Person Recognition in Low-Quality Imagery

Zhiyi Cheng

Submitted in partial fulfilment of the requirement for the degree of *Doctor of Philosophy*

School of Electronic Engineering and Computer Science

Queen Mary, University of London

2021

In memory of my beloved father.

Abstract

Person recognition aims to recognise and track the same individuals over space and time with subtle identity class information in automatically detected person images captured by unconstrained camera views. There are multi-source visual biometrical cues for person identity recognition. Specifically, compared to other widely-used cues that tend to easily change over time and space, the facial appearance is considered as a more reliable non-intrusive visual cue. Person recognition, especially the person face recognition, enables a wide range of practical applications, ranging from law enforcement and information security to business, entertainment and e-commerce. However, person recognition under realistic application scenarios remains significantly challenging, mainly due to the usual low resolutions (LR) of the images captured by low-quality cameras with unconstrained distances between cameras and people. Compared to the high-resolution (HR) images, the LR person images contain much less fine-grained discriminative details for robust identity recognition. To tackle the challenge of person recognition on low-resolution imagery data, one effective approach is to utilise the super resolution (SR) methods to recover or enhance the image details that are beneficial for identity recognition. However, this thesis reveals that conventional SR models suffer from significant performance drop when applied to low-quality LR person images, especially the natively captured surveillance facial images. Moreover, as the SR and identity recognition models advance independently, direct super resolution is less compatible with identity recognition, and hence has minor benefit or even negative effect for low-resolution person recognition.

To tackle the above problems, this thesis explores person recognition methods with improved generalisation ability to realistic low-quality person images, by adopting dedicated super-resolution algorithms. More specifically, this thesis addresses the issues for person face recognition and body recognition in low-resolution images as follows:

Chapter 3 Whilst recent person face recognition techniques have made significant progress on recognising constrained high-resolution web images, the same cannot be said on natively unconstrained low-resolution images at large scales. This chapter examines systematically this

under-studied person face recognition problem, and introduce a novel Complement Super-Resolution and Identity (CSRI) joint deep learning method with a unified end-to-end network architecture. The proposed learning mechanism is dedicated to overcome the inherent challenge of genuine low-resolution, concerning with the absence of HR facial images coupled with native LR faces, typically required for optimising image super-resolution models. This is realised by transferring the super-resolving knowledge from good-quality HR web images to the genuine LR facial data subject to the face identity label constraints of native LR faces in every mini-batch training. This chapter further constructs a new large-scale dataset TinyFace of native unconstrained low-resolution face images from selected public datasets. The extensive experiments show that there is a significant gap between the reported person face recognition performances on popular benchmarks and the results on TinyFace, and the advantages of the proposed CSRI over a variety of state-of-the-art face recognition and super-resolution deep models on solving this largely ignored person face recognition scenario. However, the lack of supervision in pixel space leads to the low-fidelity super-resolved images. which may hinder the further downstream facial analysis applications.

Chapter 4 Although with a more advanced joint-learning scheme for person face recognition by super resolution (introduced in Chapter 3), by no-means one can claim that the proposed method solves the real-world low-resolution face recognition problem, which remains a significantly challenging task. In terms of human understanding, when people are faced with a challenging face identity recognition task, they often make decisions by selecting discriminative facial features. If a recognition model can be optimised with results that can be explained in a human-understandable way, such an interpretable model may have the potential to shed light on discriminative facial features selection for better identity recognition. To achieve this, recognising faces from high-fidelity super-resolved outputs could be a viable approach. However, existing facial super-resolution methods focus mostly on improving "artificially down-sampled" low-resolution (LR) imagery. Such SR models, although strong at handling artificial LR images, often suffer from significant performance drop on genuine LR test data. Previous unsupervised domain adaptation (UDA) methods address this issue by training a model using unpaired genuine LR and HR data as well as cycle consistency loss formulation. However, this renders the model overstretched with two tasks: consistifying the visual characteristics and enhancing the image resolution. Importantly, this makes the end-to-end model training ineffective due to the difficulty of back-propagating gradients through two concatenated CNNs. To solve this problem, in this chapter, a method that joins the advantages of conventional SR and UDA models is formulated. Specifically, the optimisations for characteristics consistifying and image super-resolving are separated and controlled by introducing Characteristic Regularisation (CR) between them. This task split makes the model training more effective and computationally tractable, and enables the high-fidelity super resolution process on genuine low-resolution faces.

Chapter 5 Although the facial appearance is a more reliable visual cue for person recognition, it is often challenging or even impossible to detect the facial region in images captured by unconstrained low-quality cameras, where the faces can be of extreme poses, blur, distortion, or even invisible in the human back-view images. In such cases, the person body recognition is an important aspect for identity recognition and tracking. However, person images captured by unconstrained surveillance cameras often have low resolutions (LR). This causes the resolution mismatch problem when matched against the high-resolution (HR) gallery images, negatively affecting the performance of person body recognition. An effective approach is to leverage image super-resolution (SR) along with body recognition in a joint learning manner. However, this scheme is limited due to dramatically more difficult gradients backpropagation during training. This chapter introduces a novel model training regularisation method, called Inter-Task Association Critic (INTACT), to address this fundamental problem. Specifically, INTACT discovers the underlying association knowledge between image SR and person body recognition, and leverages it as an extra learning constraint for enhancing the compatibility of SR model with person body recognition in HR image space. This is realised by parameterising the association constraint, which can be automatically learned from the training data. Extensive experiments validate the superiority of INTACT over the state-of-the-art approaches on the cross-resolution person body recognition task using five standard datasets.

Chapter 6 draws conclusions and suggests future works on open questions arising from the studies of this thesis.

Declaration

I hereby declare that this thesis has been composed by myself and that it describes my own work. It has not been submitted, either in the same or different form, to this or any other university for a degree. All verbatim extracts are distinguished by quotation marks, and all sources of information have been acknowledged. Some works have been published or under review:

Chapter 3

1. Zhiyi Cheng, Xiatian Zhu, and Shaogang Gong. *Face Re-Identification Challenge: Are Face Recognition Models Good Enough?* Pattern Recognition, Vol. 107, 107422, 2020. (**PR**)

2. Zhiyi Cheng, Xiatian Zhu, and Shaogang Gong. *Low-resolution face recognition*. Asian Conference on Computer Vision, Perth, Australia, 2018. (**ACCV**)

Chapter 4

1. Zhiyi Cheng, Xiatian Zhu, and Shaogang Gong. *Characteristic Regularisation for Super-Resolving Face Images*. IEEE Winter Conference on Applications of Computer Vision, Snowmass, Colorado, USA, 2020. (**WACV**)

Chapter 5

1. Zhiyi Cheng, Qi Dong, Shaogang Gong, and Xiatian Zhu. *Inter-Task Association Critic for Cross-Resolution Person Re-Identification*. IEEE Conference on Computer Vision and Pattern Recognition, Seattle, Washington, USA, 2020. (**CVPR**)

Acknowledgements

The first and the most, I would like to express my deep gratitude to my supervisor Prof. Shaogang Gong for his enthusiastic and continued supervision, encouragement and support in the past three years. Meanwhile, I would like to thank Dr. Xiatian Zhu for his patient guidance and invaluable advices. They are always the most supportive and would like to offer the most inspirational suggestions over the past three years. Without their support and guidance, this thesis would not have been accomplished. It is their guidance that leads me towards a research career.

I am grateful to Prof. Tao Xiang and Dr. Miles Hansard for being my second supervisor and independent accessor, and provide inspiring comments and suggestions throughout my PhD study.

I would also like to thank the friends and collaborators in the QMUL Computer Vision Group (mostly in chronological order), including Yanbei Chen, Qian Yu, Qi Dong, Xu Lan, Jingya Wang, Conghui Hu, Jifei Song, Hanxiao Wang, Feng Liu, Kaiyue Pang, Xiaobin Chang, Da Li, Ying Zhang, Zimo Liu, Umar Riaz Muhammad, Aytac Kanaci, Hang Su, Elyor Kodirov, Weihong Li, Yali Zhao, Li Zhang, Wei Li, Peng Xu, Kaiyang Zhou, Tianyuan Yu, Jiabo Huang, Guile Wu, Minxian Li, Anran Qi, Pan Li, Qingze Yin and Guanan Wang. I am grateful to Prof. Chen Change Loy from NTU, for his warm encouragement during the difficult time through my PhD study. I give sincere thanks to all their friendship and help to support my research study over the past three years.

Most of all, I would like to express my deepest gratitude to my family members, including my parents and my brother for their enduring love and endless support. Specifically, this thesis is dedicated to my father, who passed away last year after his bravest fight with disease. His love and kindness will always encourage me.

Contents

List of Figures

Chapter 1

Introduction

With the rapid expansion of surveillance multi-camera systems around the world, associating people over space and time becomes an increasingly significant capability for a wide range of applications such as public safety, law enforcement and forensic search [57]. Person identity recognition aims to track the same individuals by the subtle class information in the person images detected under the unconstrained scenarios. To match the person identities over large distributed space and time is inherently challenging, and to extract reliable visual cues as identity information from diverse unconstrained camera views is essential for robust person recognition. Given its wide-range potential applications in real-world scenarios, person recognition has drawn growing attention from both academic and industrial researchers. Among the existing visual biometrics for person identity recognition, such as whole-body [58, 107, 114, 162, 172, 206], iris [148], gait [158], and fingerprint [128], facial appearance is considered as one of the most convenient and most reliable non-intrusive visual cues. This is due to one fact that faces, provided they are visible in captured images, are more stable cues for long-term tracking and tracing, whereas other visual appearances, e.g. clothes for whole-body cues [58], are easier to change over space and time.

Person face recognition (FR) has been extensively studied with significant advance in the literature, and FR based commercial products are increasingly appearing in our daily life, e.g. web photo-album and online e-payment. However, current FR methods generalise poorly in recognising faces in realistic noisy and low-quality images captured by unconstrained wide-field surveillance cameras, which makes the techniques far away from being satisfactory in real-world applications. This is because of the lack of sufficient detailed information for identity recognition in the low-quality inputs, as compared to the high-resolution, high-quality images (Fig. 1.1).

While being critical for public safety and law enforcement applications, person face recognition in low-quality images is significantly under-studied in comparison to the high-resolution face recognition task. It is shown in experimental studies [40] that (1) the state-of-the-art FR models trained on large scale high-quality benchmark datasets generalise poorly to the low-quality face recognition task on native low-quality surveillance facial images; (2) the performance of face recognition on artificially synthesised low-resolution images does not well reflect the true challenges of native surveillance facial images in system deployments; (3) the image super-resolution models suffer from the lack of pixel-aligned low- and high-resolution native image pairs which are necessary for model training. To facilitate solving the aforementioned problems and limitations, this thesis focuses on exploring dedicated person face recognition models in low-quality images captured by realistic unconstrained cameras, especially characterised by native low resolution.

Moreover, although the facial cue plays an essential role in person recognition, in many real-world scenarios, faces are invisible in the cases of pedestrian back-view images or occlusion. In such cases, it is much easier to access the whole-body appearance as visual cues for person recognition. However, in terms of the person body recognition, due to unconstrained distances between cameras and pedestrians, person images are often captured at various resolutions, while most existing methods assume that the probe and gallery images have similar and sufficiently high resolutions. This resolution mismatch issue brings about significant challenges to person body recognition. Therefore, towards person recognition techniques in real-world applications, in addition to person face recognition, this thesis also explores the problem of cross-resolution person body recognition.

1.1 Person Face Recognition in Low-quality Images

1.1.1 Low-Resolution Person Face Recognition by Super Resolution

Face recognition (FR) models have made significant progress on constrained good-quality images, with reported 99.63% accuracy (1:1 verification) on the LFW benchmark [80] and 99.087% rank-1 rate (1:N identification with 1,000,000 distractors in the gallery) on the MegaFace challenge [91]. Surprisingly, this thesis shows systematically that person FR remains a significant

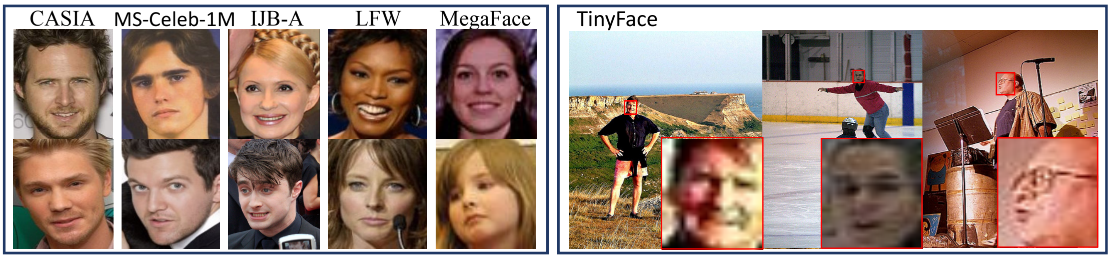

Figure 1.1: Examples of (Left) **constrained high-resolution** face images from five popular benchmarking FR datasets, and (Right) **native unconstrained low-resolution** face images typically captured in natural scenes.

challenge on *natively unconstrained low-resolution (LR)* images – *not artificially* down-sampled from high-resolution (HR) images, as typically captured in surveillance videos [40, 240] and unconstrained (unposed) snapshots from a wide field of view at distance [225, 213]. In particular, when tested against native low-resolution face images from a newly constructed tiny face dataset, this thesis reveals that the performances of current state-of-the-art deep learning FR models degrade significantly. Fig. 1.1 provides a visual comparison of the low-quality realistic images to the high-quality web photos.

In general, unconstrained low-resolution FR (LRFR) is severely under-studied versus many FR models tested on popular benchmarks of HR images, mostly captured either under constrained viewing conditions or from "posed" photoshoots including passport photo verification for airport immigration control and identity check in e-banking. Another obstacle for enabling more studies on LRFR is the lack of large scale *native LR* face image data both for model training and testing, rather than artificially down-sampled synthetic data from HR images. To collect sufficient data for deep learning, it requires to process a large amount of public domain (e.g. from the web) video and image data generated from a wide range of sources such as social-media, e.g. the MegaFace dataset [91, 137]. So far, this has only been available for HR and good quality (constrained) web face images, e.g. widely distributed celebrity images [80, 144, 123].

One potentially effective approach for achieving a robust low-resolution person face recognition model is to adopt image super-resolution (SR). Designed for recovering the finer details when super-resolving low-resolution images, existing image super-resolution methods may be beneficial for the low-resolution person recognition problem. However, as at large, the image recognition and SR studies advance independently, this thesis discovers through extensive experiments that contemporary SR deep learning models bring about very marginal performance benefit or even negative effect on person face recognition. The plausible reasons are mainly twofolds. **(1)** The inherent lacking of paired HR and their corresponding *native* LR images. Existing methods that perform well [45, 92, 99, 103, 238] all assume the provision of pixelwise-aligned LR and HR image pairs for model training. They are *not* applicable to learning for super-resolving *unpaired* native facial images, i.e., the realistic facial images for person recognition. **(2)** A significant domain gap due to different imaging noise characteristics between native low-resolution facial images and high-resolution web face photo-shoots. One may construct a SR model trained on well-conditioned HR web images with perfectly aligned corresponding synthesised (e.g. bicubic downsampling) LR images. However, a model trained by such *synthetic* LR images do not capture the unknown and significantly different imaging noise and artifacts inherent to the *native* LR images, e.g. sensor noise, compression, non-ideal point spread function, among other aliasing effects. This domain transfer discrepancy between the training data from one domain (source) and the test data from a very different domain (target) causes inherent model limitations for poor performance generalisation among existing SR algorithms. To solve this problem, one potential solution is to adopt the domain-adaptation strategy with the auxiliary data of artificial down-sampled web faces for native facial image super-resolution.

This thesis investigates the problem of super resolution for improving the performance of native low-resolution person face recognition. Extensive experiments show that existing SR models significantly degrade when applied to native low-resolution faces and can not improve the identity recognisability. To solve the aforementioned problem, a novel Super-Resolution and Identity joint learning approach to face recognition in natively LR images is proposed, with a unified deep learning network architecture. It is designed to improve the model generalisation for low-resolution face recognition tasks by enhancing the compatibility of image super-resolving and identity recognition. Compared to directly applying super-resolution algorithms to improve image details without jointly optimising for face discrimination, this method has been shown to be effective in reducing the negative effect of noisy fidelity for the low-resolution face recognition task. To overcome the inherent challenge of native low-resolution face recognition concerning with the absence of HR facial images coupled with native LR faces, typically required for optimising image super-resolution models, a Complement Super-Resolution learning mechanism is introduced. This is realised by transferring the super-resolving knowledge from good-quality HR web images to the natively LR facial data subject to the face identity label constraints of native

LR faces in every mini-batch training.

Given the absence of pixel-aligned high-resolution counterparts as strong supervision signal, supervised by the high-level facial identity constraint instead, the proposed joint-learning model design is effective to improve the compatibility of the SR module to identity recognition. However, it is shown in the qualitative results that the super-resolved outputs from the low-resolution faces fail to achieve the visual fidelity expected in the high-resolution pixel space. It hinders the further downstream facial analysis applications by face super resolution, such as expression, emotion, age, among other attributes analysis. Moreover, an interpretable recognition model is beneficial to discover more discriminative facial features for better identity recognition. To achieve this, a high-fidelity super resolution model is needed.

1.1.2 Interpretable Low-Resolution Face Recognition

Although increasing attention is drawn to the low-resolution person face recognition, to recognise identities from realistic low-resolution imaging data remains a significantly challenging task. This thesis proposes to adopt the super-resolution algorithms for identity information recovery from low-resolution images before inputting to a recognition model (Sec. 1.1.1). However, given the absence of HR facial images coupled with native LR faces, the lack of supervision in pixel space for the SR models leads to the low-fidelity resolved images.

Given that the pixel-wise alignment is unavailable, this thesis proposes to further constrain the image-space domain adaptation for SR model, by aligning the imaging characteristics of the super-resolved faces to natural high-resolution ones.

More specifically, given the lack of pixel-aligned LR and HR image pairs as the supervision signal for a SR model training, in the literature, unsupervised domain adaptation (UDA) methods are possible solutions considering genuine LR and natural HR images as two different domains. UDA techniques have achieved remarkable success [237, 76, 181, 135, 17, 215, 95, 120]. A representative modelling idea is to exploit cycle consistency loss functions between two unpaired domains (Fig 1.2(b)) [237, 215, 95]. A CNN is used to map an image from one domain to the other, which is further mapped back by another CNN. With such an encoder-decoder like architecture, one can form a reconstruction loss *jointly* for both CNN models *without* the need for paired images in each domain. The two CNN models can be trained end-to-end, inputting an image and outputting a reconstructed image per domain. This idea has been attempted in [20] for super-resolving genuine LR facial imagery.

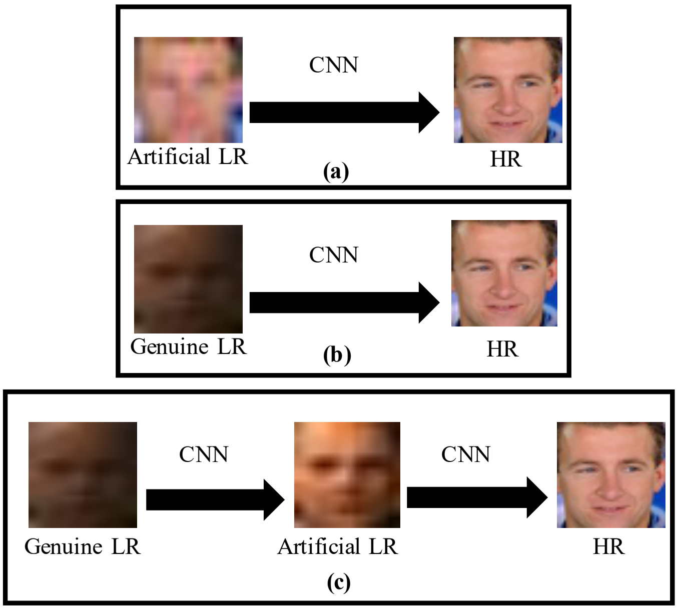

Figure 1.2: CNN architectures for facial image super-resolution. **(a)** A CNN is trained to super-resolve *artificial* LR facial images that are produced by down-sampling [45, 103]. It is a supervised learning method. **(b)** A CNN learns to adapt a *genuine* LR facial images into the HR style. Without LR-HR pairing supervision, a cycle consistency based loss function is often used for model training [237, 215, 20]. **(c)** The proposed characteristic regularisation method. The whole model training is regularised by characteristic consistifying from genuine LR facial images to artificial LR ones before super-resolved to the HR output. Best viewed in colour.

Using such cycle consistency for unsupervised domain adaptation has several adverse effects. The reconstruction loss is applicable only to the concatenation of two CNN models. This exacerbates the already challenging task of domain adaptation training. In the context of low-resolution face images for recognition, the genuine LR and HR image domains have significant differences in both image resolution and imaging conditions. Compared to a single CNN, the depth of a concatenated CNN-CNN model is effectively doubled. Existing UDA models apply the cycle consistency loss supervision at the final output of the second CNN, and propagate the supervision back to the first CNN. This gives rise to extra training difficulties in the form of *vanishing* gradients [104, 26]. In addition, jointly training two connected CNN models has to be conducted very carefully, along with the difficulty of training GAN models [59]. Moreover, the first CNN (the target model) takes responsibility of both characteristic consistifying and low-to-high resolution mapping, which further increases the model training difficulty dramatically.

This thesis solves the problem of super-resolving genuine LR facial images with high-fidelity resolved images by formulating a ***Characteristic Regularisation*** (CR) method (Fig 1.2(c)). In contrast to conventional image SR methods, this thesis particularly leverages the unpaired genuine LR images in order to take into account their characteristics information for facilitating model optimisation. Unlike cycle consistency based UDA methods, *the artificial LR images are instead leveraged as regularisation target* in order to separately learn the tasks of characteristic consistifying and image super-resolution. Specifically, the multi-task learning is performed, with the auxiliary task as *characteristic consistifying* (CC) for transforming genuine LR images into the artificial LR characteristics, and the main/target task as *image SR* for super-resolving both regularised and down-sampled LR images concurrently. Since there is no HR images coupled with genuine LR images, it is considered to align pixel content in the LR space by down-sampling the super-resolved images. This avoids the use of cycle consistency and their learning limitations. To make the super-resolved images with good facial identity information, an unsupervised semantic adaptation loss is formulated by aligning with the face recognition feature distribution of auxiliary HR images.

The CR method can be understood from two perspectives: (i) As splitting up the whole system into a model for image characteristic consistifying and a model for image SR. With the former model taking the responsibility of solving the characteristic discrepancy, the SR model can better focus on learning the resolution enhancement. This is in a divide-and-conquer principle. (ii) As a deeply supervised network [104], providing auxiliary supervision improves accuracy and convergence speed [179]. In the case of native face super resolution specifically, it allows for better and more efficient pre-training of SR module using paired artificial LR and HR images, pre-training of CC module by genuine and artificial LR images, and fast convergence in training the full CC+SR model.

1.2 Cross-Resolution Person Body Recognition

Person body recognition is an important aspect for identity recognition and tracking, in the cases of invisible facial appearance in images captured by unconstrained low-quality images, due to occlusion, back views or extreme distortion. In the literature, person body recognition is an important computer vision task and draws more and more attention recently, which aims to match the identity information in the images captured by disjoint surveillance camera views [58]. Most existing methods assume that the probe and gallery images have similar and sufficiently high resolutions. However, due to unconstrained distances between cameras and pedestrians, person images are often captured at various resolutions. This *resolution mismatch* issue brings about

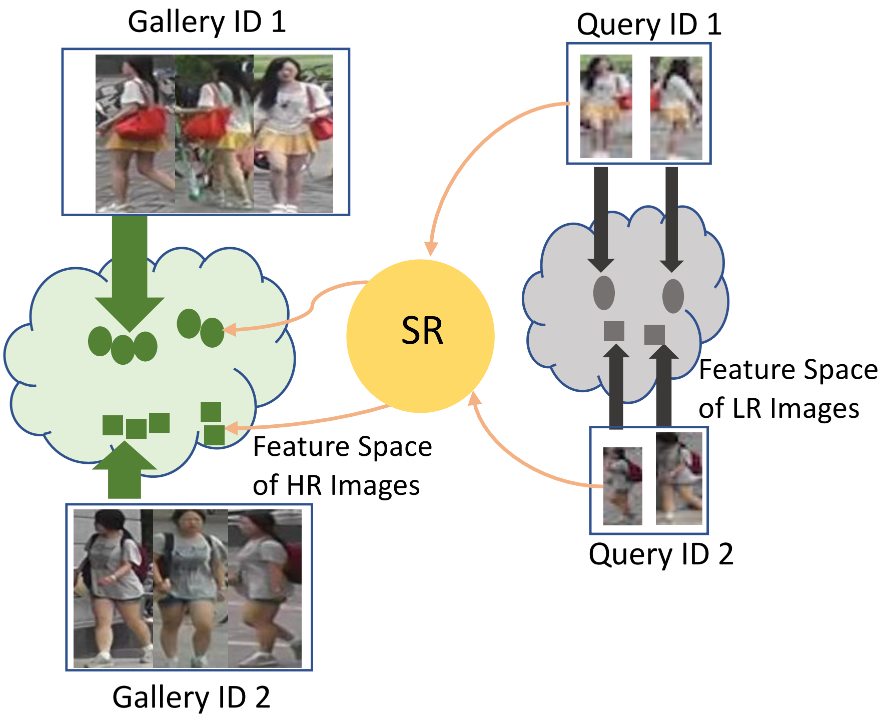

Figure 1.3: Illustration of cross-resolution person body recognition. Resolution mismatch between the low-resolution (LR) query images and the high-resolution (HR) gallery images causes unaligned feature distributions and finally inferior identity matching performance. One effective solution is using an image super-resolution (SR) model to enhance the resolution of LR query images for alleviating the distribution discrepancy with HR gallery images.

significant challenges to re-id. As low-resolution (LR) images contain much less identity detail information than high-resolution (HR) images, directly matching them across resolutions leads to substantial performance drop [86, 114]. For example, a standard person re-id model [53] can suffer up to 19.2% Rank-1 rate drop when applied to cross-resolution person re-id [114].

Existing cross-resolution person body recognition methodologies can be divided into two groups: (1) Learning resolution-invariant representation [36] and (2) Exploiting image super-resolution (SR) [86, 195]. This thesis focuses on the second category, since the first category tends to lose the fine-grained information contained in the HR images. However, directly applying the multi-task joint learning framework from the second category suffers from ineffective model training. It is because of the significantly higher difficulty of backpropagating the gradients through such a cascaded model [26]. As a consequence, as an auxiliary task, the SR module is less compatible with main task - person recognition.

This thesis proposes a novel regularisation named *Inter-Task Association Critic* (INTACT), which is dedicated to improve the compatibility between SR and recognition by an inter-task association mechanism. It is realised by two parts. In part one, the (unknown) inter-task association constraint is learned from the HR training data. In part two, the learned association module serves as a critic to supervise the SR learning.

1.3 Contributions

The contributions made in this thesis are summarised as follows:

1. Chapter 3: propose a novel Super-Resolution and Identity joint learning approach to face recognition in native LR images, with a unified deep network architecture. Unlike most existing FR methods assuming constrained HR facial images in model training and test, the proposed approach is specially designed to improve the model generalisation for LR face recognition tasks by enhancing the compatibility of face enhancement and recognition. This thesis further creates a large scale face recognition benchmark, named *TinyFace*, to facilitate the investigation of natively LR FR at large scales (large gallery population sizes) in deep learning. All the LR faces in TinyFace are collected from public web data across a large variety of imaging scenarios, captured under uncontrolled viewing conditions in pose, illumination, occlusion and background. Beyond artificially down-sampling HR facial images for LRFR performance test as in previous works, this is the first systematic study focusing specially on face recognition of native LR web images.

2. Chapter 4: proposes a novel super-resolution (SR) method for genuine low-resolution facial imagery. It combines the advantages of the existing image SR and unsupervised domain adaptation methods by a divide-and-conquer strategy. The proposed *Characteristic Regularisation* enables computationally more tractable model training and better model generalisation capability. A new unsupervised learning loss function is introduced without the limitations of cycle consistency. Extensive experiments are conducted on super-resolving both *genuine* and *artificial* LR facial imagery, with the former sampled from challenging unconstrained social-media and surveillance videos. The results validate the superiority of the model over the state-of-the-art image SR and domain adaptation methods, with high-fidelity resolved facial images.

3. Chapter 5: proposes an idea of leveraging the association between image SR and person body recognition tasks for solving the under-studied yet significant cross-resolution body recognition problem. A novel regularisation method, called *Inter-Task Association Critic*

(INTACT), is formulated for implementing the proposed inter-task association. INTACT is established on parameterising the association, and end-to-end trainable. Extensive experiments show the performance advantages of the INTACT over a wide range of state-of-the-art methods on five person body recognition benchmarks in the cross-resolution person body recognition problem.

1.4 Thesis Outline

This thesis is organised as follows:

Chapter 2 presents a review on various existing super resolution algorithms, person recognition with facial and body as biometric cues, existing face and body recognition benchmarks and models.

Chapter 3 improves the identity recognisability of super resolution for low-resolution person face recognition by introducing a novel Complement Super-Resolution and Identity (CSRI) joint deep learning method with a unified end-to-end network architecture, based on the idea of transferring the knowledge from synthetic to native SR. Extensive experiments show that such SR knowledge transfer model is able to benefit the identity matching performance.

Chapter 4 proposes an interpretable low-resolution face recognition model by formulating a high-fidelity SR model that joins the advantages of conventional SR and UDA models. The optimisations for characteristics consistifying and image super-resolving are separated and controlled by introducing Characteristic Regularisation between them, which makes the model training more effective and computationally tractable. Extensive evaluations demonstrate the performance superiority of this method in terms of both fidelity and identity recognisability.

Chapter 5 addresses the resolution mismatch issue for cross-resolution person body recognition when matched low-resolution person images against the high-resolution gallery images, by introducing a novel model training regularisation method, called Inter-Task Association Critic, to effectively leverage image super-resolution along with person body recognition in a joint learning manner. It is realised by parameterising the association constraint by automatically learning from the training data. Extensive experiments validate the superiority of INTACT on the cross-resolution body recognition task.

Chapter 6 provides conclusion and various research problems and directions to be pursued as further work.

Chapter 2

Literature Review

2.1 Person Face Recognition

This section provides a brief review on the existing face recognition algorithms, including models specially designed for low-resolution faces. This section also discusses super-resolution models for image fidelity and discriminability enhancement.

2.1.1 General Face Recognition

Early FR methods adopt hand-crafted features (e.g. Color Histogram, LBP, SIFT, Gabor) and matching model learning (e.g. discriminative margin mining, subspace learning, dictionary based sparse coding, Bayesian modelling) [10, 28, 1, 24, 223, 117]. They suffer from sub-optimal recognition generalisation, particularly with significant facial appearance variations, due to weak representation power (limited and incomplete human domain knowledge for hand-crafted features) and lack of end-to-end interaction learning between feature extraction and model inference.

Recently, deep learning based FR models [182, 96, 159, 144, 199, 122, 113, 73, 198] have achieved remarkable success. This paradigm benefits from superior network architectures [97, 167, 179, 72] and optimisation algorithms [199, 175, 159]. Deep FR methods naturally address the limitations of hand-crafted alternatives by jointly learning face representation and matching model end-to-end. A large set of labelled face images is usually necessary to train the millions of parameters of deep models. This can be commonly satisfied by large scale web face data collected and labelled (filtered) from Internet. Consequently, modern FR models are often trained, evaluated and deployed on web face datasets (Table 3.1 and Table 2.1).

This section gives a brief description of these FR models as follows:

The **DeepID2** model [173] is characterised by simultaneously learning face identification and verification supervision. Identification is to classify a face image into one ID class by softmax cross-entropy loss [97]. Formally, it predicts the posterior probability $\tilde{y}_i$ of a face image $\boldsymbol{I}_i$ over the ground-truth ID class y_i among a total n_{id} distinct training IDs:

$$p(\tilde{y}_i = y_i|\boldsymbol{I}_i) = \frac{\exp(\boldsymbol{w}_{y_i}^\top \boldsymbol{x}_i)}{\sum_{k=1}^{|n_{\text{id}}|} \exp(\boldsymbol{w}_k^\top \boldsymbol{x}_i)}, \tag{2.1}$$

where $\boldsymbol{x}_i$ specifies the DeepID2 feature vector of $\boldsymbol{I}_i$, and $\boldsymbol{W}_k$ the prediction parameter of the k-th ID class. The identification training loss is defined as:

$$l_{\text{id}} = -\log\Big(p(\tilde{y}_i = y_i|\boldsymbol{I}_i)\Big). \tag{2.2}$$

The verification signal encourages the DeepID2 features extracted from the same-ID face images to be similar so to reduce the intra-person variations. This is achieved by the pairwise contrastive loss [69]:

$$l_{\text{ve}} = \begin{cases} \frac{1}{2}\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2 & \text{if same ID,} \\ \frac{1}{2}\max\left(0, m - \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2\right)^2 & \text{otherwise.} \end{cases} \tag{2.3}$$

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

$$\mathcal{L}_{\text{DeepID2}} = l_{\text{id}} + \lambda_{\text{bln}} l_{\text{ve}}, \tag{2.4}$$

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

The **CentreFace** model [199] also adopts the softmax cross-entropy loss function (Eqn. (2.2)) to learn inter-class discrimination. However, it seeks for intra-ID compactness in a class-wise manner by posing a representation constraint that all face image features be close to the corresponding ID centre as possible. Learning this class compactness is accomplished by a centre loss function defined as:

$$l_{\text{centre}} = \frac{1}{2}\|\boldsymbol{x}_i - \boldsymbol{c}_{y_i}\|_2^2, \tag{2.5}$$

where y_i denotes the ID 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 ID are constrained to group together so that the intra-person variations can be suppressed. The final loss function is integrated with the identification supervision as:

$$\mathcal{L}_{\text{CentreFace}} = l_{\text{id}} + \lambda_{\text{bln}} l_{\text{centre}}. \tag{2.6}$$

Since the feature space is dynamic in the course of training, all class centres are progressively undated on-the-fly. The CentreFace model is implemented in a 28-layers ResNet architecture [72].

The **FaceNet** model [159] sues a triplet loss function [121] to learn a binary-class (positive *vs* negative pairs) feature embedding. The triplet loss is to induce a discriminative margin between positive and negative pairs, defined as:

$$\begin{aligned} l_{\text{tri}} = \max\left\{0,\ \alpha - \|\boldsymbol{x}_a - \boldsymbol{x}_n\|_2^2 + \|\boldsymbol{x}_a - \boldsymbol{x}_p\|_2^2\right\}, \\ \textit{subject to: } (\boldsymbol{x}_a, \boldsymbol{x}_p, \boldsymbol{x}_n) \in \mathcal{T}, \end{aligned} \tag{2.7}$$

where $\mathcal{T}$ denotes the set of triplets generated based on ID labels, and α is a pre-fixed margin for separating 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 ID are constrained to populate on an isolated manifold against other IDs by a certain distance therefore posing a discrimination capability. For fast convergence, it is critical to use triplets that violate the triplet constraint (Eqn. (2.7)). To achieve this in a scalable manner, hard positives and negatives are selected within a mini-batch.

The **VggFace** model [144] considers both identification and triplet training schemes in a sequential manner. Specifically, the model is trained by a softmax cross-entropy loss (Eqn. (2.2)). The feature embedding is then learned with a triplet loss (Eqn. (2.7)) where only the last full-connected layer is updated to implement a discriminative projection. A similar hard sample mining strategy is applied in the second step for more efficient optimisation. The VggFace adopts a 16-layers VGG16 CNN [167].

The **SphereFace** model [122] exploits a newly designed angular margin based softmax loss function. This loss differs from Euclidean distance based triplet loss (Eqn. (2.7)) by performing feature discrimination learning in a hyper-sphere manifold. The motivation is that, multi-class features learned by the identification loss exhibit an intrinsic angular distribution. Formally, the

Table 2.1: Face verification performance of state-of-the-art FR methods on the LFW challenge. "*": Results from the challenge leaderboard [79]. M: Million.

Feature Representation	Method	Accuracy (%)	Year	Training IDs	Training Images
Hand Crafted	Joint-Bayes [28]	92.42	2012	2,995	99,773
	HD-LBP [29]	95.17	2013	2,995	99,773
	TL Joint-Bayes [24]	96.33	2013	2,995	99,773
	GaussianFace [125]	98.52	2015	16,598	845,000
Deep Learning	DeepFace [182]	97.35	2014	4,030	4.18M
	DeepID [176]	97.45	2014	5,436	87,628
	LfS [214]	97.73	2014	10,575	494,414
	Fusion [183]	98.37	2015	250,000	7.5M
	VggFace [144]	98.95	2015	2,622	2.6M
	DeepID2 [173]	99.15	2014	10,177	202,599
	CentreFace [199]	99.28	2016	17,189	0.7M
	SphereFace [122]	99.42	2017	10,575	494,414
	DeepID2+ [177]	99.47	2015	12,000	290,000
	FaceNet [159]	99.63	2015	8M	200M
	TencentYouTu* [184]	99.80	2017	20,000	2M
	EasenElectron* [48]	99.83	2017	59,000	3.1M

angular softmax loss is formulated as:

$$
\begin{aligned}
l_{\text{ang}} &= -\log\Big(\frac{e^{\|\boldsymbol{x}_i\|\psi(\theta_{y_i,i})}}{e^{\|\boldsymbol{x}_i\|\psi(\theta_{y_i,i})}+\sum_{j\neq y_i} e^{\|\boldsymbol{x}_i\|\psi(\theta_{j,i})}}\Big),\\
\text{where}\ \ &\psi(\theta_{y_i},i) = (-1)^k\cos(m\theta_{y_i},i)-2k,\\
\text{subject to:}\ \ &\theta_{y_i,i}\in[\frac{k\pi}{m},\frac{(k+1)\pi}{m}],\ \ k\in[0,m-1],
\end{aligned}
\tag{2.8}
$$

where $\theta_{j,i}$ specifies the angle between normalised identification weight $\boldsymbol{W}_j$ ($\|\boldsymbol{W}_j\| = 1$) for j-th class and training sample $\boldsymbol{x}_i$, m ($m \geq 2$) the pre-set angular margin, and y_i the ground-truth class of $\boldsymbol{x}_i$. Specifically, this design manipulates the angular decision boundaries between classes and enforces a 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. Therefore, $\cos(m\theta_{y_i})$ represents a lower boundary of $\cos(\theta_{y_i})$ with larger m leading to a wider angular inter-class margin. Similar to CentreFace, a 28-layers ResNet CNN is adopted in the SphereFace implementation.

Despite advances in high-resolution FR, it remains unclear how well the state-of-the-art methods generalise to low-resolution images. Intuitively, low-resolution face recognition is extreme challenging due to two reasons: **(1)** Low-resolution faces contain much less appearance details with poorer quality and lower resolution. **(2)** Deep models are highly domain-specific and likely yield big performance degradation in cross-domain deployments, especially with large train-test domain gap, e.g. HR and LR faces. In such cases, transfer learning is challenging [142]. The scarcity of labelled surveillance data makes the problem even more challenging.

2.1.2 Low-Resolution Person Face Recognition

A challenge of face recognition in real-world applications is rooted in low-resolution (LR) [196]. Generally, existing low-resolution FR methods fall into two categories: **(1)** image super-resolution [65, 191, 74, 241, 193], and **(2)** resolution-invariant learning [41, 105, 202, 150, 161]. The first category is based on two learning criteria: pixel-level visual fidelity and ID discrimination. Existing models often focus more on appearance enhancement [65, 191]. Recent studies [74, 241, 193] attempt to unite the two sub-tasks for more discriminative learning. The second category aims to learn resolution-invariant features [41, 105] or a cross-resolution structure transformation [202, 150, 161]. The data-driven deep models can be conceptually categorised into this strategy whenever suitable training data is available for model optimisation.

However, all the existing methods have a number of limitations: (1) Considering small scale and/or *artificial* low-resolution face images in the closed-set setting, therefore unable to reflect the genuine LR face recognition challenge at scales. (2) Relying on hand-crafted features and linear/shallow model structures with suboptimal generalisation. (3) Requiring pixel-aligned low- and high-resolution training image pairs, which are unavailable for native low-resolution faces.

In low-resolution imagery face recognition deployments, two typical operational settings exist. One common setting is LR-to-HR (high-resolution) which aims to match LR probe images against HR gallery images such as passport or other document photos [14, 15, 161, 16, 150]. This is a widely used approach by Law Enforcement Agencies to matching potential candidates against a watch-list. On the other hand, there is another operational setting which requires LR-to-LR imagery face matching when both the probe and the gallery images are LR facial images [191, 65, 50, 241, 193].

Generally, LR-to-LR face recognition occurs in a less stringent deployment setting at larger scales when pre-recorded (in a controlled environment) HR facial images of a watch-list do not exist, nor there is a pre-defined watch-list. In an urban environment, there is no guarantee of controlled access points to record HR facial images of an average person in unconstrained public spaces due to the commonly used wide field of view of CCTV surveillance cameras and long distances between the cameras and people. In large, public space video surveillance data contain a very large number of “joe public” without HR facial images pre-enrolled for face recognition. Video forensic analysis often requires large scale people searching and tracking over distributed and disjoint large spaces by face recognition of *a priori* unknown (not enrolled) persons triggered by a public disturbance, when the only available facial images are from LR CCTV videos. More recently, a rapid emergence of smart shopping, such as the Amazon Go, Alibaba Hema and JD 7Fresh supermarkets, also suggests that any face recognition techniques for individualised customer in-store localisation (track-and-tag) cannot assume unrealistically stringent HR facial imagery enrollment of every single potential customer if the camera system is to be cost effective.

2.1.3 Person Face Recognition Datasets

An overview of representative face recognition challenges and benchmarks are summarised in Table 2.2.

Early challenges focus on *small-scale constrained* scenarios [148, 10, 132, 54, 166, 63, 156], with neither sufficient appearance variation for robust model training, nor practically solid test benchmarks. The seminal LFW [80] started to shift the community towards recognising unconstrained web faces, followed by even larger face benchmarks, such as CASIA [214], CelebFaces [175], VGGFace [144], MS-Celeb-1M [66], MegaFace [91] and MegaFace2 [137].

With such large benchmarks, FR accuracy in good quality images has reached an unprecedented level, e.g., 99.83% on LFW and 99.80% on MegaFace. However, this dose not scale to *native* low-resolution faces captured in unconstrained camera views, due to: (1) Existing datasets have varying degrees of data selection bias (near-frontal pose, less blur, good illumination); and (2) Deep methods are often domain-specific (only generalise well to test data similar to training set). On the other hand, there is a gap of facial images quality between a web photoshot view and a low-resolution view in-the-wild (Fig. 1.1).

Research on low-resolution face recognition has slightly advanced till recent period. It is under-studied with a very few benchmarks available. One of the major obstacles is the difficulty of establishing a large scale surveillance face dataset due to the high cost and limited feasibility in collecting surveillance faces and exhaustive ID annotation. Even in the FERET dataset, only simulated (framed) surveillance faces were collected in most cases with carefully controlled imaging settings, therefore it provides a much better facial image quality than those from native surveillance videos.

Table 2.2: Statistics of representative publicly available person face recognition benchmarks. Celeb: Celebrity.

Challenge	Year	IDs	Images	Videos	Subject	Surv?
Yale [10]	1997	15	165	0	Cooperative	No
QMUL-MultiView [56]	1998	25	4,450	5	Cooperative	No
XM2VTS [132]	1999	295	0	1,180	Cooperative	No
Yale B [54]	2001	10	5,760	0	Cooperative	No
CMU PIE [166]	2002	68	41,368	0	Cooperative	No
Multi-PIE [63]	2010	337	750,000	0	Cooperative	No
Morph [151]	2006	13,618	55,134	0	Celeb (Web)	No
LFW [80]	2007	5,749	13,233	0	Celeb (Web)	No
YouTube [201]	2011	1,595	0	3,425	Celeb (Web)	No
WDRef [28]	2012	2,995	99,773	0	Celeb (Web)	No
FaceScrub [138]	2014	530	100,000	0	Celeb (Web)	No
CASIA [214]	2014	10,575	494,414	0	Celeb (Web)	No
CelebFaces [175]	2014	10,177	202,599	0	Celeb (Web)	No
IJB-A [96]	2015	500	5,712	2,085	Celeb (Web)	No
VGGFace [144]	2015	2,622	2.6M	0	Celeb (Web)	No
UMDFaces [8]	2016	8,277	367,888	0	Celeb (Web)	No
MS-Celeb-1M [66]	2016	99,892	8,456,240	0	Celeb (Web)	No
UMDFaces-Videos [7]	2017	3,107	0	22,075	Celeb (Web)	No
IJB-B [200]	2017	1,845	11,754	7,011	Celeb (Web)	No
VGGFace2 [22]	2017	9,131	3.31M	0	Celeb (Web)	No
MegaFace2 [137]	2017	672,057	4,753,320	0	Non-Celeb (Web)	No
FERET [147]	1996	1,199	14,126	0	Cooperative	No
FRGC [146]	2004	466+	50,000+	0	Cooperative	No
CAS-PEAL [52]	2008	1,040	99,594	0	Cooperative	No
PaSC [13]	2013	293	9,376	2,802	Cooperative	No
SCface [62]	2011	130	4,160	0	Cooperative	Yes
COX [81]	2015	1,000	1,000	3,000	Cooperative	Yes
EBOLO [108]	2016	Unknown	6,135	0	Cooperative	Yes
FaceSurv [68]	2019	252	0	460	Cooperative	Yes
UCCS [64]	2017	1,732	14,016+	0	Uncooperative	Yes
SurvFace [40]	2020	15,573	463,507	0	Uncooperative	Yes

A notable recent study introduces the UCCS challenge [64], the current largest public surveillance face dataset, where faces were captured from a long-range distance without subjects' cooperation (unconstrained), with various poses, blurriness and occlusion. This benchmark represents a relatively realistic surveillance scenario compared to FERET. However, the UCCS images were captured at high-resolution from a single camera view[1], therefore providing significantly more facial details and less viewing angle variations. Moreover, UCCS is small in size, particularly the ID numbers (1,732), statistically limited for person face recognition evaluation in the context of this thesis.

Besides of a brief review of existing high-quality face recognition datasets, specifically, this section introduces a newly constructed larger scale native surveillance face recognition dataset, which is targeted for evaluating the state-of-the-art face-recognition models under the extremely low-quality (i.e., surveillance) imaging conditions, the SurvFace benchmark. It consists of 463,507 real-world surveillance face images of 15,573 different IDs captured from a diverse source of public spaces. More specifically, SurvFace explores real-world native surveillance imagery from a combination of 17 person body recognition benchmarks which were collected in different surveillance scenarios across diverse sites and multiple countries (Table 2.3). Fig 2.1 shows some images sampled from the SurvFace dataset.

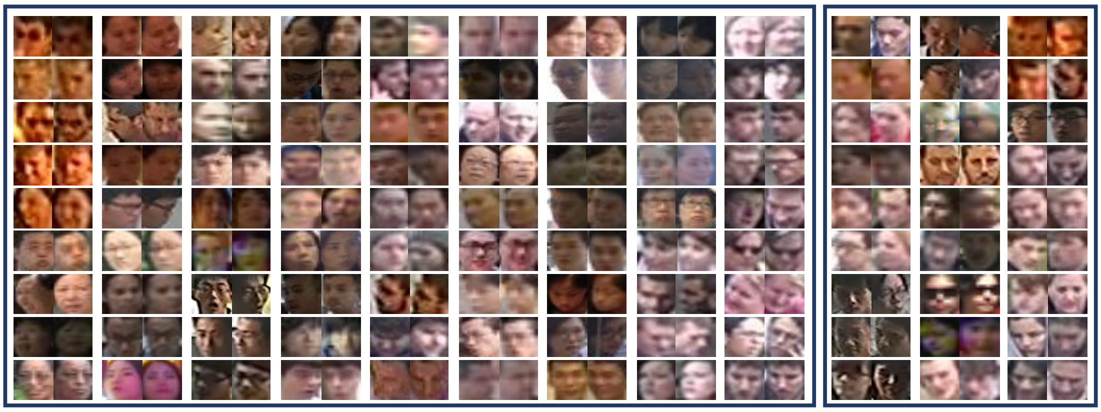

Figure 2.1: Matched (**Left**) and unmatched (**Right**) face image pairs from *SurvFace*.

With the constructed dataset, a systematic study focusing specially on face recognition in extremely low-quality imagery is conducted [40]. Some representative general face recognition methods [199, 144, 122] are evaluated. Specifically, face recognition in low-quality imagery is compared with web high-quality face recognition. For example, CentreFace achieves Rank-1

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

Table 2.3: Person body recognition datasets utilised in constructing the *SurvFace* challenge.

Person Body Recognition Dataset	IDs	Detected IDs	Bodies	Detected Faces	Nation
Shinpuhkan [90]	24	24	22,504	6,883	Japan
WARD [129]	30	11	1,436	390	Italy
RAiD [43]	43	43	6,920	3,724	US
CAVIAR4ReID [37]	50	43	1,221	141	Portugal
SARC3D [6]	50	49	200	107	Italy
ETHZ [160]	148	110	8,580	2,681	Switzerland
3DPeS [5]	192	133	1,012	366	Italy
QMUL-GRID [124]	250	242	1,275	287	UK
iLIDS-VID [190]	300	280	43,800	14,181	UK
SDU-VID [119]	300	300	79,058	67,988	China
PRID 450S [153]	450	34	900	34	Austria
VIPeR [61]	632	456	1,264	532	US
CUHK03 [109]	1,467	1,380	28,192	7,911	China
Market-1501 [230]	1,501	1,429	25,261	9,734	China
Duke4ReID [60]	1,852	1,690	46,261	17,575	US
CUHK-SYSU [207]	8,351	6,694	22,724	12,526	China
LPW [171]	4,584	2,655	590,547	318,447	China
Total	20,224	15,573	881,065	463,507	Multiple

29.9% on SurvFace, much inferior to the rate of 65.2% on MegaFace [199], i.e. a 54% (1-29.9/65.2) performance drop. This indicates that person face recognition in low-quality imagery is significantly more challenging, especially so when considering that one million distractors are used to additionally complicate the MegaFace test. It shows that existing general face recognition models generalise poorly under the low-quality imaging conditions.

Motivated by the above experimental studies, this thesis focus on addressing the limitations of existing face recognition models in low-quality imagery. Moreover, different from the SurvFace dataset, which is captured from the extremely low-quality surveillance scenarios, this thesis constructs a large-scale low-resolution face datasets collected from public web data across a large variety of imaging scenarios, for a better comparison to the popular high-quality face recognition datasets that are also collected from the web source (Sec. 3).

2.2 Person Body Recognition

This section focuses on the task of surveillance person body recognition (also known as person re-identification), which aims to track the same individuals by the subtle class information in the person images detected under the unconstrained scenarios.

2.2.1 General Person Body Recognition

There are an increasing number of studies for person body recognition in the past decade [2, 107, 114, 162, 172, 206, 231, 232, 34, 27, 31, 157]. Many of the existing works focus on addressing recognition challenges from variations in background clutter [110], human poses [118], or occlusion [133] across camera views. Specifically, Xiao et al. [206] propose a pipeline to learn generic and robust feature representations from multiple datasets to provide abundant data variations. Chen et al. [31] target to improve the generalisation capability of person recognition model by ensuring a larger inter-class variation and a smaller intra-class variation during model training. Saquib et al. [157] learn both the fine and coarse pose information of the person to fuse a discriminative embedding. Sun et al. [178] employ part-level features to provide fine-grained information. Researchers incorporate the specific domain knowledge of the whole-body biometric appearance to boost the person recognition models, such as part information [178], pose [118], and person recognition specific loss [31].

Moreover, there are extensive efforts on unsupervised learning [107, 106, 204, 32, 127, 115], domain adaptation [233, 33, 189, 145, 217, 203, 218], weak supervision [239, 131] for minimising the labelling efforts, and text-image person search [47, 216]. Specifically, the unsupervised cross-domain person body recognition targets to transfer the identity discriminative knowledge from a labelled source domain to an unlabelled target domain, by seeking a common feature space for source-target distribution alignment with discriminative learning constraints, reducing the domain discrepancy with GANs for domain styles transfer, or unifying the complementary benefits of synthetic images by GAN and feature discriminative constraints in CNN.

CNN backbone for person body recognition The deep-learning based person body recognition models usually borrow the representative CNN backbones generally designed for image recognition, such as VGG [167], Inception [179], and ResNet [72]. Recently, more advanced network structures [235, 34, 111, 178, 51, 188, 165, 170] have been specifically developed for person body recognition to boost the matching accuracy. Specifically, some networks are designed to especially exploit the upright body pose [178, 51, 188] by adding auxiliary supervision signals to features maps pooled horizontally by the last convolutional layer. Attention mechanisms are also employed to focus feature learning on the foreground person regions [165, 170]. Body part-specific CNNs are learned by means of off-the-shelf pose detectors [228, 227]. Some models branch CNNs to learn representations of global and local image regions [229].

More recently, Zhou et al. [235] learn multi-scale features explicitly at each layer of the networks as in a proposed OSNet. As the state-of-the-art person body recognition network structure, OSNet is utilised as the backbone for the cross-resolution person body recognition in this thesis (as described in Sec. 5). Specifically, omni-scale features refer to the features of both homogeneous (different spatial scales) and heterogeneous (arbitrary combination of multiple scales) scales, and a deep person body recognition CNN is designed, termed omni-scale network (OSNet), to learn the omni-scale feature. This is achieved by designing a residual block composed of multiple convolutional streams, each detecting features at a certain scale. More specifically, a unified aggregation gate is introduced to dynamically fuse multi-scale features with input-dependent channel-wise weights.

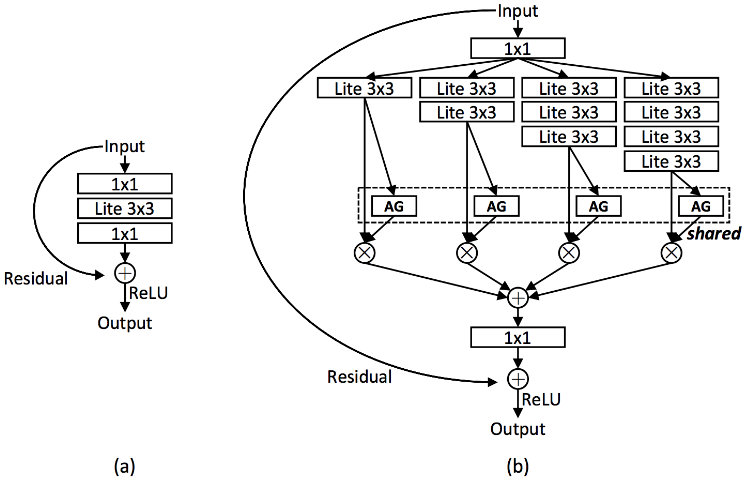

Figure 2.2: (a) Baseline bottleneck. (b) Proposed bottleneck in OSNet [235]. AG: Aggregation Gate. The first/last 1×1 layers are used to reduce/restore feature dimension.

2.2.2 Cross-resolution Person Body Recognition

However, among the existing person body recognition methods, a largely ignored aspect is that, persons images from unconstrained surveillance cameras often have varying resolutions, which would degrade the model performance if not properly handled.

To address the resolution mismatch problem, several cross-resolution person body recognition methods have been proposed [36, 86, 112, 114, 195]. They are fallen into two groups: (1) Learning resolution-invariant representation [36, 88, 112] and (2) Exploiting image super-resolution (SR) [86, 195]. In the *first* group, Jing et al. [88] propose to learn the mapping between HR and LR representations by a semi-coupled low-rank dictionary learning model; Li et al. [112] align the cross-resolution representation with a heterogeneous class mean discrepancy criterion. Chen et al. [36] learn the resolution-invariant representation by adding an adversarial loss on the representation features of HR and LR images. A weakness of these methods is that, such learned representations involve only coarse appearance information sub-optimal for body recognition. That is because fine-grained details, lacking in LR images but rich in HR images, are thrown away during learning for an agreement.

The *second* group of models, designed to exploit image super-resolution, can solve this limitation. Both methods [86, 195] adopt a joint learning strategy of SR and body recognition in a cascade, integrating identity-matching constraints with SR learning end-to-end. However, this design suffers from ineffective model training due to significantly higher difficulty of back-propagating the gradients through such a cascaded heavy model [26, 104]. Recently, Li et al. [114] combine the resolution-invariant representations with those exacted from resolution-recovered images, and achieve state-of-the-art performance. However, the above problem remains unsolved. To that end, in this work introduce a novel regularisation based on an inter-task association mechanism.

2.2.3 Person Body Recognition Datasets

This section provides a brief overview of publicly available datasets for the evaluation of person body recognition algorithms. Table 2.3 also lists the existing body recognition datasets. Specifically, this thesis utilises five representative datasets [114, 114, 114, 230, 60] for the evaluation of the proposed cross-resolution body recognition method, following the existing evaluation setting of multiple low-resolution (MLR) person body recognition [86, 114] for a better comparison. Noted that, among them, only the CAVIAR dataset [114] is captured under the real-world cross-resolution imaging condition and hence provides realistic images of multiple resolutions, i.e.a genuine MLR dataset for evaluating cross-resolution person body recognition. The other four synthetic cross-resolution datasets are constructed from the original datasets (CUHK03, VIPeR, Market-1501, DukeMTMC-reID) as follows: the query images taken from one camera are down-sampled by a randomly selected downsampling rate r $\in \{2, 3, 4\}$ (i.e.the spatial size of a down-sampled image becomes H/r × W/r), while the images taken by the other camera(s) remain unchanged. The Multiple Low Resolution (MLR) datasets are named as ***MLR-dataset*** [86].

The details of the selected evaluation datasets are listed below: The CUHK03 dataset comprises 14,097 images of 1,467 identities with 5 different camera views. As [114], the 1,367/100 training/test identity split was used. The VIPeR dataset contains 632 person image pairs captured by 2 cameras. Following [114], this dataset was randomly divided into two non-overlapping halves based on the identity labels. Namely, images of a subject belong to either the training or the test set. The CAVIAR dataset contains 1,220 images of 72 person identities captured by 2 cameras. 22 people who only appear in the closer camera were discarded, and the remaining was split into two non-overlapping halves in the identity labels as [114]. The Market-1501 dataset consists of 32,668 images of 1,501 identities captured in 6 camera views. The standard 751/750 training/test identity split was used. The DukeMTMC-reID dataset contains 36,411 images of 1,404 identities captured by 8 cameras. The standard 702/702 training/test identity split was adopted.

2.3 Biometric Cues for Person Recognition

Among existing person biometric cues, the most commonly used feature is the whole-body cue, which usually relies on the low-level clothing colour and texture feature, or mid-level attribute-based features [130]. In the early works that adopt the hand-crafted features for person recognition, the person foreground (body appearance) is segmented from the background [49, 55]. Recently, CNN-based deep learning models have been widely used for person recognition. Such deep learning models take the person images as inputs and implicitly extract the identity representations from foreground whole-body cues globally or partially[58, 107, 114, 162, 172, 206]. In addition to the whole-body cue, there are various visual biometrics based on person appearance are discovered and exploited for person matching, including iris [148], gait [158, 70, 42], and fingerprint [128].

Specifically, gait information is usually used in multi-shot recognition schemes for video-based person recognition. Han et al. [70] employ the spatio-temporal gait representation to characterise human walking properties for individual recognition by gait. Lam et al. [100] propose a gait representation - gait flow image for gait recognition. A following work [186] develop a temporal template to preserve the s temporal information in a gait sequence. Liu et al. [119]

incorporate the temporal alignment for gait representations. Recently, deep CNN networks are employed for similarity learning in gait based person recognition [205, 163]. In terms of identity recognition with iris cue, Phillips et al. [148] establish the first independent performance benchmark for iris recognition technology. A recent work [139] exploit the CNN network to extract off-the-shelf iris representations. The compactness of deep iris model and the small number of computing operations are later studied with constrained design in [140].

Compared to the other popular biometric cues for person recognition, face is probably the most reliable, visually accessible biometric to person identities, especially for the long-term tracking where clothing colour and other biometrics are easily to change. However, face cue is much less widely-used than the other cues. It is mainly due to the significant challenge brought about by the low resolution and pose variations of individuals in typical surveillance footage. Besides, the lack of large-scale surveillance facial identity recognition benchmarks characterised by realistic low resolution and other unconstrained camera conditions also limit the study of person recognition with face cue.

2.4 Image Super-Resolution

2.4.1 General Super Resolution

Image Super-Resolution (SR) aims to enhance the resolution and recover the high-resolution outputs from low-resolution images. SR enables a wide range of practical applications, including surveillance and security [149, 141] and benefits a variety of downstream tasks [71, 3, 114]. This is an inherently ill-posed problem, as a low-resolution image can be explained by many different high-resolution outputs, i.e., multiple solutions could exist for any given low-resolution pixel, which is an underdetermined inverse problem. Typically, a SR model is trained to constrain the solution space and the solutions of optimisation-based super-resolution methods is principally driven by the choice of the objective function [103].

Recently, image super-resolution has rapidly developed thanks primarily to the powerful modelling capacity of deep models especially the family of CNNs in regressing the pixel-wise loss (the mean squared error (MSE) loss) between the reconstructed and ground-truth high-resolution images [211, 44, 46, 102, 220, 92, 93, 98, 180]. Among them, several representative CNN based SR models are selected for model evaluation in the proposed benchmarks in Sec. 3. This section gives a brief introduction of the these SR models as follows:

The **SRCNN** model [44] is one of the first deep methods achieving remarkable success in super-resolution. The design is motivated by earlier sparse-coding based methods [212, 94]. By taking the end-to-end learning advantage of neural networks, SRCNN formulates originally separated components in a unified framework to realise a better mapping function learning. A mean squared error (MSE) is adopted as the loss function:

$$l_{\text{mse}} = \|f(\boldsymbol{I}^{\text{lr}};\boldsymbol{\theta}) - \boldsymbol{I}^{\text{hr}}\|_2^2 \tag{2.9}$$

where $\boldsymbol{I}^{\text{lr}}$ and $\boldsymbol{I}^{\text{hr}}$ denotes coupled low- and high-resolution training images, and $f()$ the to-be-learned super-resolution function with the parameters denoted by $\boldsymbol{\theta}$. This model takes bicubic interpolated images as input.

The **FSRCNN** model [46] is an accelerated and more accurate variant of SRCNN [44]. This is achieved by taking the original low-resolution images as input, designing a deeper hourglass (shrinking-then-expanding) shaped non-linear mapping module, and adopting a deconvolutional layer for upscaling the input. The MSR loss function (Eqn. (2.9)) is used for training.

The **VDSR** model [92] improves over SRCNN [44] by raising the network depth from 3 to 20 convolutional layers. The rational of deeper cascaded network design is to exploit richer contextual information over large image regions (e.g. 41×41 in pixel) for enhancing high frequency detail inference. To effectively train this model, the residual learning scheme is adopted. That is, the model is optimised to learn a residual image between the input and ground-truth high-resolution images. VDSR is supervised by the MSE loss (Eqn. (2.9)).

The **DRRN** model [180] constructs an even deeper (52-layers) network by jointly exploiting residual and recursive learning. 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 ID branches to mitigate the information loss across all the layers. This leads to a multi-path structured network module. Inspired by [180], 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 MSE loss function (Eqn. (2.9)) is used to supervise in model training.

The **LapSRN** model [98] consists in a multi-levels of cascaded sub-networks designed to progressively predict high-resolution reconstructions in a coarse-to-fine fashion. This scheme is hence contrary to the four one-step reconstruction models above. Same as VDSR and DRRN, the residual learning scheme is exploited along with an upscaling mapping function to alleviate the training difficulty whilst enjoying more discriminative learning. For training, it adopts a Charbonnier penalty function [18]:

$$l_{\text{cpf}} = \sqrt{\|f(\boldsymbol{I}^{\text{lr}};\boldsymbol{\theta}) - \boldsymbol{I}^{\text{hr}}\|_2^2 + \varepsilon^2} \tag{2.10}$$

where ε (e.g. set to 10^{-3}) is a pre-fixed noise constant. Compared to MST, this loss has a better potential to suppress training outliers. Each level is supervised concurrently with a separate loss against the corresponding ground-truth high-resolution images. This multi-loss structure resembles the benefits of deeply-supervised models [104, 208].

A performance summary of six representative deep super-resolution models is given on five popular benchmarks in Table 2.4, where the peak signal-to-noise ratio (PSNR) is the most widely-used reconstruction quality measurement metric for super resolution. PSNR is defined as follows:

$$\text{PSNR} = 10\log(L^2/(1/N\sum_{i=1}^{N}(\boldsymbol{I}^{\text{hr}}(i) - f(\boldsymbol{I}^{\text{lr}})(i))^2)) \tag{2.11}$$

where L refers to the maximum pixel value, and I, $\hat{I}$ are the ground-truth and reconstructed images with N pixels, respectively. Generally, as a pixel-level metric, PSNR only measures the differences between corresponding pixels instead of visual perception. Therefore, CNN models supervised by MSE loss that favours the PSNR metric often tend to suffer from poor performance in representing the reconstruction quality in real scenes, which is often related to human perceptions, and the downstream tasks like low-resolution object recognition in real-world applications. However, due to the lack of completely accurate perceptual metrics, PSNR is still currently the popular evaluation metric for SR. Moreover, the commonly-used super-resolution benchmarks generate the low-resolution images simply by *imresize* function with default settings in MATLAB (i.e., bicubic interpolation). Such benchmarks lack of realistic distracting artifacts, e.g. noises, motion blurriness, imaging compression, non-ideal point spread function, and other aliasing effects, that the native low-resolution person images captured from in-the-wild scenes, such as surveillance and social events [226, 40], usually contain, which is the focus of this thesis. Therefore, the "ideal" benchmarks with high-low resolution paired images and the benchmarking results do not ensure the superior performance of the evaluated existing state-of-the-art SR models in the context of unconstrained low-resolution person images.

Recently, to match the fidelity of the resolved faces expected at higher resolution, Generative Adversarial Networks (GANs) based image SR models [35, 103, 155, 219, 85, 11, 3, 143, 209] have been introduced which additionally exploit an unsupervised adversarial learning loss on top

Table 2.4: The performance summary of state-of-the-art image super-resolution methods on six popular benchmarks. None of these benchmarks are designed for FR due to their development independence. Metric: Peak Signal-to-Noise Ratio (PSNR), higher is better.

Model	Scale	Set5	Set14	B100	URGAN	MANGA
Bicubic	2	33.65	30.34	29.56	26.88	30.84
SRCNN [44]	2	36.65	32.29	31.36	29.52	35.72
FSRCNN [46]	2	36.99	32.73	31.51	29.87	36.62
VDSR [92]	2	37.53	33.03	31.90	30.76	-
DRCN [93]	2	37.63	32.98	31.85	30.76	37.57
LapSRN [98]	2	37.52	33.08	31.80	30.41	37.27
DRRN [180]	2	37.74	33.23	32.05	31.23	-
Bicubic	4	28.42	26.10	25.96	23.15	24.92
SRCNN [44]	4	30.49	27.61	26.91	24.53	27.66
FSRCNN [46]	4	30.71	27.70	26.97	24.61	27.89
VDSR [92]	4	31.35	28.01	27.29	25.18	-
DRCN [93]	4	31.53	28.04	27.24	25.14	28.97
LapSRN [98]	4	31.54	28.19	27.32	25.21	29.09
DRRN [180]	4	31.68	28.21	27.38	25.44	-
Bicubic	8	24.39	23.19	23.67	20.74	21.47
SRCNN [44]	8	25.33	23.85	24.13	21.29	22.37
FSRCNN [46]	8	25.41	23.93	24.21	21.32	22.39
LapSRN [98]	8	26.14	24.44	24.54	21.81	23.39

of the conventional MSE loss. These GAN methods often produce more photo-realistic and visually appealing images. Specifically, a GAN based SR model consists of a generator performing high-resolution image reconstruction from the low-resolution inputs, and a discriminator taking inputs the super-resolved images and high-resolution and determine whether the inputs are the high-resolution images or not. The generator and discriminator are trained alternatively, targeting to a generator that can generate more realistic images to fool the discriminator. Ledig et al. [103] firstly propose a GAN based SR model by exploiting an adversarial loss on top of the MSE loss as follows:

$$\mathcal{L}_{\text{gan}} = \mathbb{E}_{x_h}[\log D(x_h)] + \mathbb{E}_{x_l}[\log\left(1 - D(G(x_l))\right] \tag{2.12}$$

where x_l indicates the LR input, and x_h is the corresponding HR counterpart. More specifically, the generator G tries to minimise the objective value against an adversarial discriminator D that instead tries to maximise the value. The optimal solution is obtained as:

$$G^* = \arg\min_G \max_D \mathcal{L}_{\text{gan}}. \tag{2.13}$$

Such adversarial loss is later adopted by a variety of SR models [155, 35, 219, 85, 143, 209]. Currently, GAN training is still unstable. Although with more and more studies on GAN training stabilising [134], it remains an open research problem that how to correctly integrate and train the GANs into SR models.

2.4.2 Super-resolution for genuine imagery

Although with the rapid development of super resolution, the pairwise supervised learning, that most of the existing SR models rely on, becomes infeasible when there is no such pairwise HR-LR training data, e.g. native poor quality facial imagery data from in-the-wild social media and surveillance videos. Existing works consider mostly an *artificial* image SR problem where the LR images are *synthesised* by some pre-defined down-sampling processes that nevertheless retain essentially the same noise characteristics as their corresponding HR images, rather than super-resolving target *native* LR images of unknown and significantly different noise properties as compared to unpaired good quality source HR images. The latter is a much harder problem.

There are a few recent attempts on resolving genuine image SR [19, 20, 38, 40, 164]. In particular, Shocher et al. [164] learn an image-specific CNN model for each test time based on the internal image statistics. Whilst addressing the problem of pairwise training data limitation, this method is computationally expensive from on-the-fly per test image model learning, even with small (compact) neural networks. Zhou et al. [236] propose a kernel modelling super-resolution network that incorporates blur-kernel modelling in the training, where the constructed realistic blur kernels are generated by a generative adversarial network (GAN). However, the generated blur-kernels can be limited to model the realistic scenarios with low-quality images degraded by multiple artifacts, including blur, noise, imaging compression, non-ideal point spread function, and other aliasing effects, especially in surveillance.

Bulat and Tzimiropoulos [19] develop an end-to-end adversarial learning method for both face SR and alignment with the main idea that jointly detecting landmarks provides global structural guidance to the super-resolution process. This method is however sensitive to alignment errors. Bulat et al. [20] utilise the external training pairs, where the LR inputs are generated by simulating the real-world image degradation instead of simply down-sampling. This method presents an effective attempt on genuine LR image enhancement. However, it suffers from an issue of model input discrepancy between training (simulated genuine LR images) and test (genuine LR images). On the other hand, unsupervised domain adaptation (UDA) models [237, 215, 95] also offer a potential solution for genuine LR image super-resolution. This approach often uses some cycle consistency based loss function for model optimisation, which unfavourably makes the training difficult and ineffective. To tackle the absence of pixel-alignment between LR and HR training images, Cheng et al. [38, 40] explore facial identity information to constrain the learning of a SR model. However, this semantic regularisation fails to yield appealing visual quality.

In contrast to all the existing solutions, this thesis formulates a unified method that enjoy the strengths of both conventional SR and UDA methods in a principled manner (Sec. 4). In particular, the model separates the image characteristic consistifying (adaptation) and image super-resolution tasks by characteristic regularisation. Importantly, this makes the model training more effective and computationally more tractable, leading to superior model generalisation capability.

2.4.3 Face Image Super-resolution

As an important domain-specific super-resolution, face image 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 [4, 192, 25, 116, 84, 87, 238, 21, 222]. 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. Specifically, the Super-FAN [19] utilises FAN to constrain the alignment of facial landmarks by end-to-end multi-task learning. And the FSRNet [35] exploits both facial landmark heatmap and face parsing maps as prior constraints. Besides of the structure information, another way is to utilise the face attribute or identity information, e.g., to constrain the recovered face images to have the identical face-related attributes to ground truth. For example, the CBN [238] exploits the facial prior by alternately optimising super resolution and dense correspondence field estimation. The SICNN [224], adopts a super-identity loss function to recover the real identity during super resolution.

Besides of the explicit facial priors used in face super resolution, there are some methods implicitly adopt the facial structure information, especially for noise and low-quality inputs where the facial structure alignment is unavailable. The TDN [221] incorporates spatial transformer networks [82] to tackle the face unalignment problem. The following work TDAE [222] adopts a decoder-encoder-decoder framework to first upsample and denoise the low-resolution inputs before super resolution with facial alignment. Yang et al. [210] propose to decompose face images into facial components and background, retrieve adequate HR exemplars in external datasets by component landmarks, and fuse them to complete HR faces with generic SR for background pixels.

In addition to the face super resolution models supervised by pixel-wise loss and explicit or implicit facial priors, similar to the GAN based SR models, researchers also adopt the adversarial loss to model the global facial structure information from the target data distribution (i.e., realistic high-resolution face images) [220, 209, 219]. Moreover, researchers also improve face super resolution from other perspectives. Motivated by the human attention shifting mechanism [136], the attention-aware face super resolution [21] adopt the deep reenforcement learning to discover attention face patches for local enhancement, and thus fully exploits the global interdependency of face images.

However, this mapping relationship is typically learned from aligned low- and high-resolution image pairs. To exploit super resolution for genuine face images without paired supervision, researchers have made a few attempts [20] based on unsupervised domain adaption algorithms [237, 215, 95], which is however difficult to train and ineffective. Moreover, existing methods often require noise-free input images and assume stringent part detection and dense correspondence alignment, otherwise artifacts can be easily introduced in hallucination. These requirements may significantly restrict their usability to the low-resolution surveillance face images due to the presence of uncontrolled noise and poor quality, as well as the lack of aligned high-resolution counterparts. So far, how effective contemporary image super-resolution and face hallucination methods are for the low-resolution surveillance FR challenge remains unclear especially in large scale deployments. This thesis will carry out the corresponding model investigation and extensive experimental evaluations later on.

2.4.4 Image Super Resolution for Object Recognition

In typical surveillance scenarios, images are often captured from a long distance or wide angle, leading to the region of interest to be of a low resolution [241]. Images of low resolution brings about significant challenges to object recognition given the limited discriminative information and blur. Super resolution that targets to super resolve the high-resolution details from low-resolution inputs may provide a potential solution to the low-resolution object recognition task.

The low-resolution object recognition problem has drawn attentions in recent years [30, 38, 126, 154, 193, 224]. Among them, there are some studies that explore how to incorporate image super resolution techniques for the low-resolution image recognition. Instead of utilising the super resolved images for object recognition, Wang et al. [193] utilise the super resolution model as a pre-trained priors to train the recognition model for low-resolution inputs. There are other studies [38, 224, 39] that unite image SR and object recognition in a multi-task joint-learning framework, where the low-to-high reconstruction loss used for super resolution is adopted as auxiliary supervision constraints. Specifically, Zou and Yuen [241] propose one of the first super resolution models with specific focus on low-resolution object recognition, with a discriminative constraint for learning features useful for recognition. Singh et al. [169] propose an identity-aware face super resolution technique for generating a HR image from a given LR input. Later, Singh et al. [168] propose to adopt the super resolution reconstructed loss in the capsule network to constrain the capsules of low- and high-resolution images of the same class to be similar.

Image super resolution is also used to address the resolution mismatch problem in cross-resolution recognition, for example, the cross-resolution person body recognition task [86, 195]. Both methods [86, 195] adopt a joint learning strategy of SR and body recognition in a cascade, integrating identity-matching constraints with SR learning end-to-end. Recently, Li et al. [114] combine the resolution-invariant representations with those exacted from resolution-recovered images, and achieve state-of-the-art performance. Nonetheless, such multi-task designs in joint-learning scheme often suffers from the heavy concatenated models and hence ineffective model training due to significantly higher difficulty of back-propagating the gradients [26, 104]. To that end, this thesis introduces a novel regularisation based on an inter-task association mechanism (Sec. 5).

Although with a few works focusing on super resolution for low-resolution recognition, in large, object recognition and super-resolution advance independently, with both assuming the availability of high-resolution training data such as high-quality web images. It still remains an open research problem that how to incorporate super resolution for low-resolution object recognition. Especially in surveillance, high-resolution images are typically scarce or unavailable, which limits these existing methods to the indirect model transfer learning strategy. Besides of the problem of recognition-aware super resolution, when the training (synthesised low-resolution images) and test data (genuine low-resolution images) distributions are very different to each other (typical in reality for the surveillance facial identity matching 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 a much more challenging super-resolution task.

2.5 Summary

The preceding sections have discussed important studies in the literature in terms of super resolution, face recognition and body recognition, and the super resolution techniques in the context of low-resolution person recognition. Despite the developments achieved by existing methods, there remains many limitations and open problems. In the following chapters, novel approaches are presented to overcome the challenges as outlined below:

1. (Chapter 3) **Improve the identity recognisability of super resolution for low-resolution person face recognition**: propose a joint deep learning method with a unified end-to-end network architecture, based on the idea of transferring the knowledge from synthetic to native SR. Extensive experiments show that such SR knowledge transfer model is able to benefit the identity recognition performance.

2. (Chapter 4) **Improve the fidelity of the super resolved facial images**: formulates a method that joins the advantages of conventional SR and UDA models. The optimisations for characteristics consistifying and image super-resolving are separated and controlled by introducing Characteristic Regularisation between them, which makes the model training

more effective and computationally tractable. Extensive evaluations demonstrate the performance superiority of this method in terms of both fidelity and identity recognisability.

3. (Chapter 5) **Solve the resolution mismatch problem for person body recognition**: introduces a novel model training regularisation method to effectively leverage image super-resolution (SR) along with person body recognition in a joint learning manner. It is realised by parameterising the association constraint by automatically learning from the training data. Extensive experiments validate its superiority on the cross-resolution person body recognition task.

Chapter 3

Low-Resolution Face Recognition by Super-Resolution

3.1 Introduction

Recognising native low-resolution faces is extremely challenging, given the lack of sufficient visual information for current deep models to learn expressive feature representations. Designed for enhancing high-resolution image details, super resolution may be beneficial for this task. However, directly employing the existing SR models does not benefit the native low-resolution recognition problem. It is mainly because that the existing SR models usually require the pixelwise-aligned LR and HR image pairs for model training, and hence are trained with synthesised HR-LR image pairs by down-sampling images, instead of the native facial low-resolution images.

The SR models trained with synthesised LR data suffer from significant performance drop when applied to the realistic facial images for person recognition. This is due to the significant domain gap of different imaging noise characteristics between native low-resolution facial images and high-resolution web face photo-shoots. The SR models trained by *synthetic* LR images do not capture the unknown and significantly different imaging noise and artifacts inherent to the *native* LR images, e.g. sensor noise, compression, non-ideal point spread function, among other aliasing effects. This domain transfer discrepancy between the training data from one domain (source) and the test data from a very different domain (target) causes inherent model limitations for poor performance generalisation among existing SR algorithms. To solve this problem, one potential solution is to adopt the domain-adaptation strategy with the auxiliary data of artificial down-sampled web faces for native facial image super-resolution.

Specifically, a joint learning scheme is adopted in a unified deep network architecture, specially dedicated to improve the model generalisation to low-resolution inputs by learning the face enhancement and recognition in an end-to-end manner. With the jointly optimising, it is effective to reduce the negative effect of noisy fidelity. A complement learning mechanism is introduced considering the absence of HR facial images, by transferring the super-resolving knowledge from the auxiliary artificial super-resolution learning task to the natively LR facial data. Taken together with joint learning, the proposed method is formulated as *Complement Super-Resolution and Identity joint learning* (**CSRI**).

In the experiments, this chapter benchmarks the performance of four state-of-the-art deep learning FR models [144, 122, 174, 199] and three super-resolution methods [44, 180, 92] on the TinyFace dataset. It is observed that the existing deep learning FR models suffer from significant performance degradation when evaluated on the TinyFace challenge. The results also show the superiority of the proposed CSRI model over the state-of-the-art methods on the low-resolution face recognition tasks.

3.2 Methodology

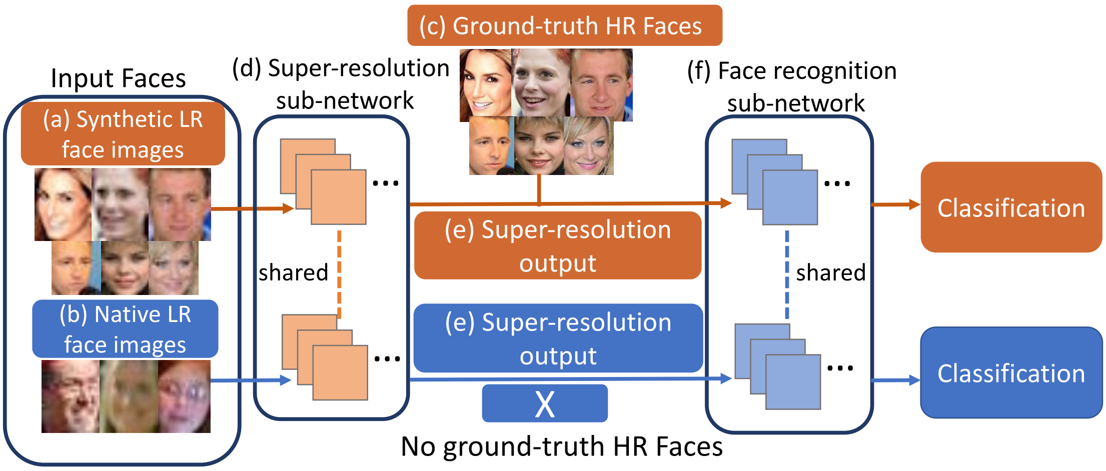
(c) Ground-truth HR Faces
Input Faces
(d) Super-resolution sub-network
(f) Face recognition sub-network
(a) Synthetic LR face images
(b) Native LR face images
shared
shared
(e) Super-resolution output
(e) Super-resolution output
X
Classification
Classification
No ground-truth HR Faces

Figure 3.1: An overview of the proposed Complement-Super-Resolution and Identity (CSRI) joint learning architecture. The CSRI contains two branches: (Orange): Synthetic LR SR-FR branch; (Blue): Native LR SR-FR branch. The two branches share parameters.

To recognise native low-resolution faces, it is essential to extract identity discriminative feature representations from LR unconstrained images. To that end, a deep neural network architecture is proposed for Complement-Super-Resolution and Identity joint learning. This approach is based on two considerations: (1) Joint learning of Super-Resolution and FR for maximising their compatibility and complementary advantages; (2) Complement-Super-Resolution learning for maximising the model discrimination on native LR face data at the absence of native HR counterparts in further SR-FR joint learning.

One major challenge in optimising the super resolution component on native low-resolution faces is that there are no coupled HR images. To address this problem, it is considered to transfer knowledge from auxiliary HR face data on which LR/HR pairs can be constructed by down-sampling.

CSRI Overview. Given the CSRI design above, a multi-branch network architecture is considered (Fig. 3.1). The CSRI contains two branches:

1. A *synthetic LR SR-FR* branch: For improving the compatibility and complementary advantages of SR and FR components by jointly learning auxiliary face data with artificially down-sampled LR/HR pairs (the top stream in Fig. 3.1);

2. A *native LR SR-FR* branch: For adapting super-resolving information of auxiliary LR/HR face pairs to the native LR facial imagery domain which lacks the corresponding HR faces by complement SR-FR learning (the bottom stream in Fig. 3.1).

In this study, the CSRI is instantiated by adopting the VDSR [92] for the SR component and the CentreFace [199] for the FR component. These CSRI components are detailed as follows.

(I) Joint Learning of Super-Resolution and Face Recognition. To adapt the image SR ability for recognition, a SR-FR joint learning strategy is considered by integrating the output of SR with the input of FR in the CSRI design so to exploit the intrinsic end-to-end deep learning advantage. To train this SR-FR joint network, both auxiliary training data with artificially down-sampled LR/HR face pairs $\{(\boldsymbol{I}^{\text{alr}}, \boldsymbol{I}^{\text{ahr}})\}$ and face identity labels $\{y\}$ (e.g. CelebA [123]) are used. Formally, a SR model represents a non-linear mapping function between LR and HR face images. For SR component optimisation, the pixel-level Mean-Squared Error (MSE) minimisation criterion is defined as

$$\mathcal{L}_{\text{sr}} = \|\boldsymbol{I}^{\text{asr}} - \boldsymbol{I}^{\text{ahr}}\|_2^2, \tag{3.1}$$

where $\boldsymbol{I}^{\text{asr}}$ denotes the super-resolved face image of $\boldsymbol{I}^{\text{alr}}$ (Fig. 3.1(a)), and $\boldsymbol{I}^{\text{ahr}}$ denotes the corresponding HR ground-truth image (Fig. 3.1(c)).

Using the MSE loss intrinsically favours the Peak Signal-to-Noise Ratio (PSNR) measurement, rather than the desired low-resolution face recognition performance. This limitation is addressed by concurrently imposing the FR criterion in optimising SR. Formally, the performance of the FR component is quantified by the softmax Cross-Entropy loss function defined as:

$$\mathcal{L}_{\text{fr}}^{\text{syn}} = -\log(p_y), \tag{3.2}$$

where y is the face identity, and p_y the prediction probability on class y by the FR component. The SR-FR joint learning objective is then formulated as:

$$\mathcal{L}_{\text{sr-fr}} = \mathcal{L}_{\text{fr}}^{\text{syn}} + \lambda_{\text{sr}} \mathcal{L}_{\text{sr}}, \tag{3.3}$$

where λ_{sr} is a weighting parameter for the SR loss quantity. $\lambda_{\text{sr}} = 0.003$ is set by cross-validation in the experiments. In doing so, the FR criterion enforces the SR learning to be identity discriminative simultaneously.

(II) Complement-Super-Resolution Learning. Given the SR-FR joint learning as above, the CSRI model learns to optimise the FR performance on the synthetic (artificially down-sampled) auxiliary LR face data. This model is likely to be sub-optimal for native LRFR due to the inherent visual appearance distribution discrepancy between synthetic and native LR face images (Fig. 3.5).

To overcome this limitation, the super-resolution and recognition joint learning is further constrained towards the native LR data by imposing the native LR face discrimination constraint into the SR component optimisation. Specifically, the SR and FR components are jointly optimised using both auxiliary (with LR/HR pairwise images) and native (with only LR images) training data for adapting the SR component learning towards native LR data. That is, the synthetic and native LR branches are concurrently optimised with the parameters shared in both SR and FR components. To enforce the discrimination of labelled native LR faces, the same Cross-Entropy loss formulation is used.

Overall Loss Function. After combining three complement SR-FR learning loss quantities, the final CSRI model objective is defined as:

$$\mathcal{L}_{\text{csrl}} = (\mathcal{L}_{\text{fr}}^{\text{syn}} + \mathcal{L}_{\text{fr}}^{\text{nat}}) + \lambda_{\text{sr}} \mathcal{L}_{\text{sr}}, \tag{3.4}$$

where $\mathcal{L}_{\text{fr}}^{\text{nat}}$ and $\mathcal{L}_{\text{fr}}^{\text{syn}}$ measure the identity discrimination constraints on the native and synthetic LR training data, respectively. With such a joint multi-task (FR and SR) formulation, the SR optimisation is specifically guided to be more discriminative for the native LR facial imagery data.

Model Training and Deployment. The CSRI can be trained by the standard Stochastic Gradient Descent algorithm in an end-to-end manner. As the auxiliary and native LR data sets are highly imbalanced in size, the CSRI is trained in two steps for improving the model convergence stability: (1) The *synthetic LR SR-FR* branch is first pre-trained on a large auxiliary face data (CelebA [123]). (2) The whole CSRI network is then trained on both auxiliary and native LR data.

In deployment, the *native LR SR-FR* branch is utilised to extract the feature vectors for face image matching with the Euclidean distance metric.

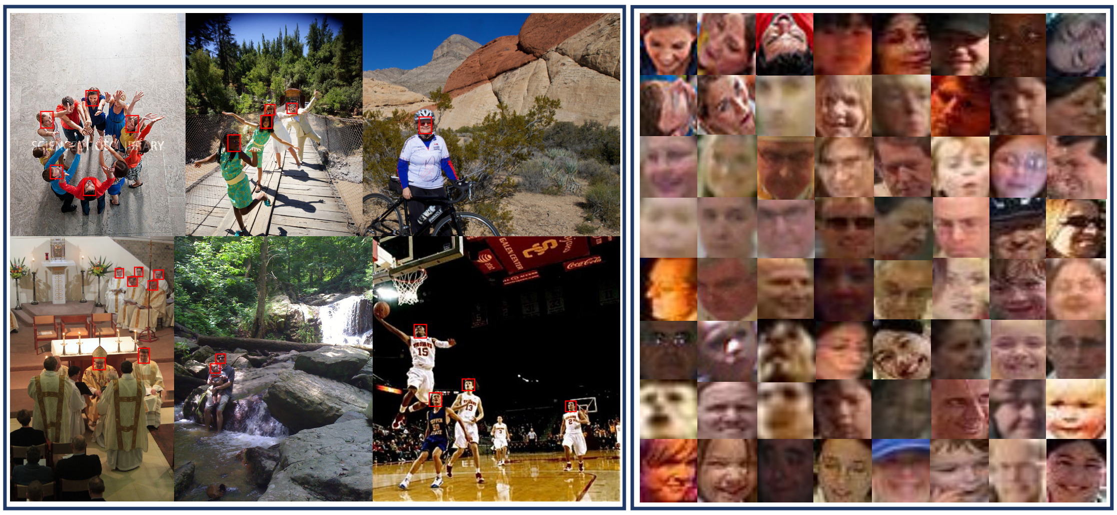

Figure 3.2: Example TinyFace images auto-detected in unconstrained images.

3.3 TinyFace: Low-Resolution Face Recognition Benchmark

3.3.1 Dataset Construction

Low-Resolution Criterion. To create a native LR face dataset, an explicit LR criterion is needed. As there is no existing standard in the literature, this thesis defines LR faces as those $\leq$32×32 pixels by following the tiny object criterion [185]. Existing FR datasets are all $>$100×100 pixels (Table 3.1).

Face Image Collection. The TinyFace dataset contains two parts, face images with *labelled* and *unlabelled* identities. The *labelled* TinyFace images were collected from the publicly available PIPA [225] and MegaFace2 [137] datasets, both of which provide unconstrained social-media web face images with large variety in facial expression/pose and imaging conditions. For the TinyFace to be realistic for LRFR test, this thesis applied the state-of-the-art HR-ResNet101 model [77] for automatic face detection, rather than manual cropping. Given the detection results, those faces with spatial extent larger than 32×32 were removed to ensure that all selected faces are of LR.

Face Image Filtering. To make a valid benchmark, it is necessary to remove the false face detections. We verified exhaustively every detection, which took approx. 280 person-hours, i.e. one labeller needs to manually verify detected tiny face images 8 hours/day consistently for a total of 35 days. Utilising multiple labellers introduces additional tasks of extra consistency checking across all the verified data by different labellers. After manual verification, all the remaining PIPA face images were then *labelled* using the identity classes available in the original data. As a result, we assembled 15,975 LR face images *with* 5,139 distinct identity labels, and 153,428 LR faces *without* identity labels. In total, we obtained 169,403 images of labelled and unlabelled faces. Fig. 3.2 shows some examples randomly selected from TinyFace.

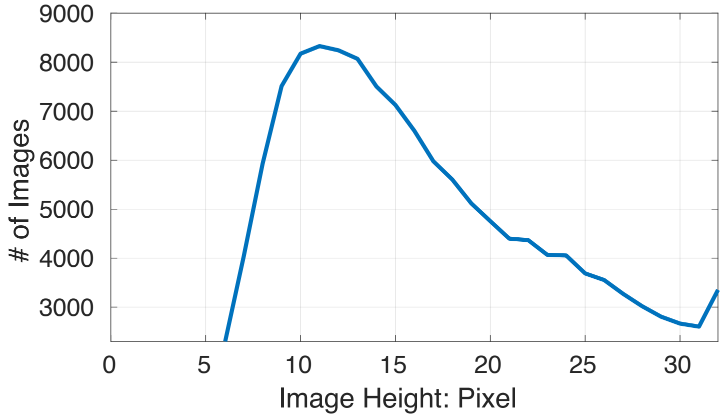

Figure 3.3: Distribution of face image height in TinyFace.

Face Image Statistics. Table 3.1 summarises the face image statistics of TinyFace in comparison to 9 existing FR benchmarks. Fig. 3.3 shows the distribution of TinyFace height resolution, ranging from 6 to 32 pixels with the average at 20. In comparison, existing benchmarks contain face images of ≥100 in average height, a ≥5× higher resolution [1].

3.3.2 Evaluation Protocol

Data Split. To establish an evaluation protocol on the TinyFace dataset, it is necessary to first define the training and test data partition. Given that both training and test data require labels with the former for model training and the latter for performance evaluation, we divided the

[1] The dataset can be downloaded at `https://qmul-tinyface.github.io/`

Table 3.1: Statistics of popular FR benchmarks.

Benchmark	Mean Height	# Identity	# Image
LFW [80]	119	5,749	13,233
VGGFace [144]	138	2,622	2.6M
MegaFace [91]	352	530	1M
CASIA [214]	153	10,575	494,414
IJB-A [96]	307	500	5,712
CelebA [123]	212	10,177	202,599
UMDFaces [8]	>100	8,277	367,888
MS-Celeb-1M [67]	>100	99,892	**8,456,240**
MegaFace2 [137]	252	**672,057**	4,753,320
TinyFace (Ours)	**20**	5,139	169,403

Table 3.2: Data partition and statistics of TinyFace.

Data	All	Training Set	Test Set		
			Probe	Gallery Match	Gallery Distractor
# Identity	5,139	2,570	2,569	2,569	Unknown
# Image	169,403	7,804	3,728	4,443	153,428

5,139 known identities into two halves: one (2,570) for training, the other (2,569) for test. All the unlabelled distractor face images are also used as test data (Table 3.2).

Face Recognition Task. In order to compare model performances on the MegaFace benchmark [137], we adopt the same face identification (1:N matching) protocol as the FR task for the TinyFace. Specifically, the task is to match a given probe face against a gallery set of enrolled face imagery with the best result being that the gallery image of a true-match is ranked at top-1 of the ranking list. For this protocol, we construct a probe and a gallery set from the test data as follows: (1) For each test face class of multiple identity labelled images, we randomly assigned half of the face images to the probe set, and the remaining to the gallery set. (2) We placed all the unlabelled disctractor images (with unknown identity) into the gallery set for enlarging the search space therefore presenting a more challenging task, similar to MegaFace [137]. The image and identity statistics of the probe and gallery sets are summarised in Table 3.2.

Performance Metrics. For FR performance evaluation, we adopt three metrics: the *Cumulative*

Matching Characteristic (CMC) curve [96], the *Precision-Recall* (PR) curve [187], and mean Average Precision (mAP). Whilst CMC measures the proportion of test probes with the true match at rank k or better, PR quantifies a trade-off between precision and recall per probe with the aim to find all true matches in the gallery [83]. To summarise the overall performance, we adopt the *mean Average Precision* (mAP), i.e. the mean value of average precision of all per-probe PR curves.

3.3.3 Training vs Testing Data Size Comparison

TinyFace is the largest native LR web face recognition benchmark (Table 3.1). It is a challenging task due to very LR face images (5× less than other benchmarks) with large variations in illumination, facial pose/expression, and background clutters. These factors represent more realistic real-world low-resolution face images for model robustness and effectiveness test.

In terms of training data size, TinyFace is smaller than some existing HR FR model *training* datasets, notably the MegaFace2 of 672,057 IDs. It is much more difficult to collect *natively* LR face images with label information. Unlike celebrities, there are much less facial images of known identity labels from the general public available for model training.

In terms of testing data size, on the other hand, the face identification *test* evaluation offered by the current largest benchmark MegaFace [91] contains *only 530 test face IDs* (from FaceScrub [138]) and 1 million gallery images, whilst TinyFace benchmark consists of 2,569 test IDs and 154,471 gallery images. Moreover, in comparison to LFW benchmark there are 5,749 face IDs in the LFW designed originally for 1:1 verification test [80], however a much smaller gallery set of 596 face IDs of LFW were adopted for 1:N matching test (open-set) with 10,090 probe images of which 596 true-matches (1-shot per ID) and 9,494 distractors [12]. Overall, TinyFace for 1:N test data has 3∼4× more test IDs than MegaFace and LFW, and 15× more distractors than LFW 1:N test data.

3.4 Experiments

In this section, we presented experimental analysis on TinyFace, the *only* large scale native LRFR benchmark, by three sets of evaluations: **(1)** Evaluation of generic FR methods *without* considering the LR challenge. We adopted the state-of-the-art deep learning FR methods (Sec. 3.4.1); **(2)** Evaluation of LRFR methods. For this test, we applied super-resolution deep learning techniques in addition to the deep learning FR models (Sec. 3.4.2); **(3)** Component analysis of the proposed CSRI method (Sec. 3.5).

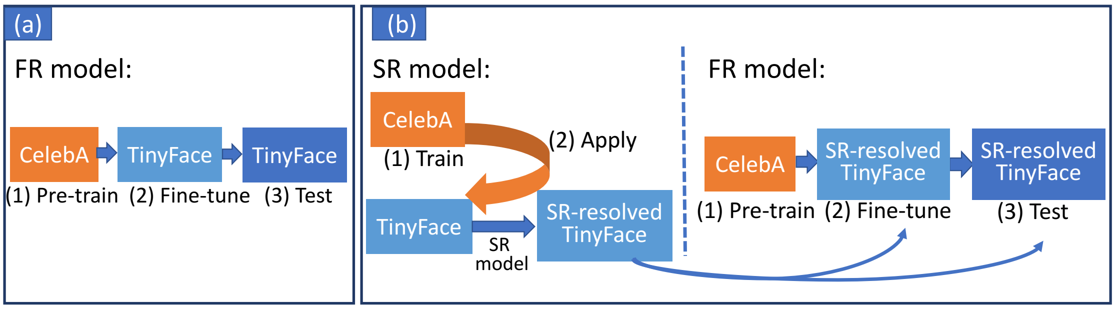

Figure 3.4: An overview of training (a) generic FR models and (b) low-resolution FR models (Independent training of Super-Resulotion (SR) and FR models).

3.4.1 Evaluation of Generic Face Recognition Methods

Table 3.3: *Generic* FR evaluation on TinyFace (Native LR face images).

Metric (%)	Rank-1	Rank-20	Rank-50	mAP
DeepID2 [174]	17.4	25.2	28.3	12.1
SphereFace [122]	22.3	35.5	40.5	16.2
VggFace [144]	30.4	40.4	42.7	23.1
CentreFace [199]	**32.1**	**44.5**	**48.4**	**24.6**

In this test, four representative deep FR models including DeepID2 [174], VggFace [144], CentreFace [199] and SphereFace [122] were evaluated. For model optimisation, a given FR model on the CelebA face data [123] was first trained before fine-tuning on the TinyFace training set[2] (see Fig. 3.4(a)). The parameter settings suggested by the original authors was adopted.

Results. Table 3.3 shows that the FR performance by any model is significantly inferior on TinyFace than on existing high-resolution FR benchmarks. For example, the best performer CentreFace yields Rank-1 32.1% on TinyFace *versus* 65.2% on MegaFace [91], i.e. more than half performance drop. This suggests that the FR problem is more challenging on natively unconstrained LR images.

Native vs Synthetic LR Face Images. For more in-depth understanding on *native* LRFR, we further compared with the FR performance on *synthetic* LR face images. For this purpose, we created a synthetic LR face dataset, which we call ***SynLR-MF2***, using 169,403 HR MegaFace2

[2] The SphereFace method fails to converge in fine-tuning on TinyFace even with careful parameter selection. The CelebA-trained SphereFace model was hence deployed .

Table 3.4: Native (TinyFace) vs. synthetic (SynLR-MF2) LR face recognition.

FR Model	Dataset	Rank-1	Rank-20	Rank-50	mAP
VggFace [144]	TinyFace	30.4	40.4	42.7	23.1
	SynLR-MF2	**34.8**	**46.8**	**49.4**	**26.0**
CentreFace [199]	TinyFace	32.1	44.5	48.4	24.6
	SynLR-MF2	**39.2**	**63.4**	**70.2**	**31.4**

images [137]. Following the data distribution of TinyFace (Table 3.2), we randomly selected 15,975 images from 5,139 IDs as the labelled test images and further randomly selected 153,428 images from the remaining IDs as the unlabelled distractors. We down-sampled all selected MegaFace2 images to the average size (20×16) of TinyFace images. To enable a like-for-like comparison, we made a random data partition on SynLR-MF2 same as TinyFace (see Table 3.2).

Table 3.4 shows that FR on synthetic LR face images is a less challenging task than that of native LR images, with a Rank-20 model performance advantage of 6.4% (46.8-40.4) by VggFace and 18.9% (63.4-44.5) by CentreFace. This difference is also visually indicated in the comparison of native and synthetic LR face images in a variety of illumination/pose and imaging quality (Fig. 3.5). This demonstrates the importance of TinyFace as a native LRFR benchmark for testing more realistic real-world FR model performances.

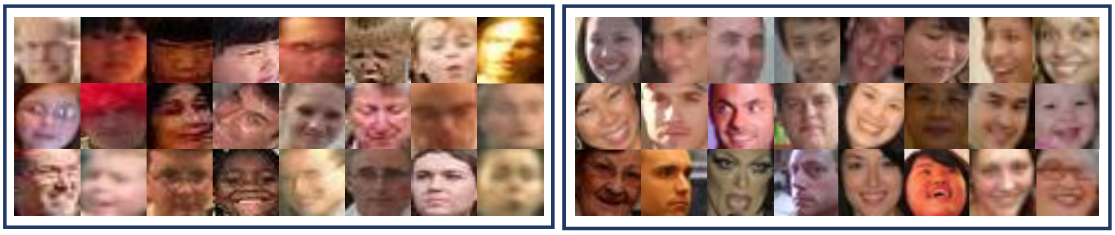

Figure 3.5: Comparison of (left) native LR face images from TinyFace and (right) synthetic LR face image from SynLR-MF2.

3.4.2 Evaluation of Low-Resolution Face Recognition Methods

In this evaluation, we explored the potential of contemporary super-resolution methods in addressing the LRFR challenge. To compare with the proposed CSRI model, we selected three representative deep learning generic-image SR models (SRCNN [44], VDSR [92] and DRRN [180]), and one LRFR deep model RPCN [193] (also using SR). We trained these SR models on the CelebA images [123] (202,599 LR/HR face pairs from 10,177 identities) with the authors sug-

FR	Method		Rank-1	Rank-20	Rank-50	mAP
CentreFace	No		**32.1**	**44.5**	**48.4**	**24.6**
	SR	SRCNN [44]	28.8	38.6	42.3	21.7
		VDSR [92]	26.0	34.5	37.7	19.1
		DRRN [180]	29.4	39.4	43.0	22.2
VggFace	No		**30.4**	**40.4**	**42.7**	**23.1**
	SR	SRCNN [44]	29.6	39.2	41.4	22.7
		VDSR [92]	28.8	38.3	40.3	22.1
		DRRN [180]	29.4	39.8	41.9	22.4
RPCN [193]			18.6	25.3	27.4	12.9
CSRI (Ours)			**44.8**	**60.4**	**65.1**	**36.2**

Table 3.5: Native *Low-Resolution* FR evaluation on TinyFace.

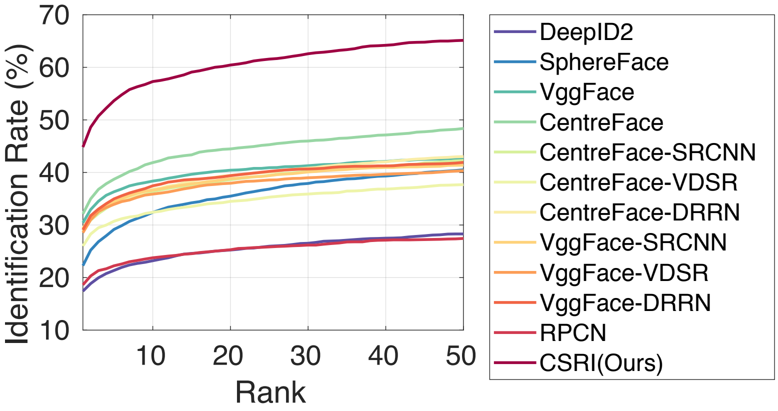

Figure 3.6: Performance comparison of different methods in CMC curves on the TinyFace dataset.

gested parameter settings for maximising their performance in the FR task (see Fig. 3.4(b)). We adopted the CentreFace and VggFace (top-2 FR models, see Table 3.3) for performing FR model training and test on super-resolved faces generated by any SR model. Since the RPCN integrates SR with FR in design, we used both CelebA and TinyFace data to train the RPCN for a fair comparison.

Results. Table 3.5 Fig. 3.6 show that: **(1)** All SR methods *degrade* the performance of a deep learning FR model. One possible explanation is that the artifacts and noise introduced in super-resolution are likely to hurt the FR model generalisation (see Fig. 3.7). This suggests that applying SR as a separate process in a simplistic approach to enhancing LRFR does not offer any benefit, and even is more likely a hindrance. **(2)** The RPCN yields the worst performance

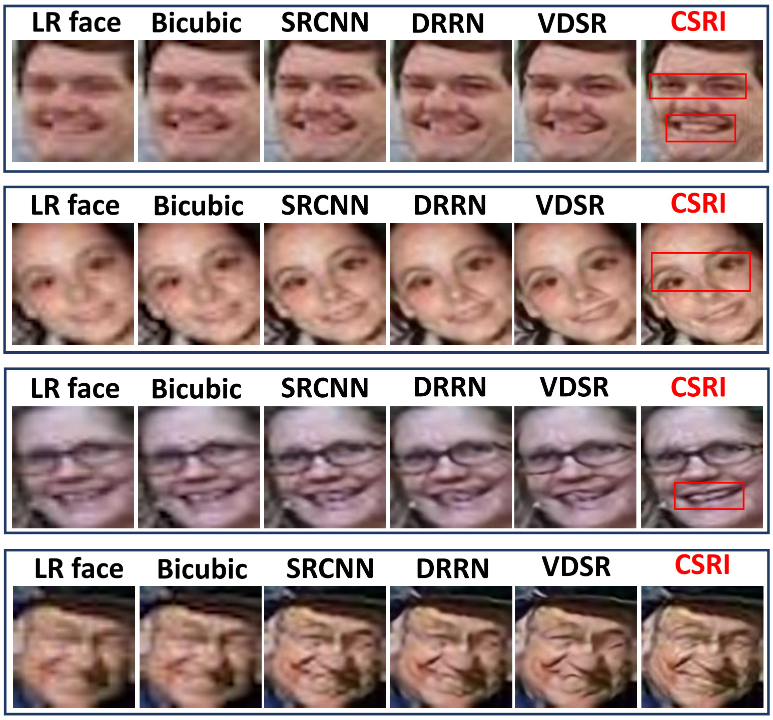

Figure 3.7: Examples of super-resolved faces. The enhanced facial feature in the resolved image is denoted by the red box.

although it was specially designed for LR face recognition. The possible reason is two-fold: (a) This method exploits the SR as model pre-training by design, which leads to insufficient FR supervision in the ID label guided model fine-tuning. (b) Adopting a weaker base network with 3 conv layers. These results suggest that existing methods are ineffective for face recognition on natively low-resolution images and when the test gallery population size becomes rather large. **(3)** The CSRI outperforms significantly all the competitors, e.g. the Rank-1 recognition performance gain by CSRI over CentreFace is significant at 12.7% (44.8-32.1). This shows the advantage of the CSRI model design in enabling FR on natively LR face images over existing generic FR models. However, despite the improvements by the proposed CSRI, due to the lack of direct supervision in the high-resolution image pixel space, the super-resolved images lack of the fidelity expected in real high-resolution domain (see Fig. 3.7), which may hinder the further down-stream tasks, e.g., facial analysis.

3.5 Component Analysis of CSRI

To better understand the CSRI's performance advantage, the individual model components were evaluated on the TinyFace benchmark by incrementally introducing individual components of the CSRI model.

SR-FR joint learning was examined in comparison to SR-FR independent learning. For fair comparison, the VDSR [92] and CentreFace [199], which are adopted the components of CSRI, were used. For SR-FR joint learning, the CSRI *synthetic LR SR-FR* branch was first trained on the CelebA data, followed by fine-tuning the FR part on TinyFace training data. Table 3.6 shows that SR-FR joint learning has a Rank-1 advantage of 10.1% (36.1-26.0) and 4.0% (36.1-32.1) over SR-FR independent learning and FR only (i.e. CentreFace in Table 3.3), respectively. This suggests the clear benefit of SR-FR joint learning due to the enhanced compatibility of SR and FR components obtained by end-to-end concurrent optimisation.

SR-FR	Rank-1	Rank-20	Rank-50	mAP
Independent Learning	26.0	34.5	37.7	19.1
Joint Learning	**36.1**	**49.8**	**54.5**	**28.2**

Table 3.6: Joint vs. independent learning of super-resolution and face recognition.

Complement SR learning was evaluated by comparing the full CSRI with the above SR-FR joint learning. Table 3.7 shows a Rank-1 boost of 8.7% (44.8-36.1), another significant benefit from the complement SR learning.

CSR	Rank-1	Rank-20	Rank-50	mAP
✗	36.1	49.8	54.5	28.2
✓	**44.8**	**60.4**	**65.1**	**36.2**

Table 3.7: Effect of complement super-resolution (CSR) learning.

3.6 Summary

This chapter presents the joint learning of Complement Super-Resolution and face Identity (CSRI) in an end-to-end trainable neural network architecture. By design, the proposed method differs significantly from most existing FR methods that assume high-resolution good quality facial imagery in both model training and testing, whereas ignoring the more challenging tasks in typical unconstrained low-resolution web imagery data.

It is shown that the proposed CSRI has significant performance advantage on the native low-resolution face recognition task. Extensive comparative evaluations show the superiority of CSRI over a range of state-of-the-art face recognition and super-resolution deep learning methods when tested on the newly introduced TinyFace benchmark. The more detailed CSRI component analysis provides further insights on the CSRI model design.

Chapter 4

Interpretable Low-Resolution Face Recognition

4.1 Introduction

It is an important computer vision task to recover the high-resolution faces from native low-resolution images, which enables a variety of downstream applications. In addition to recognising the identity of low-resolution faces, for example, super resolved faces could also benefit the facial image analysis [182, 23, 9] task. It is significant for many computer vision applications in business, law enforcement, and public security [142], but the model performance often degrades significantly when the face image resolution is very low. Chapter 3 explores how to adopt the domain-adaptation technique to optimise the super resolution model, to benefit the native low-resolution facial matching. However, the qualitative results show that the resolved images fail to match the fidelity expected at the higher resolution, which are perceptually unsatisfying and may hinder the accurate facial analysis.

This chapter aims to resolve high-fidelity faces from the realistic low-resolution images. As discussed in Chapter 3, existing state-of-the-art image SR models [35, 219, 238] mostly learn the low-to-high resolution mapping from paired *artificial LR* and HR images. The artificial LR images are usually generated by down-sampling the HR counterparts (Fig 1.2(a)). With this paradigm, existing supervised deep learning models (e.g. CNNs) can be readily applied. However, this is at a price of poor model generalisation to real-world *genuine LR* facial images, e.g. surveillance imagery captured in poor circumstances. This is because genuine LR data have rather different *imaging characteristics* from artificial LR images, often coming with *additional* unconstrained motion blur, noise, corruption, and image compression artefacts. (Fig. 4.2). This causes the distribution discrepancy between training data (artificial LR imagery) and test data (genuine LR imagery) which attributes to poor model generalisation, also known as the domain shift problem [142].

Unsupervised domain adaptation (UDA) methods are possible solutions considering genuine LR and HR images as two different domains. UDA techniques have achieved remarkable success [237, 76, 181, 135, 17, 215, 95, 120]. A representative modelling idea is to exploit cycle consistency loss functions between two unpaired domains (Fig 1.2(b)) [237, 215, 95]. It is usually implemented with a Generative Adversarial Network (GAN) - a generator (e.g., a CNN network) is used to map an image from one domain to the target domain, with a discriminator to distinguish the mapped images from the real target images. More specifically in the context of UDA with cycle consistency loss functions, a CNN is used to map an image from one domain to the other, which is further mapped back by another CNN. With such an encoder-decoder like architecture, one can form a reconstruction loss *jointly* for both CNN models *without* the need for paired images in each domain. The two CNN models can be trained end-to-end, inputting an image and outputting a reconstructed image per domain. This idea has been attempted in [20] for super-resolving genuine LR facial imagery.

Using such cycle consistency for unsupervised domain adaptation has several adverse effects. The reconstruction loss is applicable only to the concatenation of two CNN models. This exacerbates the already challenging task of domain adaptation training. In the context of native face super resolution, the genuine LR and HR image domains have significant differences in both image resolution and imaging conditions. Compared to a single CNN, the depth of a concatenated CNN-CNN model is effectively doubled. Existing UDA models apply the cycle consistency loss supervision at the final output of the second CNN, and propagate the supervision back to the first CNN. This gives rise to extra training difficulties in the form of *vanishing* gradients [104, 26]. In addition, jointly training two connected CNN models has to be conducted very carefully, along with the difficulty of training GAN models [59]. Moreover, the first CNN (the target model) takes responsibility of both characteristic consistifying and low-to-high resolution mapping, which further increases the model training difficulty dramatically.

This chapter solves the problem of super-resolving genuine LR facial images with high fidelity by formulating a ***Characteristic Regularisation*** (CR) method (Fig 1.2(c)). In contrast to conventional image SR methods, this model particularly leverages the unpaired genuine LR images in order to take into account their characteristics information for facilitating model optimisation. Unlike cycle consistency based UDA methods, the proposed model instead *leverage the artificial LR images as regularisation target* in order to separately learn the tasks of characteristic consistifying and image super-resolution. Specifically, the proposed design performs multi-task learning with the auxiliary task as *characteristic consistifying* (CC) for transforming genuine LR images into the artificial LR characteristics, and the main/target task as *image SR* for super-resolving both regularised and down-sampled LR images concurrently. Since there is no HR images coupled with genuine LR images, it is considered to align pixel content in the LR space by down-sampling the super-resolved images. This avoids the use of cycle consistency and their learning limitations. To make the super-resolved images with good facial identity information, an unsupervised semantic adaptation loss is formulated by aligning with the face recognition feature distribution of auxiliary HR images.

The proposed CR method can be understood from two perspectives: (i) As splitting up the whole system into a model for image characteristic consistifying and a model for image SR. With the former model taking the responsibility of solving the characteristic discrepancy, the SR model can better focus on learning the resolution enhancement. This is in a divide-and-conquer principle. (ii) As a deeply supervised network [104], providing auxiliary supervision improves accuracy and convergence speed [179]. In the case of super resolution specifically, it allows for better and more efficient pre-training of SR module using paired artificial LR and HR images, pre-training of CC module by genuine and artificial LR images, and fast convergence in training the full CC+SR model.

4.2 Methodology

The aim is to obtain a super-resolved HR image I^{sr} from an input genuine LR facial image I^{lr} with unknown noise characteristics. In real-world applications, there is no access to the corresponding HR counterparts for I^{lr}. This prevents the *supervised* model training of low-to-high resolution mapping between them. One solution is to leverage auxiliary HR facial image data (which is the counterparts of the artificially down-sampled LR data) I^{ahr}. Firstly, here is an overview of existing image SR models before introducing the proposed characteristic regularisation method.

4.2.1 Facial Image Super-Resolution

Given auxiliary HR facial images I^{ahr}, one can easily generate corresponding LR images I^{alr} by down-sampling. With such paired data, a common supervised image SR CNN model can be optimised by some pixel alignment loss constraint such as the Mean-Squared Error (MSE) between the resolved and ground-truth images [103]:

$$\mathcal{L}_{\text{sr}} = \|I^{\text{ahr}} - \phi_{\text{sr}}(I^{\text{alr}})\|_2^2. \tag{4.1}$$

The learned non-linear mapping function ϕ_{sr} can be then applied to super-resolve LR test images as:

$$I^{\text{asr}} = \phi_{\text{sr}}(I^{\text{alr}}). \tag{4.2}$$

This model deployment expects the test data with similar distribution as the artificial LR training facial images. If feeding genuine LR images, the model may generate much poor results due to the domain gap problem.

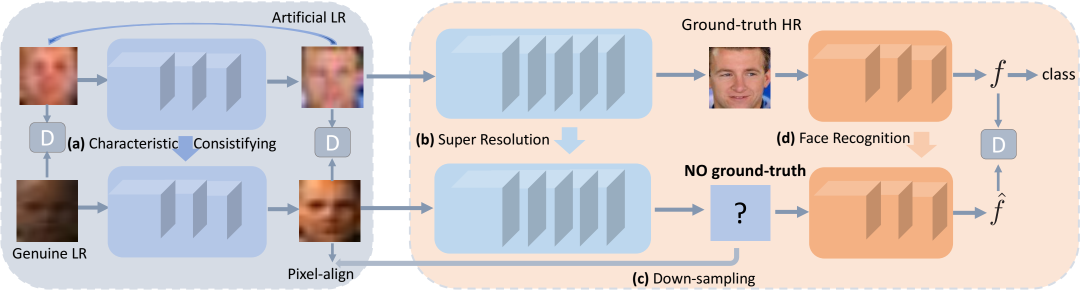

Figure 4.1: An overview of the proposed *Characteristics Regularisation* (CR) approach for super-resolving genuine LR facial imagery data. The CR model performs multi-task learning. **(a)** The auxiliary task is *characteristic consistifying* in order to transform genuine LR images into the artificial LR characteristics. **(b)** The main task is *image SR* allowing for super-resolving both regularised and down-sampled artificial LR images concurrently. **(c)** Due to no paired HR images, the model aligns pixel content in the LR space by down-sampling the super-resolved images. **(d)** To make the super-resolved images with good facial identity information, an unsupervised semantic adaptation loss term is formulated in the adversarial learning spirit, w.r.t. a supervised face recognition model trained on auxiliary HR images.

4.2.2 Characteristics Regularisation

To address the domain gap in SR, the model takes a divide-and-conquer strategy: first characteristic consistifying, then image super-resolving (Fig.4.1 gives an overview of the proposed framework). Specifically, a given genuine LR image is first transformed into that with similar appearance characteristics as artificial LR images. Then, the SR model is able to better perform image super-resolving. To that end, the unsupervised GAN learning framework is exploited

[59]. The objective is to learn a model that can synthesise facial images indistinguishable from artificial LR data with condition on genuine LR input.

Formally, the Characteristics Regularisation (CR) GAN model consists of a discriminator D that is optimised to distinguish whether the input is an artificial down-sampled LR or not, and a characteristics regularisor ϕ_{cr} that transforms a genuine LR input I^{lr} to fool the discriminator to classify the transformed $\phi_{\text{cr}}(I^{\text{lr}})$ as an artificial image. The objective function can be written as:

$$\begin{aligned}\mathcal{L}_{\text{gan}} = \mathbb{E}_{I^{\text{alr}}}[\log D(I^{\text{alr}})]+ \\ \mathbb{E}_{I^{\text{lr}}}[\log\left(1 - D(\phi_{\text{cr}}(I^{\text{lr}}))\right],\end{aligned} \tag{4.3}$$

where the characteristics regularisor ϕ_{cr} tries to minimise the objective value against an adversarial discriminator D that instead tries to maximise the value. The optimal adaptation solution is obtained as:

$$G^* = \arg\min_{\phi_{\text{cr}}}\max_{D}\mathcal{L}_{\text{gan}}. \tag{4.4}$$

To better connect the characteristics regularisation ϕ_{cr} with the super-resolving ϕ_{sr} module, an end-to-end training for the auxiliary artificial LR branch is enabled by additionally learning a mapping from the down-sampled artificial LR images to the transformed pseudo genuine LR counterparts. More specifically, *pseudo* genuine LR images are first generated by an inverse process of CR, i.e. transforming an artificial LR image to fool the discriminator to classify the transformed $\tilde{\phi}_{\text{cr}}(I^{\text{alr}})$ as a genuine LR image:

$$\begin{aligned}\arg\min_{\tilde{\phi}_{\text{cr}}}\max_{\tilde{D}}\mathbb{E}_{I^{\text{lr}}}[\log\tilde{D}(I^{\text{lr}})]+ \\ \mathbb{E}_{I^{\text{lr}}_{\text{aux}}}[\log\left(1 - \tilde{D}(\tilde{\phi}_{\text{cr}}(I^{\text{alr}}))\right].\end{aligned} \tag{4.5}$$

This is learned independently. Then, ϕ_{cr} can be jointly optimised by a loss formula as:

$$\mathcal{L}_{\text{cr}} = ||I^{\text{alr}} - \phi_{\text{cr}}(\tilde{\phi}_{\text{cr}}(I^{\text{alr}}))||_2^2 + \lambda\mathcal{L}_{\text{gan}}, \tag{4.6}$$

where λ is a weight hyper-parameter. $\lambda = 0.2$ is set in the experiment. It was found that this design improves the stability of end-to-end joint training for ϕ_{cr} and ϕ_{sr}.

4.2.3 Super-Resolving Regulated Images

If the CR module is perfect in characteristic consistifying, the SR module ϕ_{sr} trained on the auxiliary facial data can be directly applied. However, this is often not the truth in reality. So, it is helpful to further fine-tune ϕ_{sr} on the regulated data $\phi_{\text{cr}}(I^{\text{lr}})$. To do this, the model needs

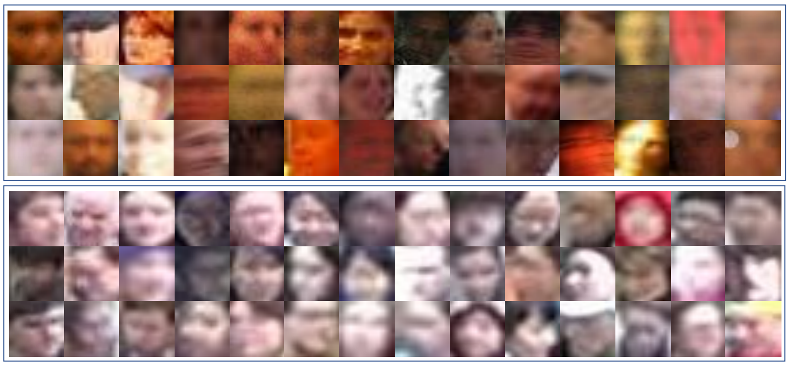

Figure 4.2: Examples of genuine facial images randomly sampled from TinyFace-s (**top**) and LR-DukeMTMC (**bottom**).

to address the problem of lacking HR supervision. Instead of leveraging the conventional cycle consistency idea, the model adopts a simple but effective pixel-wise distance constraint. The intuition is that, a good super-resolved image output, after down-sampling, should be close to the LR input. By applying this cheap condition, there is no need to access the unknown HR ground-truth. Formally, this SR loss function is designed for regulated LR images as:

$$\mathcal{L}_{\text{cr-sr}} = ||f_{\text{DS}}\Big(\phi_{\text{sr}}\big(\phi_{\text{cr}}(I^{\text{lr}})\big)\Big) - \phi_{\text{cr}}(I^{\text{lr}})||_2^2, \tag{4.7}$$

where f_{DS} refers to the down-sampling function.

4.2.4 Unsupervised Semantic Adaptation

Apart from visual fidelity, the SR output is also required to be semantically meaningful with good identity information. To this end, an unsupervised semantic adaptation loss term is formed in the adversarial learning spirit. The idea is to constrain the perceptual *feature* distribution of super-resolved facial images by matching the feature statistics of auxiliary HR images I^{ahr}. It is formally written as:

$$\begin{aligned}\mathcal{L}_{\text{cr-gan}} = \; & \mathbb{E}_{I^{\text{ahr}}}[\log D'(\phi_{\text{fr}}(I^{\text{ahr}}))] + \\ & \mathbb{E}_{\phi_{\text{sr}}(\phi_{\text{cr}}(I^{\text{lr}}))}\Big[\log\Big(1 - D'\big(\phi_{\text{fr}}(\phi_{\text{sr}}(\phi_{\text{cr}}(I^{\text{lr}})))\big)\Big)\Big],\end{aligned} \tag{4.8}$$

where ϕ_{fr} is a CentreFace [199] based feature extractor pre-trained with I^{ahr}. This loss is unsupervised without the need for identity labels of *genuine* LR training images. Compared to image based GAN loss, it is found more efficient and easier to train in a low-dimension feature space.

4.2.5 Model Training and Inference

Due to introduction of characteristic regularisation in the middle of the proposed full model, more effective model training is enabled. It facilitates a two-staged training strategy. In the first stage, the CNN for image SR is pre-trained on the auxiliary LR-HR paired facial data, the CentreFace model on HR images, and the CNN for characteristic regularisation and the inverse CR on unpaired genuine and artificial LR images in parallel. In the second stage, the cascaded CR and SR CNNs are fine-tuned together on all the training data.

CNN for image super-resolution. The image SR model ϕ_{sr} as [103] is trained by deploying the pixel-wise MSE loss function (Eq (4.1)). This model training benefits from the normal adversarial loss for achieving better perceptual quality. Other existing image SR methods [92, 45] can be readily considered in the framework.

CNN for characteristic regularisation. The CNN for characteristic regularisation ϕ_{cr} is trained by an adversarial loss and a pixel-wise loss jointly (Eq (4.6)).

Full model. In the second stage, both CNN models ϕ_{sr} and ϕ_{cr} are further fine-tuned jointly. The overall objective loss for training the full model is formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{sr}} + \lambda_{\text{cr}}\mathcal{L}_{\text{cr}} + \lambda_{\text{cr-sr}}\mathcal{L}_{\text{cr-sr}} + \lambda_{\text{cr-gan}}\mathcal{L}_{\text{cr-gan}}, \tag{4.9}$$

where $\lambda_{\text{cr}}, \lambda_{\text{cr-sr}}, \lambda_{\text{cr-gan}}$ are the weight parameters of the corresponding loss terms. In the experiment, it is set that $\lambda_{\text{cr}} = 0.06$, $\lambda_{\text{cr-sr}} = 0.01$, $\lambda_{\text{cr-gan}} = 0.03$ by cross-validation.

Model inference. Once trained, the full model is deployed for test, taking a genuine LR facial image as input, outputting a HR image.

4.3 Experiments

Datasets. For model performance evaluation, two real-world genuine LR facial image datasets were sampled from web social-media imagery and surveillance videos. Following [20], LR faces are defined as those with an average size of $\leq 16\times 16$ pixels. In particular, the web social-media based real-world face images were collected by assembling LR faces from the People In Photo Albums (PIPA) benchmark [225]. The extreme distorted facial images were further manually filtered out, making this dataset a subset of the TinyFace dataset introduced in Chapter 3, called TinyFace-s. Similarly, LR face images (small faces) were collected from a multi-target multi-camera tracking benchmark DukeMTMC [152] and built a surveillance video real-world face dataset, called LR-DukeMTMC. There are 8,641 and 7,044 face images in TinyFace-s and LR-DukeMTMC, respectively. All the face images were obtained by deploying the automatic face detector [78]. Non-face images were manually filtered out. These two new datasets consist of genuinely real-world LR facial images captured from unconstrained camera views under a large range of different viewing conditions such as expression, pose, illumination, and background clutter. Some randomly selected examples are shown in Fig 4.2.

Training and test data. To effectively train a competing model, both real-world genuine LR images and web auxiliary HR facial images are needed. For the former, 153,440 LR face images collected from the Wider Face benchmark [213] were used. This dataset offers rich facial images from a wide variety of social events, with a high degree of variability in scale, pose, lighting, and background. For the latter, the standard CelebA benchmark with 202,599 HR web facial images [175] was selected. Such a training set design ensures that each model can be trained with sufficiently diverse data to minimise the learning bias. For model test, the entire TinyFace-s and LR-DukeMTMC were utilised. Both datasets present significant test challenges, as they were drawn from unconstrained and independent data sources with arbitrary and unknown noise.

Performance evaluation metrics. Due to that there are *no* ground-truth HR data of *genuine* LR facial images, it is impossible to conduct pixel based performance evaluation and comparison. The Frechet Inception Distance (***FID***) [75] is utilised to assess the quality of resolved face images, similar to the state-of-the-art method [20]. Specifically, ***FID*** is measured by the Frechet Distance between two multivariate Gaussian distributions.

Implementation details. All the following experiments were performed in Tensorflow. The residual blocks [72] was used as the backbone unit of the network. In particular, 3 residual blocks were used in the net for the characteristics regularisation module ϕ_{cr} and $\tilde{\phi}_{\mathrm{cr}}$, and the SRGAN (3 groups containing 12/3/2 residual blocks, respectively. Resolution was increased 2 times across each group) [103] was further adapted for the facial SR module ϕ_{sr}. The adversarial discriminator for ϕ_{cr} and $\tilde{\phi}_{\mathrm{cr}}$ both consist of 6 residual blocks, followed by a fully connected layer. The adversarial discriminator D' for semantic adaptation consists of 5 fully connected layers. All LR images were sized at 16×16. The scale of real-world facial image super-resolution was 16 (4×4) times, i.e. the output size is 64×64. The learning rate was set to 10^{-4}, the batch size to 16. The SR module (ϕ_{sr} in Fig. 4.1) was pre-trained on CelebA face dataset with down-sampled artificial LR and HR image pairs for 100 epochs. And the characteristic consistifying module was trained with unpaired genuine and artificial LR images (down-sampled from CelebA dataset) for 130 epochs. The end-to-end full model was jointly trained by 10 epochs.

Dataset	TinyFace-s	LR-DukeMTMC
VDSR [92]	94.49	229.56
SRGAN [103]	103.85	232.38
FSRNet [35]	117.19	218.30
SICNN [224]	129.23	223.08
CycleGAN [237]	33.62	42.41
CSRI (Chapter 3)	104.68	240.99
LRGAN [20]	29.80	31.20
CR (proposed)	**23.09**	**25.56**

Table 4.1: Comparing the image quality on *genuine* LR facial image super-resolution. Metric: FID. **Lower is better.**

4.3.1 Test Genuine Low-Resolution Facial Images

Competitors. To evaluate the effectiveness of CR model for genuine facial image SR, the proposed model was compared with four *groups* of the state-of-the-art methods including, two generic image SR models (VDSR [92], SRGAN [103]), one image-to-image translation model (CycleGAN [237]), two *non*-genuine face SR model (FSRNet [35] and SICNN [224]), one UDA-based genuine face SR model (LRGAN [20]), and one facial identity-guided genuine SR model (CSRI, proposed in Chapter 3). Same as the proposed CR, CycleGAN, LRGAN and CSRI were trained using genuine LR images, while the others with artificial LR only as they need pixel-aligned LR and HR training image pairs.

Results. The results of these methods are compared in Table 4.1. There are observations as follows: **(1)** The proposed CR model achieves the best FID score among all the competitors, suggesting the overall performance advantage of the proposed approach on super-resolving genuine LR facial images. **(2)** Generic image SR methods (VDSR, SRGAN) perform the worst, as expected, although re-trained by the large-scale CelebA face data with artificial LR and HR image pairs. This is due to the big image characteristics difference between the source artificial LR and the target genuine LR images. **(3)** By considering the problem from image-to-image domain adaptation perspective, CycleGAN is shown to be superior than VDSR and SRGAN models. This is because of the domain gap problem. However, it is less optimal than modelling explicitly genuine LR face images in the SR process, as compared to the two specifically designed genuine LR facial image super-resolution models, CR and LRGAN. This is more so in surveillance videos (LR-DukeMTMC). **(4)** With the high-level facial identity constraint, CSRI cannot achieve satisfactory low-level visual fidelity in the pixel space. **(5)** Despite modelling facial prior explicitly, FSRGAN fails to improve meaningfully over generic SR methods (VDSR, SRGAN). This is due to the significant domain gap between the genuine and artificial LR facial images, leading to difficulty in inferring useful facial content and structural prior from the low-quality genuine LR images. **(6)** As a state-of-the-art model, LRGAN demonstrates its advantages over other models by learning explicitly the image degradation process. However, it is clearly outperformed by the proposed CR model. This suggests the overall performance advantages of the proposed method.

Qualitative evaluation. To conduct visual comparisons between different alternative methods, SR results of random genuine LR facial images are provided in Fig 4.3 . Overall, the visual examination is largely consistent with the numerical evaluation. Specifically, existing methods tend to generate images with severe blurry and artefact either globally (VDSR, SRGAN) or locally (CycleGAN, LRGAN). In contrast, CR can yield HR facial images with much better fidelity in most cases. This visually verifies the superiority of the proposed method in super-resolving genuine LR facial images.

Model complexity. The top-3 models (CR, LRGAN [20], and CycleGAN [237]) are compared in three aspects: (1) Model parameters: 2.7, 4.0, and 21 million; (2) Training time: 46, 72, and 81 hours; and (3) Per-image inference time: 7.5, 6.6, and 150 ms, using a Tesla P100 GPU. Therefore, CR is the most compact and most efficient.

4.3.2 Face Recognition on Genuine LR Face Imagery

The benefit of image SR is tested on low-resolution face recognition, on the TinyFace-s dataset. The CentreFace model trained on the auxiliary HR images and the CMC rank metrics are used. The results in Table 4.2 show that: (1) Directly using raw LR images leads to very poor recognition rate, due to lacking fine-grained facial trait details. (2) CR achieves the best performance gain as compared to all the strong competitors. (3) Interestingly, LRGAN gives a negative recognition margin, mainly due to introducing more identity-irrelevant enhancement despite good visual fidelity.

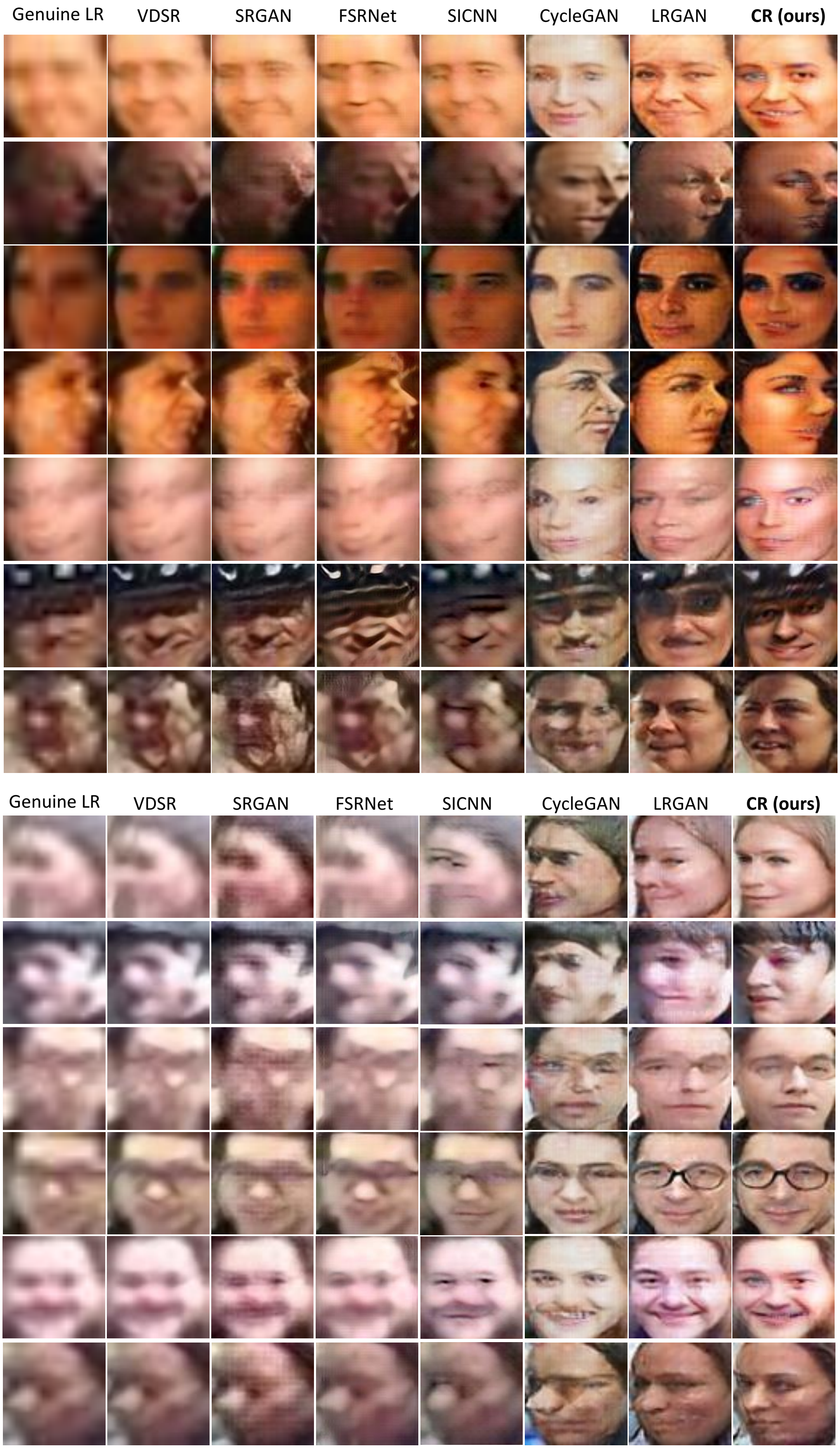

Figure 4.3: Examples of genuine LR image super-resolution on (**top**) TinyFace-s and (**bottom**) LR-DukeMTMC.

Dataset	Rank-1 (%)
VDSR [92]	25.45
SRGAN [103]	27.00
FSRNet [35]	26.50
SICNN [224]	28.85
CycleGAN [237]	25.12
CSRI (Chapter 3)	29.59
LRGAN [20]	21.99
CR (Ours)	**30.53**
Raw LR input	24.83

Table 4.2: Face recognition performance on super-resolved *genuine* LR images from TinyFace-s. Metric: Rank-1. **Higher is better.**

Metric	PSNR	SSIM
VDSR [92]	**26.31**	0.7918
SRGAN [103]	25.10	0.7873
FSRNet [35]	25.10	0.7234
SICNN [224]	26.10	0.7986
CycleGAN [237]	18.85	0.6061
CSRI (Chapter 3)	25.40	0.7388
LRGAN [20]	21.88	0.6869
CR (Ours)	25.50	**0.8184**

Table 4.3: Comparison of state-of-the-art methods on *artificial* LR facial image super-resolution. Dataset: Helen. Metric: PSNR & SSIM. **Higher is better.**

4.3.3 Test Artificial Low-Resolution Facial Images

For completeness, model performance was tested in artificial LR facial images as in conventional SR setting.

Model deployment. By design, the CR model is trained for super-resolving genuine LR facial imagery. However, it can be flexibly deployed *without* the characteristic regulation module, when artificial LR test images are given.

Dataset. In this evaluation the Helen face dataset [101] with 2,330 images was selected. The artificial LR test images were produced by bicubic down-sampling, as the conventional SR evaluation setting.

Metrics. For performance evaluation, the common Peak Signal-to-Noise Ratio (PSNR) and structure similarity index (SSIM) [197] were used. This is because, there are ground-truth HR images for pixel-level assessment in this case.

Results. Table 4.3 compares the performances on normal Helen LR facial images of CR and state-of-the-art SR methods. It is observed that CR can generate better results than all the competitors except VDSR for the PSNR metric. Interestingly, CR outperforms SRGAN which is actually the SR module in the proposed network. This implies that the model generalisation for conventional SR tasks can be improved by the proposed unsupervised SR learning objective (Eq (4.7)).

4.3.4 Component Analysis and Discussion

A series of model component analysis were conducted for giving insights to the CR performance.

Dataset	TinyFace-s	LR-DukeMTMC
W/O CR	133.30	190.73
W/ CR	**23.09**	**25.56**

Table 4.4: Effect of characteristics regularisation (CR). Metric: FID.

Characteristic regularisation. The effect and benefits of characteristic regularisation (CR) on model performance was evaluated. It is compared with a *baseline* which learns the SR module from genuine and artificial LR images *jointly*. The baseline model needs to fit heterogeneous input data distributions. The training loss function is $\mathcal{L}_{\text{base}} = \mathcal{L}_{\text{sr}} + \lambda_{\text{cr-sr}}\mathcal{L}_{\text{cr-sr}} + \lambda_{\text{cr-gan}}\mathcal{L}_{\text{cr-gan}}$. This allows for testing the exact influence of characteristic consistifying. Table 4.4 shows that CR plays a key role for enabling the model to super-resolve genuine LR facial images. Without CR, the model fails to properly accommodate the genuine data, partly due to an extreme modelling difficulty for learning such a cross-characteristics cross-resolution mapping

Dataset	TinyFace-s	LR-DukeMTMC
FID(G-LR, A-LR)	40.72	86.23
FID(R-LR, A-LR)	**19.49**	**24.32**

Table 4.5: Evaluation of the regulated LR images (R-LR). G-LR: Genuine LR images; A-LR: Artificial LR images.

The result of characteristic consistifying, i.e. the regulated LR images, was further examined. To this end, the FID between artificial and regulated LR images was measured, in comparison to that between artificial and genuine LR images. Table 4.5 shows that although regulated LR images match significantly better to artificial LR data than their genuine counterparts, the distribution difference remains. This suggests the necessity of fine-tuning the SR module on the regulated LR images (the second training stage).

Qualitative results are shown in Fig. 4.4. It is observed that compared to the genuine LR input, the regulated images have clearer contour of facial components, better lighting conditions and less blur, i.e. much closer to artificial LR data. This eases the subsequent SR job.

Dataset	TinyFace-s	LR-DukeMTMC
W/O SR-RI	111.01	113.70
W/ SR-RI	**23.09**	**25.56**

Table 4.6: Effect of the super-resolution fine-tuned on the regulated images (SR-RI). Metric: FID. **Lower is better.**

Super-resolving regulated images. In the second training stage, the SR module is fine-tuned for better super-resolving regulated LR images. The effect of this design was evaluated. Table 4.6 shows that the model performance drops noticeably without the proposed SR model fine-tuning on regulated LR images. This is consistent with the observation in Table 4.5.

Dataset	TinyFace-s	LR-DukeMTMC
W/O UL	25.30	26.11
W/ UL	**23.09**	**25.56**

Table 4.7: Effect of unsupervised loss (UL) for super-resolving regulated images. Metric: FID. **Lower is better.**

Recall that an unsupervised SR loss (Eq (4.7)) is introduced for regulated LR images due to no HR ground-truth. Pixel-wise alignment is considered in LR image space, without the need for cycle consistency. Its impact on the model performance was tested. Table 4.7 shows that applying this loss can clearly boost the fidelity quality of resolved faces. Also, it is found that the design makes the model training more stable. Further qualitative evaluation in Fig 4.5 shows that the unsupervised SR loss can help reduce the noise and distortion in SR, leading to visually more appealing results.

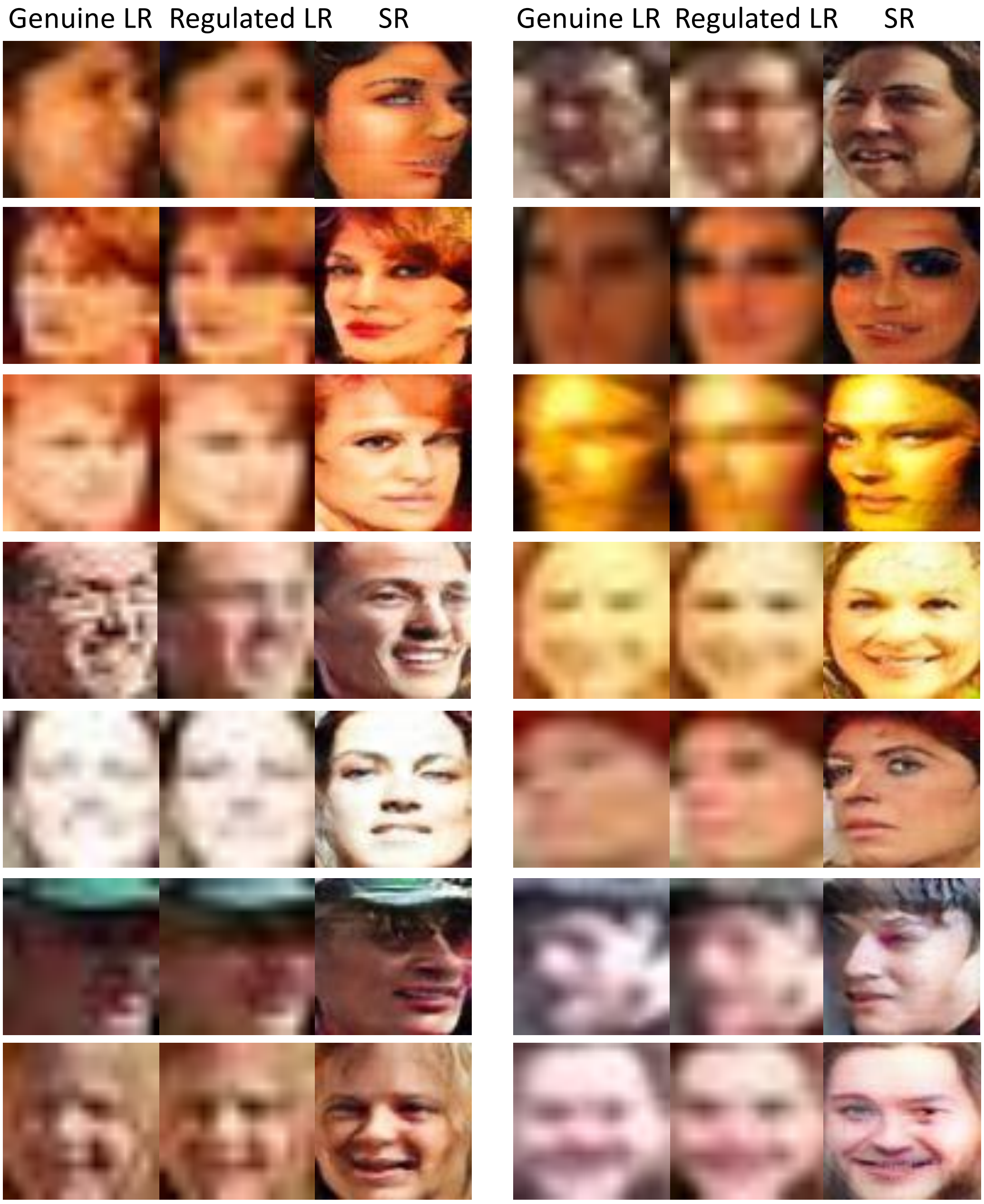

Figure 4.4: Genuine LR *vs.* regulated LR *vs.* resolved face images.

4.4 Summary

This chapter presents a Characteristic Regularised (CR) method for super-resolving genuine LR facial imagery with high fidelity. This differs from most SR studies focusing on artificial LR images with limited model generalisation on genuine LR data and UDA methods suffering ineffective training. In comparison, CR possesses the modelling merits of previous SR and UDA models end-to-end, solves both domain shift and ineffective model training, and simultaneously takes advantage of rich resolution information from abundant auxiliary training data. Extensive

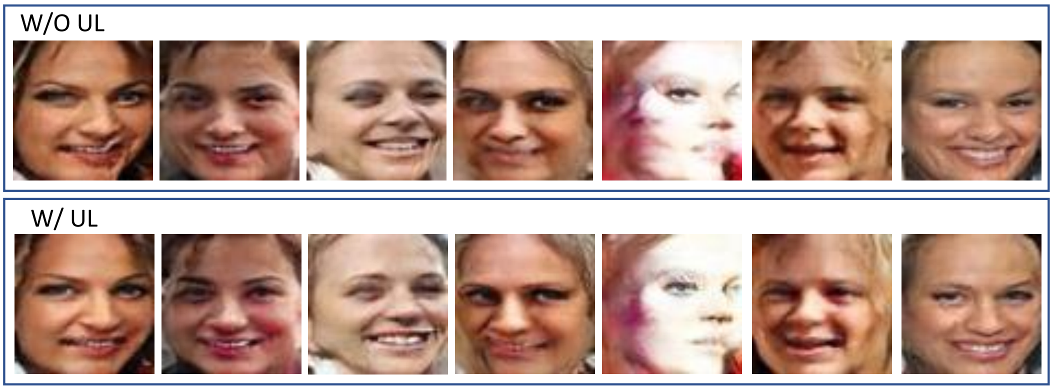

Figure 4.5: Visual examination: W/O *vs.* W/ unsupervised loss (UL) in super-resolving regulated images.

comparative experiments are conducted on both genuine and artificial LR facial images. The results show the performance and generalisation advantages of the proposed model over a variety of state-of-the-art image SR and UDA models. Detailed model component analysis is carried out for revealing the model formulation insights.

Chapter 5

Cross-Resolution Person Body Recognition

5.1 Introduction

Most existing methods assume that the probe and gallery images have similar and sufficiently high resolutions for person body recognition (person re-identification). However, due to unconstrained distances between cameras and pedestrians, person images are often captured at various resolutions. This *resolution mismatch* issue brings about significant challenges to body recognition. As low-resolution (LR) images contain much less identity detail information than high-resolution (HR) images, directly matching them across resolutions leads to substantial performance drop [86, 114]. For example, a standard person body recognition model [53] can suffer up to 19.2% Rank-1 rate drop when applied to cross-resolution person body recognition [114].

A number of cross-resolution body recognition methods have been developed for addressing the resolution mismatch problem [36, 86, 114, 195]. They are generally in two categories: (1) Learning resolution-invariant representation [36] and (2) Exploiting image super-resolution (SR) [86, 195]. The first category aims at learning a feature representation space shared by LR and HR images, but tends to lose fine-grained discriminative details due to being absent in LR images. The second category can solve this limitation often by adopting a multi-task joint learning framework which cascades SR and body recognition. However, this design suffers from ineffective model training due to significantly higher difficulty of backpropagating the gradients through such a cascaded thus heavier model [26]. As a consequence, the SR model is less compatible with person body recognition. Recently, Li et al. [114] combined the two approaches in a unified framework for improving cross-resolution body recognition performance, but still leaving the above problem unsolved.

This chapter addresses this problem by introducing a novel regularisation named *Inter-Task Association Critic* (INTACT). INTACT is an inter-task association mechanism that smooths out two unique tasks in joint learning. In design, it consists of a cascaded multi-task (SR & body recognition) network and an association critic network. The objective is to enhance the compatibility between SR and body recognition, i.e., super-resolving LR person images in such a way that the resolved images are suited for the body recognition model to perform identity matching in the HR image space. This is realised in two parts by INTACT: **(I)** The (unknown) inter-task association constraint is parameterised with a dedicated network, which enables it to be learned directly from the HR training data. **(II)** Once learned, serving in a critic role the association constraint is then applied to supervise the SR model. That means, the SR model training is further constrained to satisfy the learned inter-task association.

5.2 Methodology

Problem setting This section considers the cross-resolution person body recognition problem. The model training assumes a set of identity labelled high-resolution (HR) training images $\mathcal{D} = \{x_h, y\}$. The *objective* is to learn a person body recognition model that can tackle low-resolution (LR) query images in matching against a set of HR gallery images at test time.

The potential of image super-resolution (SR) is explored. The intuition is that an effective SR model should be able to recover the resolution of LR images so that the resolution mismatch problem between the query and gallery images can be well alleviated. To encourage that the SR model can generate such HR images that are more effective for person body recognition, a straightforward approach is to form a joint multi-task learning pipeline by cascading SR and body recognition sequentially, as exemplified in [86].

5.2.1 Joint Multi-Task Learning

Image super-resolution model To train a SR model, a set of LR-HR image pairs $\{(x_l, x_h)\}$ with pixel alignment are typically used. Often, such pairs are formed by downsampling the HR training images. Noted that in this chapter, the relatively higher resolution pedestrian images are captured under the unconstrained imaging conditions (surveillance), that is, both the higher resolution and the down-sampled lower resolution counterparts are the *genuine* images defined in the previous chapters. Therefore, in this chapter, the low resolution images are simply generated by down-sampling from the higher resolution pedestrian images. This chapter chooses the Generative Adversarial Network (GAN) model [59] for SR due to its promising performance [103].

GAN solves a min-max optimisation problem, where the discriminator D aims to distinguish the real HR from super-resolved images, while the generator G aims for generating super-resolved images that can fool the discriminator. The objective function can be defined as:

$$\mathcal{L}_{\text{gan}} = \mathbb{E}_{x_h}[\log D(x_h)] + \mathbb{E}_{x_l}[\log\big(1 - D(G(x_l))\big]. \tag{5.1}$$

More specifically, the generator G tries to minimise the objective value against an adversarial discriminator D that instead tries to maximise the value. The optimal solution is obtained as:

$$G^* = \arg\min_{G}\max_{D}\mathcal{L}_{\text{gan}}. \tag{5.2}$$

Person body recognition model With the training data $\mathcal{D}$, one can train any existing person body recognition model (e.g.[235]) by a softmax Cross-Entropy loss function:

$$\mathcal{L}_{\text{id}} = -\log(p_y), \tag{5.3}$$

where y is the ground-truth person identity of x_l, and p_y the prediction probability on class y.

Joint multi-task learning To build a joint multi-task learning pipeline, one can simply cascade SR and body recognition by using the output $G(x_l)$ of the SR as the input of body recognition model. The overall objective function is then formulated as:

$$\mathcal{L}_{\text{sr}} = \mathcal{L}_{\text{MSE}} + \lambda_g\mathcal{L}_{\text{gan}} + \lambda_c\mathcal{L}_{\text{id}}, \tag{5.4}$$

where $\mathcal{L}_{\text{MSE}}$ is the pixel-wise content loss, defined as $\mathcal{L}_{\text{MSE}} = \|x_h - G(x_l)\|_2^2$. λ_g and λ_c are weight parameters.

Limitation Despite a good solution for cross-resolution body recognition, this pipeline is intrinsically limited. This is due to significantly higher difficulty of backpropagating the gradients through two cascaded models [26, 104]. As a consequence, the SR model training is not properly constrained for maximising the person body recognition performance, i.e.the resulted SR model is not well compatible with the body recognition model.

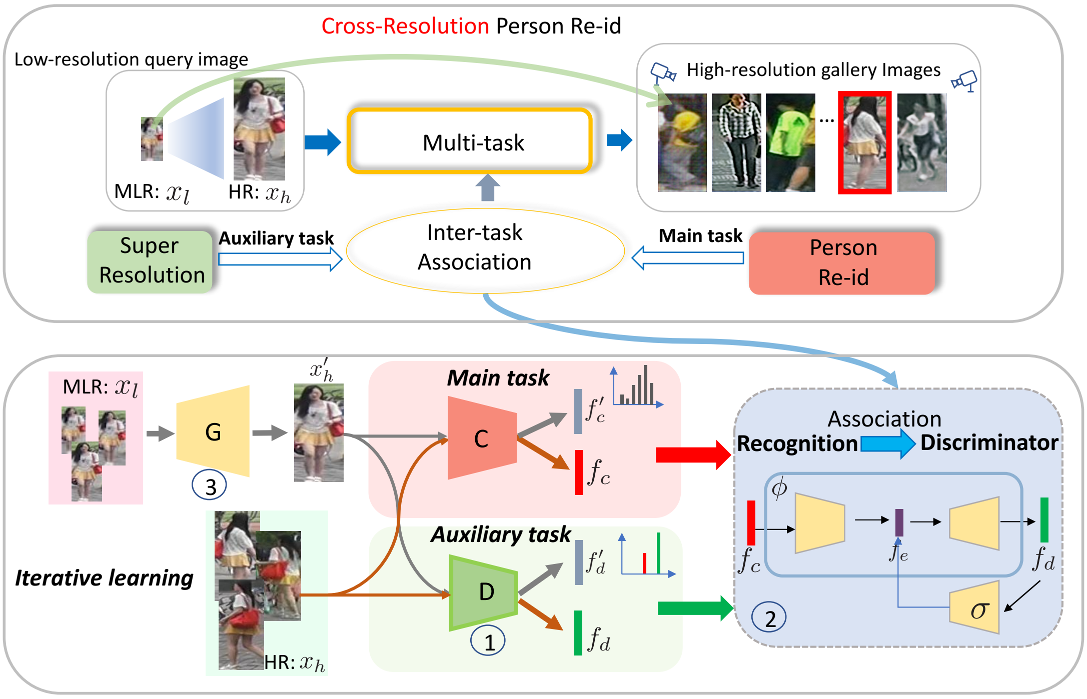

Figure 5.1: An overview of the proposed *Inter-Task Association Critic* (INTACT) method for cross-resolution person body recognition. Specifically, INTACT aims to recover the resolution of LR query images in such a way that the super-resolved images can be more accurately matched against HR gallery images for person body recognition. Joint learning of an image super-resolution (SR) model and a person body recognition model in a cascaded manner is *unsatisfactory*, due to higher difficulty of backpropagating the gradients through two cascaded models. INTACT offers a superior solution. The proposed model is trained alternatively in three steps: **(1)** Update the discriminator D of a GAN model; **(2)** Update the inter-task association module ϕ between the identity recognition representation $\boldsymbol{f}_{\text{c}}$ and discriminator representation $\boldsymbol{f}_{\text{d}}$. **(3)** Update the generator G of the GAN model, subject to the learned association regularisation on the identity recognition representation $\boldsymbol{f}'_{\text{c}}$ and discriminator representation $\boldsymbol{f}'_{\text{d}}$ of the resolved images. G: the generator of GAN model; D: the discriminator; C: the person body recognition model trained with cross-entropy loss.

5.2.2 Inter-Task Association Critic

To address this fundamental limitation, a novel regularisation, *Inter-Task Association Critic* (INTACT), is introduced. The key idea of INTACT is to exploit the intrinsic association between the SR and body recognition tasks as an extra optimisation constraint for boosting their joint learning and enhancing their compatibility. However, it is nontrivial to quantify such inter-task association which are typically complex and unknown a priori. To solve this issue, this association is parameterised using a dedicated network.

Specifically, a dedicated network is leveraged for representing the association from the main task (i.e.person body recognition) to the auxiliary task (i.e.SR). This forms the core element of INTACT. During model training, INTACT consists of two parts. In part **I**, it discovers the association using the native HR images. Concretely, it learns the association network residing between the discriminator and identity classification representations on the HR training images $\{x_h\}$. In part **II**, the learned association is then applied as a regularisation in the SR model training. Concretely, the discriminator and classification representations extracted from the resolved images are encouraged to satisfy the association constraint pre-learned from the true HR data. An overview of the INTACT method is depicted in Fig. 5.1.

Part I: Association Learning With a GAN model, the real-fake judgement task is represented by the feature activation $\boldsymbol{f}_{\mathrm{d}}$ of the discriminator. For the identity classification task, a large number of on-the-shelf person body recognition models can be adopted. This chapter exploits a very recent method presented in [235] for extracting identity classification feature $\boldsymbol{f}_{\mathrm{c}}$ to represent identity. The body recognition model is trained using HR images x_h independently to achieve the best identity representing power. It is trained one-off, frozen and served as an identity critic for the following model optimisation.

Given an input LR image x_l, the generator (SR model) is expected to output a super-resolved HR image $G(x_l)$ with high identity discrimination. To achieve this, an association constraint ϕ between the real-fake discriminator representation $\boldsymbol{f}_{\mathrm{d}}$ and identity classification $\boldsymbol{f}_{\mathrm{c}}$ representations of the image x_h is designed. Then, ϕ is represented and learned on HR training images x_h with a small network, considering that they are the target the SR images $G(x_l)$ need to approach during training.

Formally, the association is learned as the transformation from the identity recognition $\boldsymbol{f}_{\mathrm{c}}$ to discriminator $\boldsymbol{f}_{\mathrm{d}}$ representations. This is based on a hypothesis that the identity recognition representation, learned from HR training images, contains the information for general high-resolution distribution (that the real-fake discriminator tries to learn); Whilst the discriminator features are relatively less informative compared to the identity ones, due to being derived from a simpler binary classification task (real or fake). Learning such a mapping is thus more sensible. In particular, we derive an association regularisation as:

$$\mathcal{L}_{\text{intact}} = ||\phi(\boldsymbol{f}_{\mathrm{c}}) - \boldsymbol{f}_{\mathrm{d}}||_2^2. \tag{5.5}$$

It aims to optimise the parameters ϕ of the association network, using $\boldsymbol{f}_{\mathrm{c}}$ extracted from the body recognition model and $\boldsymbol{f}_{\mathrm{d}}$ extracted from the discriminator on the HR image x_h.

To facilitate learning ϕ, an additional bridging constraint is further imposed for manipulating the optimising direction. Specifically, an intermediate latent feature space $\boldsymbol{f}_{\mathrm{e}}$ is isolated from ϕ

such that a bridging operation can be implanted with a transform σ of the target $\boldsymbol{f}_{\text{d}}$, defined as:

$$\mathcal{L}_{\text{e}} = ||\sigma(\boldsymbol{f}_{\text{d}}) - \boldsymbol{f}_{\text{e}}||_2^2, \tag{5.6}$$

where $\boldsymbol{f}_{\text{e}}$ is obtained in the middle latent space of ϕ with $\boldsymbol{f}_{\text{c}}$ as input. The bridging module σ is jointly learned with the association module ϕ in a combination as:

$$\mathcal{L}_{\text{intact-e}} = \mathcal{L}_{\text{intact}} + \mathcal{L}_{\text{e}}. \tag{5.7}$$

Part II: Association Regularisation Once the inter-task association network ϕ is learned as above, it is treated as a critic to regularise the learning of the SR model (the generator) and the discriminator in the GAN based multi-task learning network. The learned association is distilled by similarly coupling the information of discriminator and identity recognition. Particularly, this distillation loss is in the same form of Eq. (5.5) but applied to the SR images $G(x_l)$ as:

$$\mathcal{L}_{\text{dis}} = ||\phi(\boldsymbol{f}'_{\text{c}}) - \boldsymbol{f}'_{\text{d}}||_2^2, \tag{5.8}$$

where $\boldsymbol{f}'_{\text{c}}$ and $\boldsymbol{f}'_{\text{d}}$ are the corresponding identity and discriminator representations of a single SR image $G(x_l)$ analogue to x_h above.

It is worth mentioning that, unlike Eq. (5.5), here the association network ϕ is fixed to functionally serve as an external *critic* in this step. This role is similar in spirit as the ImageNet pretrained VGG model of the perceptual loss [89]. Using $\mathcal{L}_{\text{dis}}$ along with GAN training, the synthesis of such HR images that respect the same association relation between identity and fidelity on the genuine HR images is essentially encouraged. This is the key drive behind the INTACT model that imposes both supervision signals and importantly their interaction in a single formulation.

Remarks Unlike the *de facto* multi-task inference using weighted loss summation for inter-task interaction learning and communicating, the underlying association between two tasks is discovered as an extra learning constraint. Significantly, once parameterised this association can be automatically learned from the original training data themselves in a data-driven manner, without any hand-crafting and the need for ad-hoc knowledge. Consequently, the intrinsic conflicts between two different tasks can be mitigated effectively, benefiting the overall model learning process towards person identity matching. Moreover, it can be also considered that INTACT takes a *soft* integration design that aims to link the underlying objectives between two different tasks by maximising their positive correlation during training. Consequently, the two learning objectives can adaptively collaborate in a unified learning process with a balanced trade-off between individual and common pursuits.

Algorithm 1 INTACT model training

```
Input: Training data $\mathcal{D} = \{x_l, x_h\}$ with identity labels $Y$.
Output: A person image super-resolution (SR) model.
Initialisation: Training a standard person body recognition model with HR images and the identity labels.
Alternating training (frozen one, and update the others):
for $i = 1$ to $iter$ do
    (1) Update the discriminator with the GAN loss (Eq. (5.2));
    (2) Update the association network $\phi$ (Eq. (5.7));
    (3) Update the generator (SR model) with the SR objective loss (Eq. (5.4)) and distillation loss (Eq. (5.8)).
end for
```

5.2.3 Model Training

In model training, the INTACT loss terms are seamlessly integrated with the standard GAN optimisation with one more step. The whole model remains end-to-end trainable. The entire training process is summarised in Algorithm 1.

5.3 Experiments

5.3.1 Datasets

This section used five person body recognition benchmarks for evaluations. The **CUHK03** dataset comprises 14,097 images of 1,467 identities with 5 different camera views. As [114], the 1,367/100 training/test identity split was used. The **VIPeR** dataset contains 632 person image pairs captured by 2 cameras. Following [114], this dataset was randomly divided into two non-overlapping halves based on the identity labels. Namely, images of a subject belong to either the training or the test set. The **CAVIAR** dataset contains 1,220 images of 72 person identities captured by 2 cameras. 22 people who only appear in the closer camera were discarded, and the remaining was split into two non-overlapping halves in the identity labels as [114]. The

Market-1501 dataset consists of 32,668 images of 1,501 identities captured in 6 camera views. The standard 751/750 training/test identity split was used. The **DukeMTMC-reID** dataset contains 36,411 images of 1,404 identities captured by 8 cameras. The standard 702/702 training/test identity split was adopted.

Following [86, 114], this section evaluated the setting of multiple low-resolution (MLR) person body recognition. Four synthetic and one real-world cross-resolution body recognition benchmarks were tested. Specifically, for the synthetic cases (Market-1501, CUHK08, VIPeR, and DukeMTMC), the query images taken from one camera are down-sampled by a randomly selected downsampling rate r $\in \{2, 3, 4\}$ (i.e.the spatial size of a downsampled image becomes H/r $\times$ W/r), while the images taken by the other camera(s) remain unchanged. The Multiple Low Resolution (MLR) datasets are named as ***MLR-dataset***. On the other hand, the CAVIAR dataset provides realistic images of multiple resolutions, i.e.a genuine MLR dataset for evaluating cross-resolution person body recognition.

5.3.2 Experimental Settings

The proposed INTACT model was evaluated using the cross-resolution person body recognition setting [86, 114], where the probe set contains LR images whilst the gallery set contains HR images. The standard single-shot person body recognition setting was adopted, and the average cumulative match characteristic was used as the evaluation metric.

5.3.3 Implementation Details

All the experiments were performed in PyTorch on a machine with a Tesla P100 GPU. During training, the varying LR images are generated by randomly down-sampling HR images by $r \in \{2, 3, 4\}$ times. All the LR images were then resized to $256\times128\times3$ for both model training and deployment. The residual blocks [72] were used as the backbone of the proposed model. For the SR generator, an encoder-decoder architecture was adopted. Specifically, it consists of 16 residual blocks equally distributed in 8 groups. The resolution drops 16 times from 256×128 to 16×8 pixels (due to the first 4 residual block groups each with a max pooling layer), and then increases back to 256×128 with the last 4 groups of residual block each with pixel shuffling. The generator's architecture is shown in Fig. 5.2. The discriminator is similar as [103]. The person body recognition network [235] was pre-trained on the HR training data. Once trained, it was frozen during the training of INTACT.

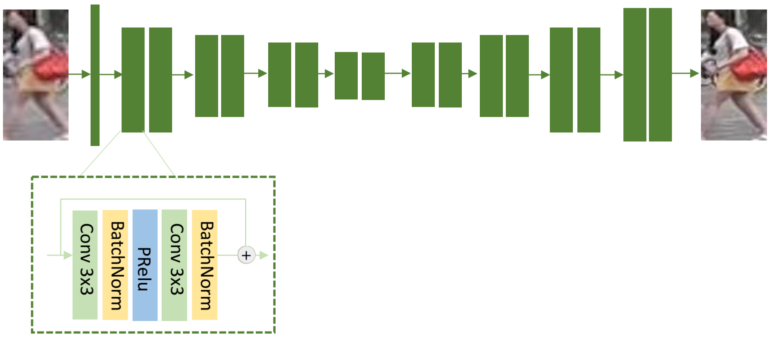

Figure 5.2: Architecture of image SR model (GAN's generator).

The inter-task association network ϕ was implemented by an encoder-decoder network, where the intermediate latent feature space $\boldsymbol{f}_{\mathrm{e}}$ is set as the encoder's output. Both encoder and decoder contain three FC layers each followed with batch-normalisation, respectively. The dimension of the latent feature space was set to 200. The bridging module σ for $\boldsymbol{f}_{\mathrm{d}}$ shares the same structure to the encoder of ϕ. The extra overhead introduced by the association network is marginal, as compared to the standard GAN training cost. Per-iteration cost increase was not noticed. Actually, it was observed that with INTACT the whole model often converges using less epochs, leading to a faster training process than the standard multi-task GAN baseline. The learning rate was set to 1×10^{-4} for generator G and discriminator D, and 1×10^{-3} for the association module ϕ. The mini-batch size was 32. The loss hyper-parameters was set consistently in all the experiments as: $\lambda_g = 0.1$, $\lambda_c = 0.3$. In practice, these parameters were selected by balancing their loss value scales to avoid any dominating term in training.

Table 5.1: Cross-resolution person body recognition performance (%). Bold and underlined numbers indicate top two results, respectively.

Model	MLR-Market-1501			MLR-CUHK03			MLR-VIPeR			MLR-DukeMTMC-reID			CAVIAR		
	Rank1	Rank5	Rank10	Rank1	Rank5	Rank10	Rank1	Rank5	Rank10	Rank1	Rank5	Rank10	Rank1	Rank5	Rank10
CamStyle [234]	74.5	88.6	93.0	69.1	89.6	93.9	34.4	56.8	66.6	64.0	78.1	84.4	32.1	72.3	85.9
FD-GAN [53]	79.6	91.6	93.5	73.4	93.8	97.9	39.1	62.1	72.5	67.5	82.0	85.3	33.5	71.4	86.5
SLD^{2}L [88]	-	-	-	-	-	-	20.3	44.0	62.0	-	-	-	18.4	44.8	61.2
SING [86]	74.4	87.8	91.6	67.7	90.7	94.7	33.5	57.0	66.5	65.2	80.1	84.8	33.5	72.7	89.0
CSR-GAN [195]	76.4	88.5	91.9	71.3	92.1	97.4	37.2	62.3	71.6	67.6	81.4	85.1	34.7	72.5	87.4
JUDEA [112]	-	-	-	26.2	58.0	73.4	26.0	55.1	69.2	-	-	-	22.0	60.1	80.8
SDF [194]	-	-	-	22.2	48.0	64.0	9.3	38.1	52.4	-	-	-	14.3	37.5	62.5
RAIN [36]	-	-	-	78.9	97.3	98.7	42.5	<u>68.3</u>	<u>79.6</u>	-	-	-	42.0	<u>77.3</u>	89.6
CAD [114]	<u>83.7</u>	<u>92.7</u>	<u>95.8</u>	<u>82.1</u>	<u>97.4</u>	**98.8**	<u>43.1</u>	68.2	77.5	<u>75.6</u>	<u>86.7</u>	<u>89.6</u>	<u>42.8</u>	76.2	<u>91.5</u>
INTACT (Ours)	**88.1**	**95.0**	**96.9**	**86.4**	**97.4**	<u>98.5</u>	**46.2**	**73.1**	**81.6**	**81.2**	**90.1**	**92.8**	**44.0**	**81.8**	**93.9**

5.3.4 Comparisons to State-of-the-Art Methods

The INTACT was compared with a wide range of state-of-the-art body recognition methods, including (1) Conventional person body recognition models: CamStyle [234] and FD-GAN [53]; (2) Super-resolution based models: SLD^{2}L [88], SING [86], CSR-GAN [195]; (3) Resolution-invariant representation learning based models: JUDEA [112], SDF [194], RAIN [36]; and (4) A hybrid method CAD [114] that combines SR and resolution-invariant representation learning.

The results comparisons are shown in Table 5.1. There are the following observations:

(1) INTACT achieves the state-of-the-art performance on all the five datasets, consistently outperforming the best competitor [114] by up to 6% at Rank-1.

(2) Compared to the SR based cross-resolution person body recognition methods (SLD^{2}L [88], SING [86], CSR-GAN [195]), INTACT achieves significant improvement, e.g.up to 15.1% Rank-1 performance boost. This validates that the proposed model can effectively address the inferior compatibility issue between image SR and person body recognition as suffered by these previous state-of-the-art methods.

(3) Compared to resolution-invariant representation learning models (JUDEA [112], SDF [194], RAIN [36]), INTACT achieves the best performance on both the small datasets (MLR-VIPeR and CAVIAR, which is generally very challenging for deep learning methods due to no sufficient training data), and the large dataset (MLR-CUHK03), often by a large margin. This suggests that image SR based methods provide more superior solutions.

(4) Compared to the best competitor [114] that exploits both image SR and resolution-invariant representation learning, INTACT remains a better method by only using image SR as the core strategy.

(5) The standard person body recognition models (CamStyle [234] and FD-GAN [53]) suffer from significant performance drop on MLR person body recognition datasets, as compared to their reported results on standard HR person body recognition datasets. This shows that the resolution mismatch problem is typically ignored by most existing body recognition methods.

5.3.5 Inter-Task Association Analysis

A multi-task learning framework was adopted as the base model, where the SR module serves as a preprocessing step to recover the essential details originally missing in LR images in order to more accurately match HR gallery images. It is considered that the SR task inherently may be

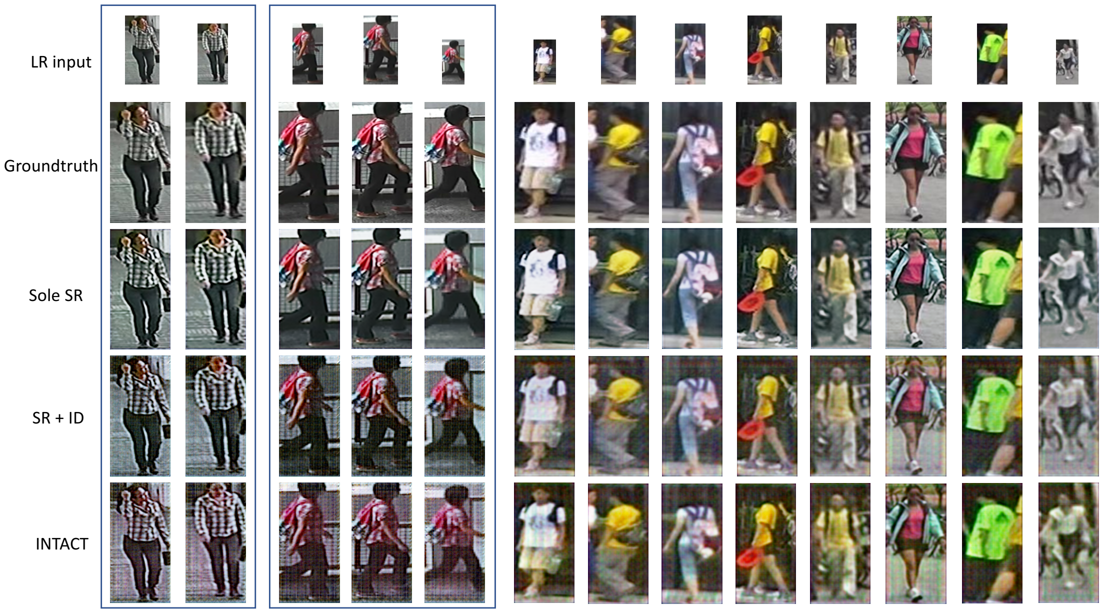

Figure 5.3: Examples of recovered images from the test set of MLR-CUHK03 and MLR-Market1501.

not compatible to the identity matching task, which is the reason why we introduced INTACT as an explicit regularisation for SR. Here, an experiment was conducted to examine the association between the two different tasks (image SR & person body recognition) and its effect on the overall model performance.

Evaluation metrics The pixel-wise SR quality of the recovered images by INTACT was measured on the MLR-CUHK03 test set. The SSIM and PSNR metrics were used, together with Rank-1 as person body recognition performance metric.

Competitors INTACT was compared with: (1) Sole SR: the proposed method without person body recognition constraint, (2) SR+ID: jointly learning image SR and person body recognition, and (3) four state-of-the-art body recognition models (CycleGAN [237], SING [86], CSR-GAN [195] and CAD [114]).

Results The performance comparisons in Table 5.2 show the follows observations:

(1) The sole SR module of INTACT (supervised by MSE loss only) achieves the best pixel-wise SR performance, i.e.the highest PSNR and SSIM scores. This verifies the effectiveness of the SR generator in the proposed model.

(2) Although the sole SR model achieves the highest PSNR and SSIM performance, its resolved images yield the worst accuracy for cross-resolution person body recognition. This indicates that as the low-level image quality metrics, both SSIM and PSNR are unsuited for evaluating

high-level semantic recognition tasks such as person body recognition in our case.

(3) The models that favour the super-resolution performance do not provide improvements on cross-resolution person body recognition performance. This suggests that the SR supervision is not directly relevant to person body recognition.

(4) The SR modules that do benefit the person body recognition task enhance only the identity matching related details whilst ignores other fine-grained details. This actually produces inferior pixel-level fidelity.

Several qualitative comparison of the recovered images for visual examination is provided in Fig. 5.3. Whilst the INTACT notably outperforms all the baselines in numerical evaluation, this performance difference is however seldom reflected in the low-level image space. This implies that high-level semantic objective is less interpretive due to high functional complexity of deep network models.

Table 5.2: Comparison of super-resolution and cross-resolution person body recognition performance on the MLR-CUHK03 test set.

Model	SSIM	PSNR	Rank1
CycleGAN [237]	0.55	14.1	62.1
SING [86]	0.65	18.1	67.7
CSR-GAN [195]	0.76	21.5	71.3
CAD [114]	0.73	20.2	82.1
Sole SR	**0.82**	**26.6**	23.0
SR + ID	0.77	23.3	82.7
INTACT	0.73	22.8	**86.4**

5.3.6 Ablation Study

Loss component analysis The INTACT is jointly trained with image SR, person body recognition and the inter-task association loss functions (cf. Eq. (5.7) & (5.8)). This section examines their performance effects on the MLR-Market-1501 dataset. (Noted that similar ablation study was also conducted on other MLR datasets and achieved similar results. Here the most typical case is selected for a simplified comparison.) Table 5.3 reports the ablation results. It is observed that:

(1) With identity classification alone, the model achieves the poorest matching performance. This

verifies that jointly learning the multi-task framework with the standard cascaded SR and person body recognition model is unsatisfactory.

(2) After adding GAN loss, the model achieves slightly better performance. The plausible reason is that the adversarial loss helps to align the statistics of resolved images to the native HR data. However, the improvement is fairly marginal.

(3) Importantly, the proposed association loss brings a significant improvement, verifying the effectiveness of our regularisation scheme based on the idea of exploiting the underlying inter-task correlation.

Table 5.3: Loss component analysis of INTACT on MLR-Market-1501. MSE: pixel-wise content loss, ID: identity classification loss (Eq. (5.3)), Association: our association loss (Eq. (5.7) & (5.8)).

Supervision	Rank1	Rank5	Rank10
MSE+ID	83.7	93.0	95.6
MSE+ID+GAN	84.7	93.9	96.1
MSE+ID+GAN+Association	**88.1**	**95.0**	**96.9**

Association design For the association learning between the discrimination and recognition representations in INTACT, a recognition-to-discriminator design was adopted. This is based on a hypothesis that the identity recognition representations learned from the HR images should contain the desired information of high resolution (that the real-fake HR discriminator tries to learn); And the real-fake HR discriminator representations, derived by a simple binary classification task, are relatively simpler.

This section examines the effect of association design by additionally testing two more formulations: (i) Common space association (Fig. 5.4 (a)), and (ii) Discrimination-to-recognition association (Fig. 5.4 (b)) which is the inverse of the recognition-to-discriminator design (Fig. 5.4 (c)) adopted in INTACT. Table 5.4 shows that different designs present fairly similar performances, and the recognition-to-discriminator is the best choice. This verifies the proposed association strategy.

Bridging constraint To facilitate the training of inter-task association between the discriminator and identity classification representations, an intermediate latent feature space f_e was isolated from ϕ to bridge the association target (Eq. (5.6) and (5.7)), implemented by an encoder-decoder structure. The result in Table 5.5 shows that the introduction of such a bridging constraint helps

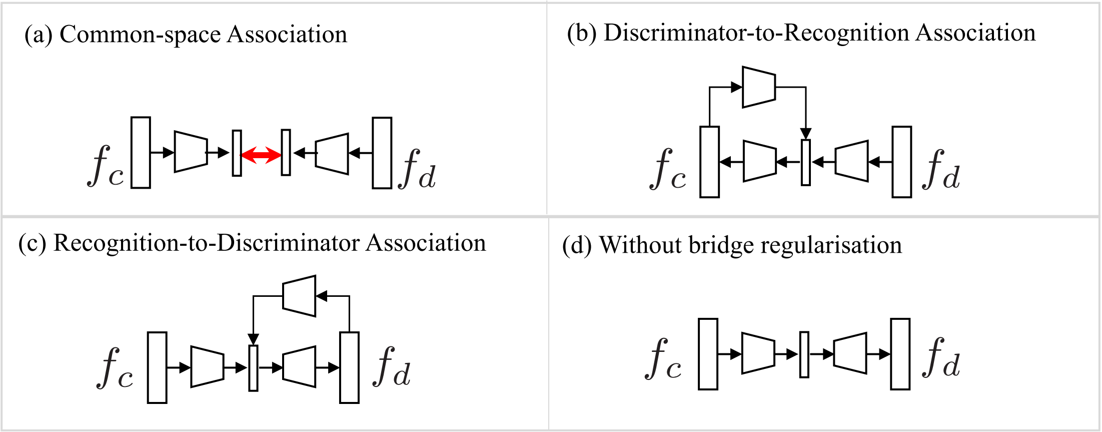
(a) Common-space Association
f_c
f_d
(b) Discriminator-to-Recognition Association
f_c
f_d
(c) Recognition-to-Discriminator Association
f_c
f_d
(d) Without bridge regularisation
f_c
f_d

Figure 5.4: Schematics of different association designs.

Table 5.4: Association designs. R-to-D: from identity recognition representation to discriminator representation (used in INTACT); D-to-R: the inverse.

Association Space	Rank1	Rank5	Rank10
Common Space (a)	84.3	94.0	95.3
D-to-R (b)	83.4	93.5	95.0
R-to-D (c, ours)	**88.1**	**95.0**	**96.9**

to better constrain the associative learning.

Table 5.5: Effect of the bridge constraint (Eq. (5.6)).

Bridge constraint	Rank1	Rank5	Rank10
W/O (Fig. 5.4 (d))	84.3	93.5	95.8
W (Fig. 5.4 (c))	**88.1**	**95.0**	**96.9**

5.4 Summary

This chapter presents a novel deep learning regularisation, named *Inter-Task Association Critic* (INTACT), for solving the under-studied yet important cross-resolution person body recognition problem. As a generic learning constraint, INTACT is designed specially for improving the training of existing multi-task (image SR and person body recognition) models, by alleviating properly the difficulty of gradients backpropagation through two cascaded networks. During training, INTACT discovers the underlying association knowledge between image SR and person body recognition by learning from the HR training data, and uses the self-discovered association information to further guide the learning behaviour of SR model alternatively. Thus the compatibility of SR with body recognition matching can be maximised. This is built up on parameterising the inter-task association with a dedicated network. Extensive experimental results have demonstrated the performance superiority of the proposed model over a wide variety of existing cross-resolution and standard person body recognition methods on five challenging benchmarks. Component analysis of the model provides insights into the formulation of INTACT.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

This thesis has presented various super resolution (SR) based models for person recognition in low-quality imagery, with applications to face recognition and person body recognition. The model optimisation difficulties by low resolution is challenging due to the much less fine-grained discriminative details for robust identity matching. And conventional SR models fail to enhance the essential image details beneficial for identity recognition, since direct super resolution is less compatible with identity matching, and hence has minor benefit or even negative effect for low-resolution person recognition. Specifically, to solve this problem, in this thesis:

1. In Chapter 3, a novel Complement Super-Resolution and Identity (CSRI) joint deep learning method with a unified end-to-end network architecture is introduced, to improve the identity recognisability of super resolution for low-resolution facial matching. This model is based on knowledge transfer for super resolution towards native LR.

2. Chapter 4 focuses on improving the fidelity of the super resolved facial images to enable an interpretable face identity recognition model and more downstream facial analysis tasks. To solve the ineffective training problems by unsupervised domain adaptation (UDA) adopted by previous unpaired super resolution models, in this chapter, a method that joins the advantages of conventional SR and UDA models is formulated, where Characteristic Regularisation (CR) is introduced between the SR and UDA to separate and control their optimisations. Such design makes the model training more effective and

computationally tractable.

3. Chapter 5 focuses on the person body recognition task, to solve the inherent challenges by the resolution mismatch problem when the pedestrian images are captured in various resolutions. This chapter introduces a novel model training regularisation method, called Inter-Task Association Critic (INTACT) to overcome the gradients backpropagation difficulties in the previous cascaded SR and person body recognition joint-learning schemes. This is realised by discovering the underlying association knowledge between SR and person body recognition, as an extra learning constraint for enhancing the compatibility of SR with person body recognition in HR image space.

6.2 Future Work

The potential research directions for future work beyond the proposed methods are summarised as follows to end this thesis.

1. (Chapter 3) **Low-resolution person face recognition by super resolution**: The evaluation results show that knowledge transfer for super resolution toward native low resolution is able to improve the model generalisation ability for realistic low-resolution face recognition. In Chapter 3, the joint-learning scheme is adopted to realise the knowledge transfer idea. Recently, a variety of works have proposed more advanced transfer learning techniques, e.g., domain adaptation and style transfer. It is an important future effort to explore more effective transfer learning methods for realistic super resolution and the downstream low-resolution face recognition task.

2. (Chapter 4) **Interpretable low-resolution face recognition**: Although the proposed model make a step towards the high-fidelity and discriminative super resolution for real-world low-resolution faces, it remains largely unsolved to deal with the extreme challenging cases, characterised by blur, occlusion and large poses. And the super-resolved faces are still far from satisfactory for high-accuracy face recognition. It is essential in the future works to adopt more advanced super resolution models, constrained by more effective recognition supervision signals. Also, other techniques adopted by broader computer vision tasks could be explored, e.g., image deblurring, unpaired image translation and zero-shot learning.

3. (Chapter 5) **Cross-resolution person body recognition**: The association between different tasks is explored and adopted as an extra constraint to supervise the optimisation in the multi-task scheme. In addition to solving the resolution mismatch issue for person body recognition, such design has potentials for dealing with other similar multi-task learning tasks in computer vision. It would enable wider research applications to further explore its benefits for other problems or more complicated multi-task frameworks in more general settings. Besides, it would largely benefit the current extremely challenging low-quality person recognition task, to enable a multi-source identity recognition system, by combining the face, body, and other biometric cues automatically. To achieve this, the correlation among different cues, along with their corresponding convincing scores need to be considered. The inter-task association technique proposed in Chapter 5 could be further adopted to learn such inter-cues correlation, considering the representation learning of each cue as one task in a multi-task system.

Bibliography

[1] Timo Ahonen, Abdenour Hadid, and Matti Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):2037–2041, 2006.

[2] Song Bai, Peng Tang, Philip HS Torr, and Longin Jan Latecki. Re-ranking via metric fusion for object retrieval and person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 740–749, Long Beach, CA, USA, 2019.

[3] Yancheng Bai, Yongqiang Zhang, Mingli Ding, and Bernard Ghanem. Sod-mtgan: Small object detection via multi-task generative adversarial network. In *European Conference on Computer Vision*, pages 206–221, Munich, Germany, 2018.

[4] Simon Baker and Takeo Kanade. Hallucinating faces. In *IEEE International Conference on Automatic Face and Gesture Recognition*, Grenoble, France, 2000.

[5] Davide Baltieri, Roberto Vezzani, and Rita Cucchiara. 3dpes: 3d people dataset for surveillance and forensics. In *Joint ACM workshop on Human gesture and behavior understanding*, pages 59–64, Scottsdale, Arizona, USA, 2011.

[6] Davide Baltieri, Roberto Vezzani, and Rita Cucchiara. Sarc3d: a new 3d body model for people tracking and re-identification. In *International Conference on Image Analysis and Processing*, pages 197–206, Ravenna, Italy, 2011.

[7] Ankan Bansal, Carlos Castillo, Rajeev Ranjan, and Rama Chellappa. The do's and don'ts for cnn-based face verification. In *Workshop of IEEE International Conference on Computer Vision*, pages 2545–2554, Venice, Italy, 2017.

[8] Ankan Bansal, Anirudh Nanduri, Carlos Castillo, Rajeev Ranjan, and Rama Chellappa. Umdfaces: An annotated face dataset for training deep networks. *International Joint Conference on Biometrics*, pages 464–473, 2017.

[9] Marian Stewart Bartlett, Gwen Littlewort, Ian Fasel, and Javier R Movellan. Real time face detection and facial expression recognition: Development and applications to human computer interaction. In *Workshop of IEEE Conference on Computer Vision and Pattern Recognition*, volume 5, pages 53–53, Madison, WI, USA, 2003.

[10] Peter N. Belhumeur, João P Hespanha, and David J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997.

[11] Sefi Bell-Kligler, Assaf Shocher, and Michal Irani. Blind super-resolution kernel estimation using an internal-gan. In *Advances in Neural Information Processing Systems*, pages 284–293, Vancouver, BC, Canada, 2019.

[12] Lacey Best-Rowden, Hu Han, Charles Otto, Brendan F Klare, and Anil K Jain. Unconstrained face recognition: Identifying a person of interest from a media collection. 9(12):2144–2157, 2014.

[13] J Ross Beveridge, P Jonathon Phillips, David S Bolme, Bruce A Draper, Geof H Givens, Yui Man Lui, Mohammad Nayeem Teli, Hao Zhang, W Todd Scruggs, Kevin W Bowyer, et al. The challenge of face recognition from digital point-and-shoot cameras. In *IEEE International Conference on Biometrics: Theory, Applications, and Systems*, pages 1–8, Arlington, VA, USA, 2013.

[14] Emil Bilgazyev, Boris A Efraty, Shishir K Shah, and Ioannis A Kakadiaris. Sparse representation-based super resolution for face recognition at a distance. In *British Machine Vision Conference*, Dundee, UK, 2011.

[15] Soma Biswas, Kevin W Bowyer, and Patrick J Flynn. Multidimensional scaling for matching low-resolution facial images. In *IEEE International Conference on Biometrics: Theory, Applications, and Systems*, pages 1–6, Washington, DC, USA, 2010.

[16] Soma Biswas, Kevin W Bowyer, and Patrick J Flynn. Multidimensional scaling for matching low-resolution face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(10):2019–2030, 2012.

[17] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks.

In *IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, 2017.

[18] Andrés Bruhn, Joachim Weickert, and Christoph Schnörr. Lucas/kanade meets horn/schunck: Combining local and global optic flow methods. *International Journal of Computer Vision*, 61(3):211–231, 2005.

[19] Adrian Bulat and Georgios Tzimiropoulos. Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans. In *IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018.

[20] Adrian Bulat, Jing Yang, and Georgios Tzimiropoulos. To learn image super-resolution, use a gan to learn how to do image degradation first. In *European Conference on Computer Vision*, Munich, Germany, 2018.

[21] Qingxing Cao, Liang Lin, Yukai Shi, Xiaodan Liang, and Guanbin Li. Attention-aware face hallucination via deep reinforcement learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 690–698, Honolulu, Hawaii, USA, 2017.

[22] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 67–74, Xi'an, China, 2018.

[23] Xudong Cao, Yichen Wei, Fang Wen, and Jian Sun. Face alignment by explicit shape regression. *International Journal of Computer Vision*, 107(2):177–190, 2014.

[24] Xudong Cao, David Wipf, Fang Wen, Genquan Duan, and Jian Sun. A practical transfer learning algorithm for face verification. In *IEEE International Conference on Computer Vision*, pages 3208–3215, Sydney, NSW, Australia, 2013.

[25] Ayan Chakrabarti, AN Rajagopalan, and Rama Chellappa. Super-resolution of face images using kernel pca-based prior. *IEEE Transactions on Multimedia*, 9(4):888–892, 2007.

[26] Sarath Chandar, Chinnadhurai Sankar, Eugene Vorontsov, Samira Ebrahimi Kahou, and Yoshua Bengio. Towards non-saturating recurrent units for modelling long-term dependencies. In *AAAI Conference on Artificial Intelligence*, Honolulu, Hawaii, USA, 2019.

[27] Xiaobin Chang, Timothy M Hospedales, and Tao Xiang. Multi-level factorisation net for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2109–2118, Salt Lake City, UT, USA, 2018.

[28] Dong Chen, Xudong Cao, Liwei Wang, Fang Wen, and Jian Sun. Bayesian face revisited: A joint formulation. In *European Conference on Computer Vision*, pages 566–579, Florence, Italy, 2012.

[29] Dong Chen, Xudong Cao, Fang Wen, and Jian Sun. Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. In *IEEE Conference on Computer Vision and Pattern Recognition*, Portland, OR, USA, 2013.

[30] Jiawei Chen, Jonathan Wu, Janusz Konrad, and Prakash Ishwar. Semi-coupled two-stream fusion convnets for action recognition at extremely low resolutions. In *IEEE Winter Conference on Applications of Computer Vision*, pages 139–147, Santa Rosa, CA, USA, 2017.

[31] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 403–412, Columbus, OH, USA, 2017.

[32] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. Deep association learning for unsupervised video person re-identification. *arXiv e-print*, 2018.

[33] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. Instance-guided context rendering for cross-domain person re-identification. In *IEEE International Conference on Computer Vision*, pages 232–242, Seoul, Korea, 2019.

[34] Ying-Cong Chen, Xiatian Zhu, Wei-Shi Zheng, and Jian-Huang Lai. Person re-identification by camera correlation aware feature augmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(2):392–408, 2017.

[35] Yu Chen, Ying Tai, Xiaoming Liu, Chunhua Shen, and Jian Yang. Fsrnet: End-to-end learning face super-resolution with facial priors. In *IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018.

[36] Yun-Chun Chen, Yu-Jhe Li, Xiaofei Du, and Yu-Chiang Frank Wang. Learning resolution-invariant deep representations for person re-identification. In *AAAI Conference on Artificial Intelligence*, volume 33, pages 8215–8222, Honolulu, Hawaii, USA, 2019.

[37] Dong Seon Cheng, Marco Cristani, Michele Stoppa, Loris Bazzani, and Vittorio Murino. Custom pictorial structures for re-identification. In *British Machine Vision Conference*, volume 1, page 6, Dundee, UK, 2011.

[38] Zhiyi Cheng, Xiatian Zhu, and Shaogang Gong. Low-resolution face recognition. In *Asian Conference on Computer Vision*, pages 605–621, Perth, Australia, 2018.

[39] Zhiyi Cheng, Xiatian Zhu, and Shaogang Gong. Characteristic regularisation for super-resolving face images. In *IEEE Winter Conference on Applications of Computer Vision*, pages 2435–2444, Snowmass Village, CO, USA, 2020.

[40] Zhiyi Cheng, Xiatian Zhu, and Shaogang Gong. Face re-identification challenge: Are face recognition models good enough? *Pattern Recognition*, page 107422, 2020.

[41] Jae Young Choi, Yong Man Ro, and Konstantinos N Plataniotis. Color face recognition for degraded face images. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(5):1217–1230, 2009.

[42] Patrick Connor and Arun Ross. Biometric recognition by gait: A survey of modalities and features. *Computer Vision and Image Understanding*, 167:1–27, 2018.

[43] Abir Das, Anirban Chakraborty, and Amit K Roy-Chowdhury. Consistent re-identification in a camera network. In *European Conference on Computer Vision*, pages 330–345, Zurich, Switzerland, 2014.

[44] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *European Conference on Computer Vision*, pages 184–199, Zurich, Switzerland, 2014.

[45] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):295–307, 2016.

[46] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In *European Conference on Computer Vision*, pages 391–407, Amsterdam, The Netherlands, 2016.

[47] Qi Dong, Shaogang Gong, and Xiatian Zhu. Person search by text attribute query as zero-shot learning. In *IEEE International Conference on Computer Vision*, pages 3652–3661, Seoul, Korea, 2019.

[48] Easen Electron. Easenelectron. `http://english.easen-electron.com/`, 2017.

[49] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. Person re-identification by symmetry-driven accumulation of local features. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2360–2367, San Francisco, CA, USA, 2010.

[50] Clinton Fookes, Frank Lin, Vinod Chandran, and Sridha Sridharan. Evaluation of image resolution and super-resolution on face recognition performance. *Journal of Visual Communication and Image Representation*, 23(1):75–93, 2012.

[51] Yang Fu, Yunchao Wei, Yuqian Zhou, Honghui Shi, Gao Huang, Xinchao Wang, Zhiqiang Yao, and Thomas Huang. Horizontal pyramid matching for person re-identification. In *AAAI Conference on Artificial Intelligence*, volume 33, pages 8295–8302, Honolulu, Hawaii, USA, 2019.

[52] Wen Gao, Bo Cao, Shiguang Shan, Xilin Chen, Delong Zhou, Xiaohua Zhang, and Debin Zhao. The cas-peal large-scale chinese face database and baseline evaluations. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 38(1):149–161, 2007.

[53] Yixiao Ge, Zhuowan Li, Haiyu Zhao, Guojun Yin, Shuai Yi, Xiaogang Wang, et al. Fd-gan: Pose-guided feature distilling gan for robust person re-identification. In *Advances in Neural Information Processing Systems*, pages 1222–1233, Montral, Canada, 2018.

[54] Athinodoros S. Georghiades, Peter N. Belhumeur, and David J. Kriegman. 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, 2001.

[55] Niloofar Gheissari, Thomas B Sebastian, and Richard Hartley. Person reidentification using spatiotemporal appearance. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 1528–1535, New York, NY, USA, 2006.

[56] S. Gong, S. MeKenna, and A. Psarrou. *Dynamic Vision: From Images to Face Recognition*. Imperial College Press, World Scientific, May 2000.

[57] Shaogang Gong, Marco Cristani, Chen Change Loy, and Timothy M Hospedales. The re-identification challenge. In *Person re-identification*, pages 1–20. Springer, 2014.

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

[59] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, Montreal, Quebec, Canada, 2014.

[60] Mengran Gou, Srikrishna Karanam, Wenqian Liu, Octavia Camps, and Richard J Radke. Dukemtmc4reid: A large-scale multi-camera person re-identification dataset. In *Workshop of IEEE Conference on Computer Vision and Pattern Recognition*, pages 10–19, Honolulu, Hawaii, USA, 2017.

[61] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *European Conference on Computer Vision*, pages 262–275, Marseille, France, 2008.

[62] Mislav Grgic, Kresimir Delac, and Sonja Grgic. Scface–surveillance cameras face database. *Multimedia Tools and Applications*, 51(3):863–879, 2011.

[63] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010.

[64] Manuel Günther, Peiyun Hu, Christian Herrmann, Chi-Ho Chan, Min Jiang, Shufan Yang, Akshay Raj Dhamija, Deva Ramanan, Jürgen Beyerer, Josef Kittler, et al. Unconstrained face detection and open-set face recognition challenge. In *International Joint Conference on Biometrics*, pages 697–706, Denver, Colorado, USA, 2017.

[65] Bahadir K Gunturk, Aziz Umit Batur, Yucel Altunbasak, Monson H Hayes, and Russell M Mersereau. Eigenface-domain super-resolution for face recognition. *IEEE Transactions on Image Processing*, 12(5):597–606, 2003.

[66] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision*, pages 87–102, Amsterdam, The Netherlands, 2016.

[67] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision*, Amsterdam, The Netherlands, 2016.

[68] Sanchit Gupta, Nikita Gupta, Soumyadeep Ghosh, Maneet Singh, Shruti Nagpal, Mayank Vatsa, and Richa Singh. Facesurv: A benchmark video dataset for face detection and recognition across spectra and resolutions. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 1–7, Lille, France, 2019.

[69] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *IEEE Conference on Computer Vision and Pattern Recognition*, New York, NY, USA, 2006.

[70] Ju Han and Bir Bhanu. Individual recognition using gait energy image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(2):316–322, 2005.

[71] Muhammad Haris, Greg Shakhnarovich, and Norimichi Ukita. Task-driven super resolution: Object detection in low-resolution images. *arXiv e-print*, 2018.

[72] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, Las Vegas, NV, USA, 2016.

[73] Mingjie He, Jie Zhang, Shiguang Shan, Meina Kan, and Xilin Chen. Deformable face net for pose invariant face recognition. *Pattern Recognition*, 100:107113, 2020.

[74] Pablo H Hennings-Yeomans, Simon Baker, and BVK Vijaya Kumar. Simultaneous super-resolution and feature extraction for recognition of low-resolution faces. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, Anchorage, Alaska, USA, 2008.

[75] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equi-

librium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, Long Beach, California, USA, 2017.

[76] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. CyCADA: Cycle-consistent adversarial domain adaptation. In *International Conference on Machine learning*, Stockholm, Sweden, 2018.

[77] Peiyun Hu and Deva Ramanan. Finding tiny faces. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 951–959, Honolulu, Hawaii, USA, 2017.

[78] Peiyun Hu and Deva Ramanan. Finding tiny faces. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1522–1530, Columbus, OH, USA, 2017.

[79] Gary B. Huang and Erik Learned-Miller. Labeled faces in the wild. `http://vis-www.cs.umass.edu/lfw/results.html`, 2017.

[80] Gary B Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, University of Massachusetts, 2007.

[81] Zhiwu Huang, Shiguang Shan, Ruiping Wang, Haihong Zhang, Shihong Lao, Alifu Kuerban, and Xilin Chen. A benchmark and comparative study of video-based face recognition on cox face database. *IEEE Transactions on Image Processing*, 24(12):5967–5981, 2015.

[82] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in Neural Information Processing Systems*, pages 2017–2025, Montreal, Quebec, Canada, 2015.

[83] Herve Jegou, Matthijs Douze, and Cordelia Schmid. Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1):117–128, 2011.

[84] Kui Jia and Shaogang Gong. Generalized face super-resolution. *IEEE Transactions on Image Processing*, 17(6):873–886, 2008.

[85] Kui Jiang, Zhongyuan Wang, Peng Yi, Guangcheng Wang, Tao Lu, and Junjun Jiang. Edge-enhanced gan for remote sensing image superresolution. *IEEE Transactions on Geoscience and Remote Sensing*, 57(8):5799–5812, 2019.

[86] Jiening Jiao, Wei-Shi Zheng, Ancong Wu, Xiatian Zhu, and Shaogang Gong. Deep low-resolution person re-identification. In *AAAI Conference on Artificial Intelligence*, New Orleans, Louisiana USA, 2018.

[87] Yonggang Jin and Christos-Savvas Bouganis. Robust multi-image based blind face hallucination. In *IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, 2015.

[88] Xiao-Yuan Jing, Xiaoke Zhu, Fei Wu, Xinge You, Qinglong Liu, Dong Yue, Ruimin Hu, and Baowen Xu. Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 695–704, Boston, MA, USA, 2015.

[89] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711, Amsterdam, The Netherlands, 2016.

[90] Yasutomo Kawanishi, Yang Wu, Masayuki Mukunoki, and Michihiko Minoh. Shinpuhkan2014: A multi-camera pedestrian dataset for tracking people across multiple cameras. In *Korea-Japan Joint Workshop on Frontiers of Computer Vision*, volume 5, Okinawa, Japan, 2014.

[91] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4873–4882, Las Vegas, NV, USA, 2016.

[92] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1646–1654, Las Vegas, NV, USA, 2016.

[93] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016.

[94] Kwang In Kim and Younghee Kwon. Single-image super-resolution using sparse regression and natural image prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(6):1127–1133, 2010.

[95] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *International Conference on Machine learning*, Sydney, Australia, 2017.

[96] Brendan F Klare, Ben Klein, Emma Taborsky, Austin Blanton, Jordan Cheney, Kristen Allen, Patrick Grother, Alan Mah, and Anil K Jain. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1931–1939, Boston, MA, USA, 2015.

[97] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, Lake Tahoe, Nevada, USA, 2012.

[98] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 624–632, Columbus, OH, USA, 2017.

[99] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate superresolution. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, page 5, Columbus, OH, USA, 2017.

[100] Toby HW Lam, King Hong Cheung, and James NK Liu. Gait flow image: A silhouette-based gait representation for human identification. *Pattern Recognition*, 44(4):973–987, 2011.

[101] Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev, and Thomas S Huang. Interactive facial feature localization. In *European Conference on Computer Vision*, pages 679–692, Florence, Italy, 2012.

[102] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4681–4690, Columbus, OH, USA, 2017.

[103] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew P Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al.

Photo-realistic single image super-resolution using a generative adversarial network. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, page 4, Columbus, OH, USA, 2017.

[104] Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In *Artificial Intelligence and Statistics*, pages 562–570, San Diego, California, USA, 2015.

[105] Zhen Lei, Timo Ahonen, Matti Pietikäinen, and Stan Z Li. Local frequency descriptor for low-resolution face recognition. In *Face and Gesture*, pages 161–166, Santa Barbara, CA, USA, 2011.

[106] Minxian Li, Xiatian Zhu, and Shaogang Gong. Unsupervised person re-identification by deep learning tracklet association. In *European Conference on Computer Vision*, pages 737–753, Munich, Germany, 2018.

[107] Minxian Li, Xiatian Zhu, and Shaogang Gong. Unsupervised tracklet person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[108] Pei Li, Joel Brogan, and Patrick J Flynn. Toward facial re-identification: Experiments with data from an operational surveillance camera plant. In *IEEE International Conference on Biometrics: Theory, Applications, and Systems*, pages 1–8, Niagara Falls, NY, USA, 2016.

[109] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 152–159, Columbus, OH, USA, 2014.

[110] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2285–2294, Salt Lake City, UT, USA, 2018.

[111] Wei Li, Xiatian Zhu, and Shaogang Gong. Scalable person re-identification by harmonious attention. *International Journal of Computer Vision*, pages 1–19, 2019.

[112] Xiang Li, Wei-Shi Zheng, Xiaojuan Wang, Tao Xiang, and Shaogang Gong. Multi-scale learning for low-resolution person re-identification. In *IEEE International Conference on Computer Vision*, pages 3765–3773, Santiago, Chile, 2015.

[113] Ya Li, Guangrun Wang, Lin Nie, Qing Wang, and Wenwei Tan. Distance metric optimization driven convolutional neural network for age invariant face recognition. *Pattern Recognition*, 75:51–62, 2018.

[114] Yu-Jhe Li, Yun-Chun Chen, Yen-Yu Lin, Xiaofei Du, and Yu-Chiang Frank Wang. Recover and identify: A generative dual model for cross-resolution person re-identification. *arXiv e-print*, 2019.

[115] Yutian Lin, Xuanyi Dong, Liang Zheng, Yan Yan, and Yi Yang. A bottom-up clustering approach to unsupervised person re-identification. In *AAAI Conference on Artificial Intelligence*.

[116] Ce Liu, Heung-Yeung Shum, and William T Freeman. Face hallucination: Theory and practice. *International Journal of Computer Vision*, 75(1):115–134, 2007.

[117] Chengjun Liu and Harry Wechsler. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. *IEEE Transactions on Image Processing*, 11(4):467–476, 2002.

[118] Jinxian Liu, Bingbing Ni, Yichao Yan, Peng Zhou, Shuo Cheng, and Jianguo Hu. Pose transferrable person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4099–4108, 2018.

[119] Kan Liu, Bingpeng Ma, Wei Zhang, and Rui Huang. A spatio-temporal appearance representation for viceo-based pedestrian re-identification. In *IEEE International Conference on Computer Vision*, pages 3810–3818, Santiago, Chile, 2015.

[120] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems*, Long Beach, CA, USA, 2017.

[121] Tie-Yan Liu et al. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval*, 3(3):225–331, 2009.

[122] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 212–220, Honolulu, HI, USA, 2017.

[123] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *IEEE International Conference on Computer Vision*, Santiago, Chile, 2015.

[124] Chen Change Loy, Tao Xiang, and Shaogang Gong. Multi-camera activity correlation analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1988–1995, Miami, Florida, USA, 2009.

[125] Chaochao Lu and Xiaoou Tang. Surpassing human-level face verification performance on lfw with gaussianface. In *AAAI Conference on Artificial Intelligence*, Austin, Texas, USA, 2015.

[126] Ze Lu, Xudong Jiang, and Alex Kot. Deep coupled resnet for low-resolution face recognition. *IEEE Signal Processing Letters*, 25(4):526–530, 2018.

[127] Xiaolong Ma, Xiatian Zhu, Shaogang Gong, Xudong Xie, Jianming Hu, Kin-Man Lam, and Yisheng Zhong. Person re-identification by unsupervised video matching. *Pattern Recognition*, 65:197–210, 2017.

[128] Davide Maltoni, Dario Maio, Anil Jain, and Salil Prabhakar. *Handbook of fingerprint recognition*. Springer Science & Business Media, 2009.

[129] Niki Martinel and Christian Micheloni. Re-identify people in wide area camera network. In *Workshop of IEEE Conference on Computer Vision and Pattern Recognition*, pages 31–36, Providence, RI, USA, 2012.

[130] Daniel Martinho-Corbishley, Mark S Nixon, and John N Carter. Super-fine attributes with crowd prototyping. 41(6):1486–1500, 2018.

[131] Jingke Meng, Sheng Wu, and Wei-Shi Zheng. Weakly supervised person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 760–769, Long Beach, CA, USA, 2019.

[132] Kieron Messer, Jiri Matas, Josef Kittler, Juergen Luettin, and Gilbert Maitre. Xm2vtsdb: The extended m2vts database. In *International conference on audio and video-based biometric person authentication*, volume 964, pages 965–966, Washington, D.C., USA, 1999.

[133] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. Pose-guided feature alignment for occluded person re-identification. In *IEEE International Conference on Computer Vision*, pages 542–551, Seoul, Korea, 2019.

[134] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv e-print*, 2018.

[135] Zak Murez, Soheil Kolouri, David Kriegman, Ravi Ramamoorthi, and Kyungnam Kim. Image to image translation for domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018.

[136] Jiri Najemnik and Wilson S Geisler. Optimal eye movement strategies in visual search. *Nature*, 434(7031):387–391, 2005.

[137] Aaron Nech and Ira Kemelmacher-Shlizerman. Level playing field for million scale face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7044–7053, Columbus, OH, USA, 2017.

[138] Hong-Wei Ng and Stefan Winkler. A data-driven approach to cleaning large face datasets. In *IEEE International Conference on Image Processing*, pages 343–347, Paris, France, 2014.

[139] Kien Nguyen, Clinton Fookes, Arun Ross, and Sridha Sridharan. Iris recognition with off-the-shelf cnn features: A deep learning perspective. *IEEE Access*, 6:18848–18855, 2017.

[140] Kien Nguyen, Clinton Fookes, and Sridha Sridharan. Constrained design of deep iris networks. *arXiv e-print*, 2019.

[141] Kien Nguyen, Clinton Fookes, Sridha Sridharan, Massimo Tistarelli, and Mark Nixon. Super-resolution for biometrics: A comprehensive survey. *Pattern Recognition*, 78:23–42, 2018.

[142] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10), 2010.

[143] Seong-Jin Park, Hyeongseok Son, Sunghyun Cho, Ki-Sang Hong, and Seungyong Lee.

Srfeat: Single image super-resolution with feature discrimination. In *European Conference on Computer Vision*, pages 439–455, Munich, Germany, 2018.

[144] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *British Machine Vision Conference*, Swansea, UK, 2015.

[145] Peixi Peng, Tao Xiang, Yaowei Wang, Massimiliano Pontil, Shaogang Gong, Tiejun Huang, and Yonghong Tian. Unsupervised cross-dataset transfer learning for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016.

[146] P Jonathon Phillips, Patrick J Flynn, Todd Scruggs, Kevin W Bowyer, Jin Chang, Kevin Hoffman, Joe Marques, Jaesik Min, and William Worek. Overview of the face recognition grand challenge. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 947–954, San Diego, CA, USA, 2005.

[147] P Jonathon Phillips, Hyeonjoon Moon, Syed A Rizvi, and Patrick J Rauss. The feret evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(10):1090–1104, 2000.

[148] P Jonathon Phillips, W Todd Scruggs, Alice J O'Toole, Patrick J Flynn, Kevin W Bowyer, Cathy L Schott, and Matthew Sharpe. Frvt 2006 and ice 2006 large-scale experimental results. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5):831–846, 2010.

[149] Pejman Rasti, Tonis Uiboupin, Sergio Escalera, and Gholamreza Anbarjafari. Convolutional neural network super resolution for face recognition in surveillance monitoring. In *International conference on articulated motion and deformable objects*, pages 175–184, Palma de Mallorca, Spain, 2016.

[150] Chuan-Xian Ren, Dao-Qing Dai, and Hong Yan. Coupled kernel embedding for low-resolution face image recognition. *IEEE Transactions on Image Processing*, 21(8):3770–3783, 2012.

[151] Karl Ricanek and Tamirat Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 341–345, Southampton, UK, 2006.

[152] Ergys Ristani and Carlo Tomasi. Tracking multiple people online and in real time. In *Asian Conference on Computer Vision*, pages 444–459, Singapore, Singapore, 2014.

[153] Peter M. Roth, Martin Hirzer, Martin Köstinger, Csaba Beleznai, and Horst Bischof. Mahalanobis Distance Learning for Person Re-Identification. In *Person Re-Identification*, pages 247–267. Springer, 2014.

[154] Michael S Ryoo, Kiyoon Kim, and Hyun Jong Yang. Extreme low resolution activity recognition with multi-siamese embedding learning. In *AAAI Conference on Artificial Intelligence*, New Orleans, Louisiana USA, 2018.

[155] Mehdi SM Sajjadi, Bernhard Scholkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *IEEE International Conference on Computer Vision*, pages 4491–4500, Venice, Italy, 2017.

[156] Ferdinando S Samaria and Andy C Harter. Parameterisation of a stochastic model for human face identification. In *IEEE workshop on applications of computer vision*, pages 138–142, Sarasota, FL, USA, 1994.

[157] M Saquib Sarfraz, Arne Schumann, Andreas Eberle, and Rainer Stiefelhagen. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 420–429, Salt Lake City, UT, USA, 2018.

[158] Sudeep Sarkar, P Jonathon Phillips, Zongyi Liu, Isidro Robledo Vega, Patrick Grother, and Kevin W Bowyer. The humanid gait challenge problem: Data sets, performance, and analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(2):162–177, 2005.

[159] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, Boston, MA, USA, 2015.

[160] William Robson Schwartz and Larry S Davis. Learning discriminative appearance-based models using partial least squares. In *XXII Brazilian symposium on computer graphics and image processing*, pages 322–329, Rio de Janeiro, Brazil, 2009.

[161] Sumit Shekhar, Vishal M Patel, and Rama Chellappa. Synthesis-based recognition of low resolution faces. In *International Joint Conference on Biometrics*, pages 1–6, Washington, DC, USA, 2011.

[162] Yantao Shen, Hongsheng Li, Shuai Yi, Dapeng Chen, and Xiaogang Wang. Person re-identification with deep similarity-guided graph neural network. In *European Conference on Computer Vision*, pages 486–504, Munich, Germany, 2018.

[163] Kohei Shiraga, Yasushi Makihara, Daigo Muramatsu, Tomio Echigo, and Yasushi Yagi. Geinet: View-invariant gait recognition using a convolutional neural network. pages 1–8, Halmstad, Sweden, 2016.

[164] Assaf Shocher, Nadav Cohen, and Michal Irani. Zero-shot? super-resolution using deep internal learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018.

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

[166] Terence Sim, Simon Baker, and Maan Bsat. The cmu pose, illumination, and expression (pie) database. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 53–58, Washington, D.C., USA, 2002.

[167] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[168] Maneet Singh, Shruti Nagpal, Richa Singh, and Mayank Vatsa. Dual directed capsule network for very low resolution image recognition. In *IEEE International Conference on Computer Vision*, pages 340–349, Seoul, Korea, 2019.

[169] Maneet Singh, Shruti Nagpal, Mayank Vatsa, Richa Singh, and Angshul Majumdar. Identity aware synthesis for cross resolution face recognition. In *Workshop of IEEE Conference on Computer Vision and Pattern Recognition*, pages 479–488, Salt Lake City, UT, USA, 2018.

[170] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Mask-guided contrastive attention model for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1179–1188, Salt Lake City, UT, USA, 2018.

[171] Guanglu Song, Biao Leng, Yu Liu, Congrui Hetang, and Shaofan Cai. Region-based quality estimation network for large-scale person re-identification. In *AAAI Conference on Artificial Intelligence*, New Orleans, Louisiana, USA, 2018.

[172] Jifei Song, Yongxin Yang, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Generalizable person re-identification by domain-invariant mapping network. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 719–728, Long Beach, CA, USA, 2019.

[173] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation by joint identification-verification. In *Advances in Neural Information Processing Systems*, Montreal, Quebec, Canada, 2014.

[174] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation by joint identification-verification. In *Advances in Neural Information Processing Systems*, Montreal, Quebec, Canada, 2014.

[175] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation from predicting 10,000 classes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1891–1898, Columbus, OH, USA, 2014.

[176] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation from predicting 10,000 classes. In *IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, 2014.

[177] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deeply learned face representations are sparse, selective, and robust. In *IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, 2015.

[178] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *European Conference on Computer Vision*, pages 480–496, Munich, Germany, 2018.

[179] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, Boston, MA, USA, 2015.

[180] Ying Tai, Jian Yang, and Xiaoming Liu. Image super-resolution via deep recursive residual network. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3147–3155, Columbus, OH, USA, 2017.

[181] Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. In *International Conference on Learning Representations*, Toulon, France, 2017.

[182] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, Columbus, OH, USA, 2014.

[183] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Web-scale training for face identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, 2015.

[184] Tencent. Youtu lab, tencent. `http://bestimage.qq.com/`, 2017.

[185] Antonio Torralba, Rob Fergus, and William T Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(11), 2008.

[186] Chen Wang, Junping Zhang, Liang Wang, Jian Pu, and Xiaoru Yuan. Human identification using temporal information preserving gait template. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2164–2176, 2011.

[187] Dayong Wang, Charles Otto, and Anil K Jain. Face search at scale. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1122–1136, 2016.

[188] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *ACM International Conference on Multimedia*, pages 274–282, Seoul, Korea, 2018.

[189] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2275–2284, Salt Lake City, Utah, USA, 2018.

[190] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by video ranking. In *European Conference on Computer Vision*, pages 688–703, Zurich, Switzerland, 2014.

[191] Xiaogang Wang and Xiaoou Tang. Face hallucination and recognition. In *International Conference on Audio-and Video-Based Biometric Person Authentication*, pages 486–494, Rye Brook, NY, USA, 2003.

[192] Xiaogang Wang and Xiaoou Tang. Hallucinating face by eigentransformation. 35(3), 2005.

[193] Zhangyang Wang, Shiyu Chang, Yingzhen Yang, Ding Liu, and Thomas S Huang. Studying very low resolution recognition using deep networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4792–4800, Las Vegas, NV, USA, 2016.

[194] Zheng Wang, Ruimin Hu, Yi Yu, Junjun Jiang, Chao Liang, and Jinqiao Wang. Scale-adaptive low-resolution person re-identification via learning a discriminating surface. In *International Joint Conference of Artificial Intelligence*, pages 2669–2675, New York, NY, USA, 2016.

[195] Zheng Wang, Mang Ye, Fan Yang, Xiang Bai, and Shin'ichi Satoh. Cascaded sr-gan for scale-adaptive low resolution person re-identification. In *International Joint Conference of Artificial Intelligence*, pages 3891–3897, Stockholm, Sweden, 2018.

[196] Zhifei Wang, Zhenjiang Miao, QM Jonathan Wu, Yanli Wan, and Zhen Tang. Low-resolution face recognition: a review. *The Visual Computer*, 30(4):359–386, 2014.

[197] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[198] Xin Wei, Hui Wang, Bryan Scotney, and Huan Wan. Minimum margin loss for deep face recognition. *Pattern Recognition*, 97:107012, 2020.

[199] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, pages 499–515, Amsterdam, The Netherlands, 2016.

[200] Cameron Whitelam, Emma Taborsky, Austin Blanton, Brianna Maze, Jocelyn Adams, Tim Miller, Nathan Kalka, Anil K Jain, James A Duncan, Kristen Allen, et al. Iarpa janus benchmark-b face dataset. In *Workshop of IEEE Conference on Computer Vision and Pattern Recognition*, pages 90–98, Honolulu, Hawaii, USA, 2017.

[201] Lior Wolf, Tal Hassner, and Itay Maoz. Face recognition in unconstrained videos with matched background similarity. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 529–534, Colorado Springs, CO, USA, 2011.

[202] Yongkang Wong, Conrad Sanderson, Sandra Mau, and Brian C Lovell. Dynamic amelioration of resolution mismatches for local feature based identity inference. In *IEEE International Conference on Pattern Recognition*, pages 1200–1203, Istanbul, Turkey, 2010.

[203] Ancong Wu, Wei-Shi Zheng, and Jian-Huang Lai. Unsupervised person re-identification by camera-aware similarity consistency learning. In *IEEE International Conference on Computer Vision*, pages 6922–6931, Seoul, Korea, 2019.

[204] Guile Wu, Xiatian Zhu, and Shaogang Gong. Tracklet self-supervised learning for unsupervised person re-identification. In *AAAI Conference on Artificial Intelligence*, New York City, NY, USA, 2020.

[205] Zifeng Wu, Yongzhen Huang, Liang Wang, Xiaogang Wang, and Tieniu Tan. A comprehensive study on cross-view gait based human identification with deep cnns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(2):209–226, 2016.

[206] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1249–1258, Las Vegas, NV, USA, 2016.

[207] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3415–3424, Columbus, OH, USA, 2017.

[208] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *IEEE International Conference on Computer Vision*, pages 1395–1403, Santiago, Chile, 2015.

[209] Xiangyu Xu, Deqing Sun, Jinshan Pan, Yujin Zhang, Hanspeter Pfister, and Ming-Hsuan Yang. Learning to super-resolve blurry face and text images. In *IEEE International Conference on Computer Vision*, pages 251–260, Venice, Italy, 2017.

[210] Chih-Yuan Yang, Sifei Liu, and Ming-Hsuan Yang. Hallucinating compressed face images. *International Journal of Computer Vision*, 126(6):597–614, 2018.

[211] Chih-Yuan Yang, Chao Ma, and Ming-Hsuan Yang. Single-image super-resolution: A benchmark. In *European Conference on Computer Vision*, pages 372–386, Zurich, Switzerland, 2014.

[212] Jianchao Yang, John Wright, Thomas S Huang, and Yi Ma. Image super-resolution via sparse representation. *IEEE Transactions on Image Processing*, 19(11), 2010.

[213] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016.

[214] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

[215] Zili Yi, Hao (Richard) Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *IEEE International Conference on Computer Vision*, Venice, Italy, 2017.

[216] Zhou Yin, Wei-Shi Zheng, Ancong Wu, Hong-Xing Yu, Hai Wan, Xiaowei Guo, Feiyue Huang, and Jianhuang Lai. Adversarial attribute-image person re-identification. *arXiv e-print*, 2017.

[217] Hong-Xing Yu, Ancong Wu, and Wei-Shi Zheng. Cross-view asymmetric metric learning for unsupervised person re-identification. In *IEEE International Conference on Computer Vision*, pages 994–1002, Venice, Italy, 2017.

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

Conference on Computer Vision and Pattern Recognition, pages 2148–2157, Long Beach, CA, USA, 2019.

[219] Xin Yu, Basura Fernando, Richard Hartley, and Fatih Porikli. Super-resolving very low-resolution face images with supplementary attributes. In *IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018.

[220] Xin Yu and Fatih Porikli. Ultra-resolving face images by discriminative generative networks. In *European Conference on Computer Vision*, pages 318–333, Amsterdam, The Netherlands, 2016.

[221] Xin Yu and Fatih Porikli. Face hallucination with tiny unaligned images by transformative discriminative neural networks. In *AAAI Conference on Artificial Intelligence*, San Francisco, California, USA, 2017.

[222] Xin Yu and Fatih Porikli. Hallucinating very low-resolution unaligned and noisy face images by transformative discriminative autoencoders. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3760–3768, Honolulu, Hawaii, USA, 2017.

[223] Baochang Zhang, Shiguang Shan, Xilin Chen, and Wen Gao. Histogram of gabor phase patterns (hgpp): A novel object representation approach for face recognition. *IEEE Transactions on Image Processing*, 16(1):57–68, 2006.

[224] Kaipeng Zhang, Zhanpeng Zhang, Chia-Wen Cheng, Winston H Hsu, Yu Qiao, Wei Liu, and Tong Zhang. Super-identity convolutional neural network for face hallucination. In *European Conference on Computer Vision*, pages 183–198, Munich, Germany, 2018.

[225] Ning Zhang, Manohar Paluri, Yaniv Taigman, Rob Fergus, and Lubomir Bourdev. Beyond frontal faces: Improving person recognition using multiple cues. In *IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, 2015.

[226] Ning Zhang, Manohar Paluri, Yaniv Taigman, Rob Fergus, and Lubomir Bourdev. Beyond frontal faces: Improving person recognition using multiple cues. In *IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, 2015.

[227] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Densely semantically aligned person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 667–676, Long Beach, CA, USA, 2019.

[228] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, and Xiaoou Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1077–1085, Columbus, OH, USA, 2017.

[229] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. Deeply-learned part-aligned representations for person re-identification. In *IEEE International Conference on Computer Vision*, pages 3219–3228, Venice, Italy, 2017.

[230] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *IEEE International Conference on Computer Vision*, pages 1116–1124, Santiago, Chile, 2015.

[231] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Reidentification by relative distance comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[232] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Towards open-world person re-identification by one-shot group-based verification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(3):591–606, 2015.

[233] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 598–607, Long Beach, CA, USA, 2019.

[234] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. Camera style adaptation for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5157–5166, Salt Lake City, UT, USA, 2018.

[235] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *IEEE International Conference on Computer Vision*, Seoul, Korea, 2019.

[236] Ruofan Zhou and Sabine Susstrunk. Kernel modeling super-resolution on real low-resolution images. In *IEEE International Conference on Computer Vision*, pages 2433–2443, Seoul, Korea, 2019.

[237] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision*, pages 2223–2232, Venice, Italy, 2017.

[238] Shizhan Zhu, Sifei Liu, Chen Change Loy, and Xiaoou Tang. Deep cascaded bi-network for face hallucination. In *European Conference on Computer Vision*, pages 614–630, Amsterdam, The Netherlands, 2016.

[239] Xiangping Zhu, Xiatian Zhu, Minxian Li, Vittorio Murino, and Shaogang Gong. Intra-camera supervised person re-identification: A new benchmark. In *Workshop of IEEE International Conference on Computer Vision*, pages 0–0, Seoul, Korea, 2019.

[240] Xiatian Zhu. *Semantic Structure Discovery in Surveillance Videos*. PhD thesis, Queen Mary University of London, 2016.

[241] Wilman WW Zou and Pong C Yuen. Very low resolution face recognition problem. *IEEE Transactions on Image Processing*, 21(1):327–340, 2011.