by the super-resolution methods in the unscaling process. However, the overall performance of VggFace is very similar to the one when trained on the CASIA faces alone (the middle of Table 7), without a clear improvement either.

(2) Independently trained super-resolution model is significantly inferior to jointly trained. This means that

additionally considering the FR discrimination in super-resolution model training is useful and effective. This is also visually validated to some degree as in Figure 14, for example, the jointly trained VDSR model generates less blurred faces. However, the overall performance of joint model training is still worse than that without using super-resolution. This further indicates the low suitability of existing image quality enhancing algorithms in handling the low-resolution surveillance FR problem. Overall, these evidences above suggest a clear necessity of developing more advanced algorithms for more effectively addressing the genuine large scale surveillance FR problem beyond the *joint learning* approach.

Table 14: Effect of image super-resolution on the simulated low-resolution *MegaFace2* web face images (Nech and K-S, 2017). These test web faces were generated by down-sampling. The joint training scheme was adopted. Protocol: Open-Set. Metric: TPIR20@FPIR and AUC. SR: Super-Resolution.

Metrics	TPIR20@FPIR			AUC
	0.3	0.2	0.1	
FR / SR	CentreFace (Wen et al, 2016)			
No SR	0.399	0.280	0.133	0.460
SRCNN	0.266	0.192	0.090	0.365
FSRCNN	0.263	0.195	0.107	0.353
VDSR	**0.400**	**0.283**	**0.141**	**0.475**

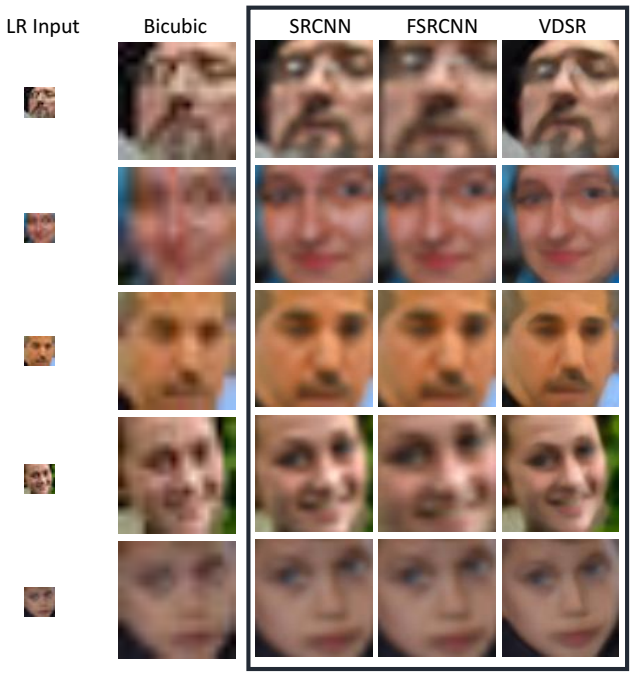


Fig. 15: Examples of super-resolved faces of simulated low-resolution (LR) MegaFace2 web images (Nech and K-S, 2017). The *CentreFace* model is used in the joint training scheme along with three deep learning super-resolution models.

(II) Surveillance Faces *versus* Web Faces The results that image super-resolution cannot improve low-resolution surveillance FR are somewhat contrary to the conclusions in some existing works (Huang and He, 2011, Liu et al, 2005, Wang and Tang, 2005, Wang et al, 2014b). To further examine the effect of contemporary super-resolution models in low-resolution featured surveillance FR, we evaluated them on simulated down-sampled web faces. This eliminates significantly the domain drift issue as encountered by QMUL-SurvFace.

Evaluation setting For this evaluation, we constructed a web face training/test set with the same identity as the QMUL-SurvFace dataset (Table 5) by randomly sampling the MegaFace2 data (Nech and K-S, 2017). All selected MegaFace2 web faces were then artificially down-sampled to the size of 24×20 pixels in face height and width, the average size of the QMUL-SurveFace images (Figure 13). This customised MegaFace2 set contains only non-celebrity people and therefore ensures no identity overlap with the model training data CASIA celebrity face images. We built a similar open-set face identification test setting with 3,000 identities (51,949 images) in the gallery and 10,254 identities (176,990 images) in the probe (Table 6). We also randomly and similarly sampled a training subset with 81,355 face images from 5,319 people without identity overlap against the test set. As such, we created a like-for-like comparison setting using low-resolution web faces to the QMUL-SurvFace setting. We adopted the most effective *joint training* strategy.

Results Table 14 shows that only the VSDR model can slightly boost the simulated low-resolution web face identification performance, whilst other two models similarly yield negative effects. This suggests that the contemporary super-resolution methods generally are not clearly effective in synthesising FR discriminative high frequency details (Figure 15) even trained under a concurrent supervision of face identity classification and pixel value regression in an end-to-end joint manner. Relatively, the overall identification performance on the simulated low-resolution web data is better than on the genuine native counterpart. This is also qualitatively reflected when comparing the super-resolved images (Figure 15 vs Figure 14). This further indicates that the low-resolution surveillance FR is more challenging in that we have no access to groundtruth high-resolution surveillance face images.

Table 15: Evaluating the effect of image super-resolution in surveillance face verification on the *QMUL-SurvFace* challenge. The joint training scheme was adopted. Metrics: TAR@FAR and AUC. SR: Super-Resolution. All the models are fine-tuned on the QMUL-SurvFace training subset.

Metrics	TAR@FAR				AUC
	0.3	0.1	0.01	0.001	
SR \ FR	CentreFace (Wen et al, 2016)				
No SR	0.954	0.858	0.624	0.407	0.949
SRCNN	0.965	0.893	0.618	0.389	0.959
FSRCNN	0.959	0.886	0.602	0.366	0.956
VDSR	**0.972**	**0.911**	**0.651**	**0.440**	**0.966**

5.2.2 Face Verification Evaluation

Apart from the face identification evaluation as above, we further evaluated the effect of image super-resolution for pairwise face verification on the QMUL-SurvFace challenge. In this analysis, we selected the best FR model CentreFace (Wen et al, 2016) and the most effective *joint training* scheme.

Table 15 shows that the VDSR model yields a clear improvement in the overall performance and the other two slightly improve the AUC score. This observation is unexpected and different from that in face identification (Table 13). This is a second inconsistency aspect discovered in our evaluations in addition to the FR model relative superiority (Table 9 vs. Table 12), partially due to the discrepancy in FR operational means (an under-studied research issue in the literature). However, the verification performance remains unsatisfactory at strict false alarm rates (FAR), e.g. 0.001. Yet, this condition actually is highly desired in real-world operations since it is closely relevant to the system usability criterion. This indicates that image super-resolution still requires further investigations for improving face verification in native surveillance facial images.

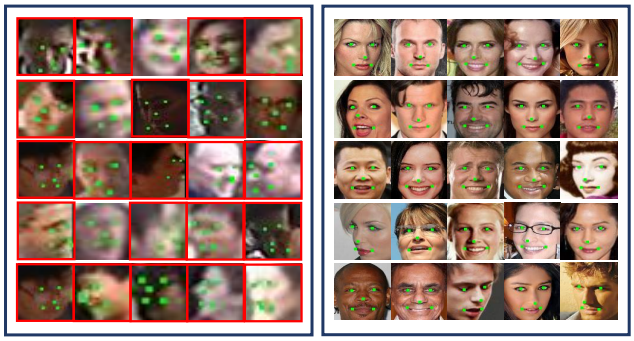
6 Discussion and Conclusion

In this work, we have presented a large scale benchmark with native low-resolution facial images for a true Surveillance Face Recognition Challenge, the *QMUL-SurvFace* dataset, including a large real-world surveillance face dataset and extensive benchmarking results with in-depth discussions and analysis. In contrast to existing FR challenges on which the FR performance has saturated, this challenge shows that state-of-the-art algorithms remains unsatisfactory in handling poor surveillance face images, even by knowledge transfer from abundant auxiliary web faces. For concluding remarks, we discuss a number of research directions which we consider to be worthwhile investigating for the future researches.

Transfer Learning From the benchmarking results (Table 13), it is evident that transferring the knowledge of auxiliary web datasets is beneficial to boost the surveillance FR performance, e.g. by model pre-training. However, more research is needed to make this knowledge transfer more effective. Given the clear domain discrepancy in appearance between surveillance and non-surveillance face images, effective domain adaptation algorithms (Ganin and Lempitsky, 2015, Ghifary et al, 2016, Pan and Yang, 2010, Saenko et al, 2010, Tzeng et al, 2015) shall be important in transferring discriminative information across domains. Among existing techniques, style transfer (Gatys et al, 2016, Li and Wand, 2016, Li et al, 2017, Zhu et al, 2017) is one intuitive approach to transfer learning across distinct domains. The basic idea is to transform the source images into the target domain style (e.g. surveillance imaging style) so that the richly labelled discriminative information in the source domains (e.g. web images) can be directly exploited to train supervised FR models. Whilst solving the FR discriminative style transfer problem is inherently challenging, it may promise worthwhile potential for the surveillance FR problem.

Resolution Restoration As mentioned earlier, the low-resolution issue can hinder the surveillance FR performance. Image super-resolution is one intuitive and natural approach. However, our extensive evaluations suggest that current super-resolution algorithms are still not satisfactory in addressing this issue in surveillance FR. Two main reasons are: (1) We have no access to native high-resolution surveillance face images required by training the super-resolution models (Yang et al, 2014). (2) It is non-trivial to explore the resolution transformation learning from the web data despite at scale due to the wide appearance domain gap (Pan and Yang, 2010). Although image super-resolution techniques have been adopted to enhance low-resolution FR in previous works (Wang et al, 2014b), they rely on hand-crafted feature representations with limited evaluations on only small simulated test settings but no clear efficacy evidence on large scale genuine surveillance FR deployments. Therefore, one interesting research topic in the future can be unsupervised transfer deep learning of image super-resolution for restoring native low-resolution surveillance faces.

Semantic Attributes Face attributes have served as a type of meaningful mid-level representation for boosting face verification and identification in normal or high resolution images (Berg and Belhumeur, 2013, Kumar



(a) QMUL-SurvFace (b) CASIA web faces

Fig. 16: Qualitative comparisons of facial landmark detection (Zhang et al, 2016) on randomly sampled (a) QMUL-SurvFace images and (b) CASIA web faces images. The detected five landmarks (left/right eye centres, the nose tip, left/right mouth corners) are denoted by green dots. Test samples with incorrect landmark detections are indicated by the red box. It is evident that the deep learning model fails on the most low-resolution surveillance faces, whilst succeeds on all high-resolution web faces.

et al, 2009, Manyam et al, 2011, Song et al, 2014). Attribute recognition in face images is challenging due to complex appearance variations in pose, expression, lighting, occlusion, and background (Farhadi et al, 2009, Parikh and Grauman, 2011) as well as imbalanced class distributions (Dong et al, 2017, Huang et al, 2016). Clearly, this task can be much harder given poor quality surveillance face images. However, this research has been made more plausible by the introduction of large face attribute datasets (Liu et al, 2015c). Encouraged by the existing non-surveillance FR methods above, we anticipate that semantic attributes can be important part of learning facial representation for improving surveillance FR performance.

Face Alignment Face alignment or facial landmark detection is an important preprocessing step in face recognition (Chen et al, 2012, Wen et al, 2016, Zhang et al, 2016). Although great progress has been made (Cao et al, 2014, Zhang et al, 2014, Zhu et al, 2015), robust alignment remains a formidable challenge in poor quality surveillance images (see Figure 16 for qualitative evaluations in the context of comparing with web face images). Similar to super-resolution and attribute detection, this task also suffers from the severe domain shift challenge because the landmark labels are typically only available on non-surveillance web images in the public domain (Burgos-Artizzu et al, 2013, Le et al, 2012, Zhu et al, 2015). One straightforward method is therefore to construct a large surveillance face landmark dataset by deploying exhaustive manual annotation efforts so that existing deep learning methods can be effectively exploited. Besides, integrating landmark detection and super-resolution jointly into the FR models is also conceptually feasible and interesting.

Contextual Constraints Considering the incomplete and noisy information involved in surveillance face images, it is also essential to discover and model the context knowledge as auxiliary constraints to assist the FR matching in open spaces. In real-world scenes, people often travel in groups. The group structure therefore provides meaningful information (social force) that can be useful in reasoning identity classes of individual members (Gallagher and Chen, 2009, Helbing and Molnar, 1995, Hughes, 2002, Zheng et al, 2009). To enable the modelling of such social context, there is a need for further extending this new challenge with additional face group images.

Open-Set Recognition Surveillance FR is essentially an open-set recognition problem (Bendale and Boult, 2015, 2016) since it requires to identify if a probe person belongs to one of the target identities or not. In reality, most probes can be non-target considering the vast surveillance video search space. It is thus potentially beneficial when the model can be discriminatively learned to construct a decision boundary on all target people (Zheng et al, 2016). At large, open set recognition is developing independently of FR in the literature. We thus expect a future direction on solving the two problems jointly.

Zero-Shot Learning More broadly, FR is also related to zero-shot learning (Fu et al, 2017) where no training samples of test identity classes are available. However, zero-shot learning focuses on the knowledge transfer across known and unknown classes vis some intermediate representation such as semantic attributes (Fu et al, 2014) and word vectors (Frome et al, 2013) in the classification setup. Both known and unknown classes are usually pre-defined with assigned representations, e.g. class prototypes. On the contrary, FR typically does not rely on any bridging information but aims to learn an identity-sensitive representation from the training face classes with the hope to universally generalise to an arbitrary number of new identities. In this sense, FR can be considered as a generalised zero-shot problem. Nonetheless, this does not prevent from borrowing some technical ideas from zero-shot learning to benefit open-set surveillance FR formulation.

Imagery Data Scalability Compare to the existing web FR benchmarks (K-S et al, 2016, Liu et al, 2015c, Nech and K-S, 2017, Yi et al, 2014), the proposed QMUL-SurvFace challenge is still relatively smaller. An impor-

tant future effort is therefore to expand the challenge for more effectively training deep FR models and further scaling up the open-set evaluations. We consider this one of our progressive future tasks.

Final Remark In the context that the web FR performance has been largely saturated on most existing challenges, this work presents at the right time a more challenging benchmark for continuously motivating innovative algorithm development especially dedicated for solving the practically critical surveillance FR problem whilst simultaneously benefiting other related research areas, such as transfer learning and open-set recognition. Overall, this study aims at calling for more collective research efforts from the entire community to contribute and accelerate the development of the challenging yet practically crucial surveillance FR problem.

Acknowledgements This work was partially supported by the Royal Society Newton Advanced Fellowship Programme (NA150459), InnovateUK Industrial Challenge Project on Developing and Commercialising Intelligent Video Analytics Solutions for Public Safety (98111-571149), Vision Semantics Ltd., and SeeQuestor Ltd.

References

Abate AF, Nappi M, Riccio D, Sabatino G (2007) 2d and 3d face recognition: A survey. Pattern Recognition Letters 28(14):1885–1906

Abiantun R, Savvides M, Kumar BV (2006) How low can you go? low resolution face recognition study using kernel correlation feature analysis on the frgcv2 dataset. In: Symposium on research at the Biometric consortium conference

Adini Y, Moses Y, Ullman S (1997) Face recognition: The problem of compensating for changes in illumination direction. IEEE Transactions on Pattern Analysis and Machine Intelligence 19(7):721–732

Ahonen T, Hadid A, Pietikainen M (2006) Face description with local binary patterns: Application to face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(12):2037–2041

Ahonen T, Rahtu E, Ojansivu V, Heikkila J (2008) Recognition of blurred faces using local phase quantization. In: IEEE International Conference on Pattern Recognition

Baker S, Kanade T (2000) Hallucinating faces. In: IEEE Conference on Automatic Face and Gesture Recognition

Baltieri D, Vezzani R, Cucchiara R (2011a) 3dpes: 3d people dataset for surveillance and forensics. In: ACM workshop on Human Gesture and Behavior Understanding

Baltieri D, Vezzani R, Cucchiara R (2011b) Sarc3d: a new 3d body model for people tracking and reidentification. International Conference on Image Analysis and Processing pp 197–206

Bansal A, Nanduri A, Castillo C, Ranjan R, Chellappa R (2016) Umdfaces: An annotated face dataset for training deep networks. arXiv preprint arXiv:161101484

Bansal A, Castillo C, Ranjan R, Chellappa R (2017) The do's and don'ts for cnn-based face verification. In: Workshop of IEEE International Conference on Computer Vision

Belhumeur PN, Hespanha JP, Kriegman DJ (1997) Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. IEEE Transactions on Pattern Analysis and Machine Intelligence 19(7):711–720

Bendale A, Boult T (2015) Towards open world recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 1893–1902

Bendale A, Boult TE (2016) Towards open set deep networks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 1563–1572

Berg T, Belhumeur PN (2013) Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 955–962

Best-Rowden L, Han H, Otto C, Klare BF, Jain AK (2014) Unconstrained face recognition: Identifying a person of interest from a media collection 9(12):2144–2157

Beveridge JR, Phillips PJ, Bolme DS, Draper BA, Givens GH, Lui YM, Teli MN, Zhang H, Scruggs WT, Bowyer KW, et al (2013) The challenge of face recognition from digital point-and-shoot cameras. In: IEEE International Conference on Biometrics: Theory, Applications, and Systems

Bilgazyev E, Efraty BA, Shah SK, Kakadiaris IA (2011) Sparse representation-based super resolution for face recognition at a distance. In: British Machine Vision Conference

Biswas S, Bowyer KW, Flynn PJ (2010) Multidimensional scaling for matching low-resolution facial images. In: IEEE International Conference on Biometrics: Theory, Applications, and Systems, pp 1–6

Biswas S, Bowyer KW, Flynn PJ (2012) Multidimensional scaling for matching low-resolution face images. IEEE Transactions on Pattern Analysis and Machine Intelligence 34(10):2019–2030

Bottou L (2010) Large-scale machine learning with stochastic gradient descent. In: COMPSTAT

Bruhn A, Weickert J, Schnörr C (2005) Lucas/kanade meets horn/schunck: Combining local and global op-

tic flow methods. International Journal of Computer Vision 61(3):211–231

Burgos-Artizzu XP, Perona P, Dollár P (2013) Robust face landmark estimation under occlusion. In: IEEE International Conference on Computer Vision, pp 1513–1520

Cao Q, Lin L, Shi Y, Liang X, Li G (2017a) Attention-aware face hallucination via deep reinforcement learning. In: IEEE Conference on Computer Vision and Pattern Recognition

Cao Q, Shen L, Xie W, Parkhi OM, Zisserman A (2017b) Vggface2: A dataset for recognising faces across pose and age. arXiv preprint arXiv:171008092

Cao X, Wipf D, Wen F, Duan G, Sun J (2013) A practical transfer learning algorithm for face verification. In: IEEE International Conference on Computer Vision

Cao X, Wei Y, Wen F, Sun J (2014) Face alignment by explicit shape regression. International Journal of Computer Vision 107(2):177–190

Chakrabarti A, Rajagopalan A, Chellappa R (2007) Super-resolution of face images using kernel pca-based prior. IEEE Transactions on Multimedia 9(4):888–892

Chang KI, Bowyer KW, Flynn PJ (2005) An evaluation of multimodal 2d+ 3d face biometrics. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(4):619–624

Chellappa R, Wilson CL, Sirohey S (1995) Human and machine recognition of faces: A survey. Proceedings of the IEEE 83(5):705–741

Chen D, Cao X, Wang L, Wen F, Sun J (2012) Bayesian face revisited: A joint formulation. European Conference on Computer Vision

Chen D, Cao X, Wen F, Sun J (2013) Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. In: IEEE Conference on Computer Vision and Pattern Recognition

Cheng DS, Cristani M, Stoppa M, Bazzani L, Murino V (2011) Custom pictorial structures for re-identification. In: British Machine Vision Conference

Choi JY, Ro YM, Plataniotis KN (2008) Feature subspace determination in video-based mismatched face recognition. In: IEEE Conference on Automatic Face and Gesture Recognition

Choi JY, Ro YM, Plataniotis KN (2009) Color face recognition for degraded face images 39(5):1217–1230

Das A, Chakraborty A, Roy-Chowdhury AK (2014) Consistent re-identification in a camera network. In: European Conference on Computer Vision

Daugman J (1997) Face and gesture recognition: Overview. IEEE Transactions on Pattern Analysis and Machine Intelligence 19(7):675–676

Dong C, Loy CC, He K, Tang X (2014) Learning a deep convolutional network for image super-resolution. In: European Conference on Computer Vision

Dong C, Loy CC, Tang X (2016) Accelerating the super-resolution convolutional neural network. In: European Conference on Computer Vision

Dong Q, Gong S, Zhu X (2017) Class rectification hard mining for imbalanced deep learning. In: IEEE International Conference on Computer Vision, pp 1851–1860

Electron E (2017) Easenelectron. `http://english.easen-electron.com/`

Ersotelos N, Dong F (2008) Building highly realistic facial modeling and animation: a survey. Image and Vision Computing 24(1):13–30

Farhadi A, Endres I, Hoiem D, Forsyth D (2009) Describing objects by their attributes. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 1778–1785

Fookes C, Lin F, Chandran V, Sridharan S (2012) Evaluation of image resolution and super-resolution on face recognition performance. Journal of Visual Communication and Image Representation 23(1):75–93

Frome A, Corrado GS, Shlens J, Bengio S, Dean J, Mikolov T, et al (2013) Devise: A deep visual-semantic embedding model. In: Advances in Neural Information Processing Systems, pp 2121–2129

Fu Y, Hospedales TM, Xiang T, Gong S (2014) Learning multimodal latent attributes. IEEE transactions on pattern analysis and machine intelligence 36(2):303–316

Fu Y, Xiang T, Jiang YG, Xue X, Sigal L, Gong S (2017) Recent advances in zero-shot recognition. IEEE Signal Processing Magazine

Gallagher AC, Chen T (2009) Understanding images of groups of people. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 256–263

Ganin Y, Lempitsky V (2015) Unsupervised domain adaptation by backpropagation. In: International Conference on Machine learning, pp 1180–1189

Gatys LA, Ecker AS, Bethge M (2016) Image style transfer using convolutional neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 2414–2423

Georghiades AS, Belhumeur PN, Kriegman DJ (2001) From few to many: Illumination cone models for face recognition under variable lighting and pose. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(6):643–660

Ghifary M, Kleijn WB, Zhang M, Balduzzi D, Li W (2016) Deep reconstruction-classification networks for unsupervised domain adaptation. In: European Conference on Computer Vision, pp 597–613

Gong S, Ong EJ, McKenna SJ (1998) Learning to associate faces across views in vector space of similarities to prototypes. In: British Machine Vision Conference, vol 1, pp 54–64

Gong S, MeKenna S, Psarrou A (2000) Dynamic Vision: From Images to Face Recognition. Imperial College Press, World Scientific

Gong S, Cristani M, Yan S, Loy CC (2014) Person re-identification. Springer

Gou M, Karanam S, Liu W, Camps O, Radke RJ (2017) Dukemtmc4reid: A large-scale multi-camera person re-identification dataset. In: Workshop of IEEE Conference on Computer Vision and Pattern Recognition

Gray D, Tao H (2008) Viewpoint invariant pedestrian recognition with an ensemble of localized features. European Conference on Computer Vision

Grgic M, Delac K, Grgic S (2011) Scface–surveillance cameras face database. Multimedia Tools and Applications 51(3):863–879

Gross R, Matthews I, Cohn J, Kanade T, Baker S (2010) Multi-pie. Image and Vision Computing 28(5):807–813

Grother P, Ngan M (2014) Face recognition vendor test (frvt): Performance of face identification algorithms. NIST Interagency Report 8009(5)

Grother P, Ngan M, Hanaoka K (2017) Face recognition vendor test (frvt) ongoing

Günther M, Hu P, Herrmann C, Chan CH, Jiang M, Yang S, Dhamija AR, Ramanan D, Beyerer J, Kittler J, et al (2017) Unconstrained face detection and open-set face recognition challenge. In: International Joint Conference on Biometrics

Gunturk BK, Batur AU, Altunbasak Y, Hayes MH, Mersereau RM (2003) Eigenface-domain super-resolution for face recognition. IEEE Transactions on Image Processing 12(5):597–606

Guo Y, Zhang L, Hu Y, He X, Gao J (2016) Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In: European Conference on Computer Vision

Hadsell R, Chopra S, LeCun Y (2006) Dimensionality reduction by learning an invariant mapping. In: IEEE Conference on Computer Vision and Pattern Recognition

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition

He X, Cai D, Yan S, Zhang HJ (2005) Neighborhood preserving embedding. In: IEEE International Conference on Computer Vision

Helbing D, Molnar P (1995) Social force model for pedestrian dynamics. Physical review E 51(5):4282

Hennings-Yeomans PH, Baker S, Kumar BV (2008) Simultaneous super-resolution and feature extraction for recognition of low-resolution faces. In: IEEE Conference on Computer Vision and Pattern Recognition

Hu P, Ramanan D (2016) Finding tiny faces. arXiv e-print

Huang C, Li Y, Change Loy C, Tang X (2016) Learning deep representation for imbalanced classification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 5375–5384

Huang GB, Ramesh M, Berg T, Learned-Miller E (2007) Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Tech. rep., University of Massachusetts

Huang H, He H (2011) Super-resolution method for face recognition using nonlinear mappings on coherent features. IEEE Transactions on Neural Networks 22(1):121–130

Hughes RL (2002) A continuum theory for the flow of pedestrians. Transportation Research Part B: Methodological 36(6):507–535

Jia K, Gong S (2005) Multi-modal tensor face for simultaneous super-resolution and recognition. In: IEEE International Conference on Computer Vision

Jia K, Gong S (2008) Generalized face super-resolution. IEEE Transactions on Image Processing 17(6):873–886

Jiang J, Hu R, Wang Z, Han Z, Ma J (2016) Facial image hallucination through coupled-layer neighbor embedding. IEEE Transactions on Circuits and Systems for Video Technology 26(9):1674–1684

Jin Y, Bouganis CS (2015) Robust multi-image based blind face hallucination. In: IEEE Conference on Computer Vision and Pattern Recognition

K-S I, Seitz SM, Miller D, Brossard E (2016) The megaface benchmark: 1 million faces for recognition at scale. In: IEEE Conference on Computer Vision and Pattern Recognition

Kawanishi Y, Wu Y, Mukunoki M, Minoh M (2014) Shinpuhkan2014: A multi-camera pedestrian dataset for tracking people across multiple cameras. In: Korea-Japan Joint Workshop on Frontiers of Computer Vision

Kim J, Kwon Lee J, Mu Lee K (2016a) Accurate image super-resolution using very deep convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition

Kim J, Kwon Lee J, Mu Lee K (2016b) Deeply-recursive convolutional network for image super-resolution. In: IEEE Conference on Computer Vision and Pattern Recognition

Kim KI, Kwon Y (2010) Single-image super-resolution using sparse regression and natural image prior.

IEEE Transactions on Pattern Analysis and Machine Intelligence 32(6):1127–1133

Klare BF, Klein B, Taborsky E, Blanton A, Cheney J, Allen K, Grother P, Mah A, Jain AK (2015) Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In: IEEE Conference on Computer Vision and Pattern Recognition

Kong SG, Heo J, Abidi BR, Paik J, Abidi MA (2005) Recent advances in visual and infrared face recognitiona review. Computer Vision and Image Understanding 97(1):103–135

Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems

Kumar N, Berg AC, Belhumeur PN, Nayar SK (2009) Attribute and simile classifiers for face verification. In: IEEE International Conference on Computer Vision, pp 365–372

Lai WS, Huang JB, Ahuja N, Yang MH (2017) Deep laplacian pyramid networks for fast and accurate super-resolution. In: IEEE Conference on Computer Vision and Pattern Recognition

Le V, Brandt J, Lin Z, Bourdev L, Huang TS (2012) Interactive facial feature localization. In: European Conference on Computer Vision, pp 679–692

Ledig C, Theis L, Huszár F, Caballero J, Cunningham A, Acosta A, Aitken A, Tejani A, Totz J, Wang Z, et al (2016) Photo-realistic single image super-resolution using a generative adversarial network. arXiv e-print

Lee CY, Xie S, Gallagher P, Zhang Z, Tu Z (2015) Deeply-supervised nets. In: Artificial Intelligence and Statistics, pp 562–570

Lei Z, Ahonen T, Pietikäinen M, Li SZ (2011) Local frequency descriptor for low-resolution face recognition. In: IEEE Conference on Automatic Face and Gesture Recognition

Li B, Chang H, Shan S, Chen X, Gao W (2008) Hallucinating facial images and features. In: IEEE International Conference on Pattern Recognition

Li B, Chang H, Shan S, Chen X (2009) Coupled metric learning for face recognition with degraded images. Advances in Machine Learning

Li B, Chang H, Shan S, Chen X (2010) Low-resolution face recognition via coupled locality preserving mappings 17(1):20–23

Li C, Wand M (2016) Precomputed real-time texture synthesis with markovian generative adversarial networks. In: European Conference on Computer Vision, pp 702–716

Li S, Jain A (2011) Handbook of Face Recognition. Springer

Li W, Zhao R, Xiao T, Wang X (2014) Deepreid: Deep filter pairing neural network for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition

Li Y, Fang C, Yang J, Wang Z, Lu X, Yang MH (2017) Diversified texture synthesis with feed-forward networks. In: IEEE Conference on Computer Vision and Pattern Recognition

Liao S, Lei Z, Yi D, Li SZ (2014) A benchmark study of large-scale unconstrained face recognition. In: International Joint Conference on Biometrics

Liu C, Wechsler H (2002) Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. IEEE Transactions on Image Processing 11(4):467–476

Liu C, Shum HY, Freeman WT (2007) Face hallucination: Theory and practice. International Journal of Computer Vision 75(1):115

Liu J, Deng Y, Bai T, Wei Z, Huang C (2015a) Targeting ultimate accuracy: Face recognition via deep embedding. arXiv e-print

Liu K, Ma B, Zhang W, Huang R (2015b) A spatio-temporal appearance representation for viceo-based pedestrian re-identification. In: IEEE International Conference on Computer Vision

Liu TY, et al (2009) Learning to rank for information retrieval. Foundations and Trends® in Information Retrieval 3(3):225–331

Liu W, Lin D, Tang X (2005) Hallucinating faces: Tensorpatch super-resolution and coupled residue compensation. In: IEEE Conference on Computer Vision and Pattern Recognition

Liu W, Wen Y, Yu Z, Li M, Raj B, Song L (2017) Sphereface: Deep hypersphere embedding for face recognition. arXiv e-print

Liu Z, Luo P, Wang X, Tang X (2015c) Deep learning face attributes in the wild. In: IEEE International Conference on Computer Vision

Loy CC, Xiang T, Gong S (2009) Multi-camera activity correlation analysis. In: IEEE Conference on Computer Vision and Pattern Recognition

Lu C, Tang X (2015) Surpassing human-level face verification performance on lfw with gaussianface. In: AAAI Conference on Artificial Intelligence

Maltoni D, Maio D, Jain A, Prabhakar S (2009) Handbook of fingerprint recognition. Springer Science & Business Media

Manyam OK, Kumar N, Belhumeur P, Kriegman D (2011) Two faces are better than one: Face recognition in group photographs. In: International Joint Conference on Biometrics, pp 1–8

Martinel N, Micheloni C (2012) Re-identify people in wide area camera network. In: Workshop of IEEE

Conference on Computer Vision and Pattern Recognition

Masi I, Rawls S, Medioni G, Natarajan P (2016) Pose-aware face recognition in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition

Messer K, Matas J, Kittler J, Luettin J, Maitre G (1999) Xm2vtsdb: The extended m2vts database. In: International Conference on Audio and Video-based Biometric Person Authentication

Nech A, K-S I (2017) Level playing field for million scale face recognition. In: IEEE Conference on Computer Vision and Pattern Recognition

Ng HW, Winkler S (2014) A data-driven approach to cleaning large face datasets. In: IEEE International Conference on Image Processing

Ortiz EG, Becker BC (2014) Face recognition for web-scale datasets. Computer Vision and Image Understanding 118:153–170

Pan SJ, Yang Q (2010) A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering 22(10):1345–1359

Parikh D, Grauman K (2011) Relative attributes. In: IEEE International Conference on Computer Vision, pp 503–510

Parkhi OM, Vedaldi A, Zisserman A (2015) Deep face recognition. In: British Machine Vision Conference

Phillips PJ, Moon H, Rizvi SA, Rauss PJ (2000) The feret evaluation methodology for face-recognition algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(10):1090–1104

Phillips PJ, Flynn PJ, Scruggs T, Bowyer KW, Chang J, Hoffman K, Marques J, Min J, Worek W (2005) Overview of the face recognition grand challenge. In: IEEE Conference on Computer Vision and Pattern Recognition

Phillips PJ, Scruggs WT, O'Toole AJ, Flynn PJ, Bowyer KW, Schott CL, Sharpe M (2010) Frvt 2006 and ice 2006 large-scale experimental results. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(5):831–846

Prado KS, Roman NT, Silva VF, Bernardes Jr JL, Digiampietri LA, Ortega EM, Lima CA, Cura LM, Antunes MM (2016) Automatic facial recognition: A systematic review on the problem of light variation

Ren CX, Dai DQ, Yan H (2012) Coupled kernel embedding for low-resolution face image recognition. IEEE Transactions on Image Processing 21(8):3770–3783

Ricanek K, Tesafaye T (2006) Morph: A longitudinal image database of normal adult age-progression. In: IEEE Conference on Automatic Face and Gesture Recognition

Roth PM, Hirzer M, Köstinger M, Beleznai C, Bischof H (2014) Mahalanobis Distance Learning for Person Re-Identification. In: Person Re-Identification, Springer, pp 247–267

Rumelhart DE, Hinton GE, Williams RJ, et al (1988) Learning representations by back-propagating errors. Cognitive Modeling 5(3):1

Saenko K, Kulis B, Fritz M, Darrell T (2010) Adapting visual category models to new domains. European Conference on Computer Vision pp 213–226

Samal A, Iyengar PA (1992) Automatic recognition and analysis of human faces and facial expressions: A survey. Pattern Recognition 25(1):65–77

Samaria FS, Harter AC (1994) Parameterisation of a stochastic model for human face identification. In: IEEE Workshop on Applications of Computer Vision

Sarkar S, Phillips PJ, Liu Z, Vega IR, Grother P, Bowyer KW (2005) The humanid gait challenge problem: Data sets, performance, and analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(2):162–177

Schroff F, Kalenichenko D, Philbin J (2015) Facenet: A unified embedding for face recognition and clustering. In: IEEE Conference on Computer Vision and Pattern Recognition

Schwartz W, Davis L (2009) Learning discriminative appearance-based models using partial least squares. In: Brazilian Symposium on Computer Graphics and Image Processing

Sengupta S, Chen JC, Castillo C, Patel VM, Chellappa R, Jacobs DW (2016) Frontal to profile face verification in the wild. In: IEEE Winter Conference on Applications of Computer Vision

Shekhar S, Patel VM, Chellappa R (2011) Synthesis-based recognition of low resolution faces. In: International Joint Conference on Biometrics

Sim T, Baker S, Bsat M (2002) The cmu pose, illumination, and expression (pie) database. In: IEEE Conference on Automatic Face and Gesture Recognition, pp 53–58

Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations

Song F, Tan X, Chen S (2014) Exploiting relationship between attributes for improved face verification. Computer Vision and Image Understanding 122:143–154

Song G, Leng B, Liu Y, Hetang C, Cai S (2017) Region-based quality estimation network for large-scale person re-identification. arXiv preprint arXiv:171108766

Sun Y, Chen Y, Wang X, Tang X (2014a) Deep learning face representation by joint identification-verification. In: Advances in Neural Information Processing Systems

Sun Y, Wang X, Tang X (2014b) Deep learning face representation from predicting 10,000 classes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1891–1898

Sun Y, Wang X, Tang X (2014c) Deep learning face representation from predicting 10,000 classes. In: IEEE Conference on Computer Vision and Pattern Recognition

Sun Y, Liang D, Wang X, Tang X (2015a) Deepid3: Face recognition with very deep neural networks. arXiv e-print

Sun Y, Wang X, Tang X (2015b) Deeply learned face representations are sparse, selective, and robust. In: IEEE Conference on Computer Vision and Pattern Recognition

Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition

Szegedy C, Ioffe S, Vanhoucke V, Alemi AA (2017) Inception-v4, inception-resnet and the impact of residual connections on learning. In: AAAI Conference on Artificial Intelligence

Tai Y, Yang J, Liu X (2017) Image super-resolution via deep recursive residual network. In: IEEE Conference on Computer Vision and Pattern Recognition

Taigman Y, Yang M, Ranzato M, Wolf L (2014) Deepface: Closing the gap to human-level performance in face verification. In: IEEE Conference on Computer Vision and Pattern Recognition

Taigman Y, Yang M, Ranzato M, Wolf L (2015) Web-scale training for face identification. In: IEEE Conference on Computer Vision and Pattern Recognition

Tan X, Triggs B (2010) Enhanced local texture feature sets for face recognition under difficult lighting conditions. IEEE Transactions on Image Processing 19(6):1635–1650

Tan X, Chen S, Zhou ZH, Zhang F (2006) Face recognition from a single image per person: A survey. Pattern Recognition 39(9):1725–1745

Tencent (2017) Youtu lab, tencent. `http://bestimage.qq.com/`

Turk M, Pentland A (1991) Eigenfaces for recognition. Journal of Cognitive Neuroscience 3(1):71–86

Tzeng E, Hoffman J, Darrell T, Saenko K (2015) Simultaneous deep transfer across domains and tasks. In: IEEE International Conference on Computer Vision, pp 4068–4076

Wang D, Otto C, Jain AK (2016a) Face search at scale. IEEE Transactions on Pattern Analysis and Machine Intelligence

Wang T, Gong S, Zhu X, Wang S (2014a) Person re-identification by video ranking. In: European Conference on Computer Vision

Wang X, Tang X (2003) Face hallucination and recognition. In: International Conference on Audio-and Video-Based Biometric Person Authentication

Wang X, Tang X (2005) Hallucinating face by eigentransformation 35(3):425–434

Wang Z, Miao Z (2008) Scale invariant face recognition using probabilistic similarity measure. In: IEEE International Conference on Pattern Recognition

Wang Z, Miao Z, Wu QJ, Wan Y, Tang Z (2014b) Low-resolution face recognition: a review. Image and Vision Computing 30(4):359–386

Wang Z, Chang S, Yang Y, Liu D, Huang TS (2016b) Studying very low resolution recognition using deep networks. In: IEEE Conference on Computer Vision and Pattern Recognition

Wechsler H (2009) Reliable face recognition methods: system design, implementation and evaluation, vol 7. Springer Science & Business Media

Wechsler H, Philips J, Bruce V, Fogelman-Soulie F, Huang T (1998) Face Recognition: From Theory to Applications. Springer-Verlag

Wechsler H, Phillips JP, Bruce V, Soulie FF, Huang TS (2012) Face recognition: From theory to applications, vol 163. Springer Science & Business Media

Wen Y, Zhang K, Li Z, Qiao Y (2016) A discriminative feature learning approach for deep face recognition. In: European Conference on Computer Vision

Whitelam C, Taborsky E, Blanton A, Maze B, Adams J, Miller T, Kalka N, Jain AK, Duncan JA, Allen K, et al (2017) Iarpa janus benchmark-b face dataset. In: CVPR Workshop on Biometrics

Wolf L, Levy N (2013) The svm-minus similarity score for video face recognition. In: IEEE Conference on Computer Vision and Pattern Recognition

Wolf L, Hassner T, Maoz I (2011) Face recognition in unconstrained videos with matched background similarity. In: IEEE Conference on Computer Vision and Pattern Recognition

Wong Y, Sanderson C, Mau S, Lovell BC (2010) Dynamic amelioration of resolution mismatches for local feature based identity inference. In: IEEE International Conference on Pattern Recognition

Wright J, Yang AY, Ganesh A, Sastry SS, Ma Y (2009) Robust face recognition via sparse representation. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(2):210–227

Xiao T, Li S, Wang B, Lin L, Wang X (2016) End-to-end deep learning for person search. arXiv e-print

Xie S, Tu Z (2015) Holistically-nested edge detection. In: IEEE International Conference on Computer Vision, pp 1395–1403

Xu X, Liu W, Li L (2013) Face hallucination: How much it can improve face recognition. In: Australian Conference

Yang CY, Ma C, Yang MH (2014) Single-image super-resolution: A benchmark. In: European Conference on Computer Vision

Yang J, Wright J, Huang TS, Ma Y (2010) Image super-resolution via sparse representation. IEEE Transactions on Image Processing 19(11):2861–2873

Yi D, Lei Z, Liao S, Li SZ (2014) Learning face representation from scratch. arXiv e-print

Yu X, Porikli F (2016) Ultra-resolving face images by discriminative generative networks. In: European Conference on Computer Vision

Yu X, Porikli F (2017) Hallucinating very low-resolution unaligned and noisy face images by transformative discriminative autoencoders. In: IEEE Conference on Computer Vision and Pattern Recognition

Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks. In: European Conference on Computer Vision

Zhang B, Shan S, Chen X, Gao W (2007) Histogram of gabor phase patterns (hgpp): A novel object representation approach for face recognition. IEEE Transactions on Image Processing 16(1):57–68

Zhang K, Zhang Z, Li Z, Qiao Y (2016) Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters 23(10):1499–1503

Zhang Z, Luo P, Loy CC, Tang X (2014) Facial landmark detection by deep multi-task learning. In: European Conference on Computer Vision, pp 94–108

Zhao W, Chellappa R (2011) Face Processing: Advanced modeling and methods. Academic Press

Zhao W, Chellappa R, Phillips PJ, Rosenfeld A (2003) Face recognition: A literature survey. ACM Computing Surveys 35(4):399–458

Zheng L, Shen L, Tian L, Wang S, Wang J, Tian Q (2015) Scalable person re-identification: A benchmark. In: IEEE International Conference on Computer Vision

Zheng WS, Gong S, Xiang T (2009) Associating groups of people. In: British Machine Vision Conference, vol 2

Zheng WS, Gong S, Xiang T (2016) Towards open-world person re-identification by one-shot group-based verification. IEEE Transactions on Pattern Analysis and Machine Intelligence 38(3):591–606

Zhou C, Zhang Z, Yi D, Lei Z, Li SZ (2011) Low-resolution face recognition via simultaneous discriminant analysis. In: International Joint Conference on Biometrics

Zhou SK, Chellappa R, Zhao W (2006) Unconstrained face recognition, vol 5. Springer Science & Business Media

Zhu JY, Park T, Isola P, Efros AA (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: IEEE International Conference on Computer Vision

Zhu S, Li C, Change Loy C, Tang X (2015) Face alignment by coarse-to-fine shape searching. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 4998–5006

Zhu S, Liu S, Loy CC, Tang X (2016) Deep cascaded bi-network for face hallucination. In: European Conference on Computer Vision

Zou WW, Yuen PC (2012) Very low resolution face recognition problem. IEEE Transactions on Image Processing 21(1):327–340

Zou X, Kittler J, Messer K (2007) Illumination invariant face recognition: A survey. In: IEEE International Conference on Biometrics: Theory, Applications, and Systems, pp 1–8