Characteristic Regularisation for Super-Resolving Face Images

Zhiyi Cheng
Queen Mary University of London
z.cheng@qmul.ac.uk

Xiatian Zhu
Vision Semantics Limited, London, UK
eddy.zhuxt@gmail.com

Shaogang Gong
Queen Mary University of London
s.gong@qmul.ac.uk

Abstract

Existing facial image super-resolution (SR) 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, we formulate a method that joins the advantages of conventional SR and UDA models. Specifically, we separate and control the optimisations for characteristics consistifying and image super-resolving by introducing Characteristic Regularisation (CR) between them. This task split makes the model training more effective and computationally tractable. Extensive evaluations demonstrate the performance superiority of our method over state-of-the-art SR and UDA models on both genuine and artificial LR facial imagery data.

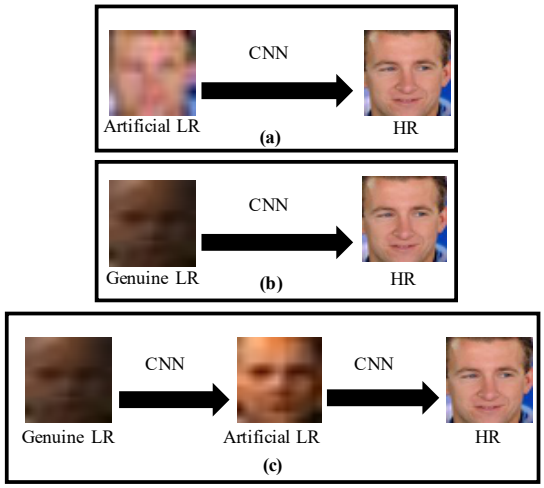

Figure 1. CNN architectures for facial image super-resolution. **(a)** A CNN is trained to super-resolve *artificial* LR facial images that are produced by down-sampling [11, 21]. It is a supervised learning method. **(b)** A CNN learns to adapt a *genuine* LR facial images into the HR style. Without LR-HR pairing supervision, a cycle consistency based loss function is often used for model training [40, 36, 5]. **(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.

1. Introduction

Facial image analysis [32, 6, 2] is significant for many computer vision applications in business, law enforcement, and public security [26]. However, the model performance often degrades significantly when the face image resolution is very low. Face image super-resolution (SR) [1] provides a viable solution by recovering a high-resolution (HR) face image from its low-resolution (LR) counterpart. Existing state-of-the-art image SR models [8, 37, 41] 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(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. 3). 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 [26].

Unsupervised domain adaptation (UDA) methods are possible solutions considering genuine LR and HR images as two different domains. UDA techniques have achieved remarkable success [40, 15, 31, 25, 3, 36, 18, 23]. A representative modelling idea is to exploit cycle consistency loss functions between two unpaired domains (Fig 1(b)) [40, 36, 18]. 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 [5] 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 our context, 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 [22, 7]. In addition, jointly training two connected CNN models has to be conducted very carefully, along with the difficulty of training GAN models [12]. 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.

In this work, we solve the problem of super-resolving genuine LR facial images by formulating a ***Characteristic Regularisation*** (CR) method (Fig 1(c)). In contrast to conventional image SR methods, we particularly leverage the unpaired genuine LR images in order to take into account their characteristics information for facilitating model optimisation. Unlike cycle consistency based UDA methods, we instead *leverage the artificial LR images as regularisation target* in order to separately learn the tasks of characteristic consistifying and image super-resolution. Specifically, we perform 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, we consider 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, we formulate an unsupervised semantic adaptation loss by aligning with the face recognition feature distribution of auxiliary HR images.

Our 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 [22], providing auxiliary supervision improves accuracy and convergence speed [30]. In our case 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.

The **contributions** of this work are as follows: (1) We propose 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. (2) The proposed *Characteristic Regularisation* enables computationally more tractable model training and better model generalisation capability. (3) We introduce a new unsupervised learning loss function without the limitations of cycle consistency. We conduct extensive experiments 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 our model over the state-of-the-art image SR and domain adaptation methods.

2. Related Work

Image super-resolution. Most state-of-the-art image SR methods usually learn the resolution mapping functions with artificially down-sampled LR and ground-truth HR image pairs [11, 17, 19, 24, 38, 41]. Such pairwise supervised learning becomes infeasible when there is no HR-LR training pairs, e.g. genuine low-quality facial imagery data from in-the-wild social media and surveillance videos. Recently, Generative Adversarial Networks (GANs) based image SR models [8, 21, 37] have been developed. They additionally exploit an unsupervised adversarial learning loss on top of the conventional MSE loss. These GAN methods often produce more photo-realistic and visually appealing images. While the GAN loss is unsupervised, these methods still heavily rely on the paired LR and HR training data therefore remaining unsuitable for genuine facial image SR. Most of the existing works consider mostly an *artificial* image SR

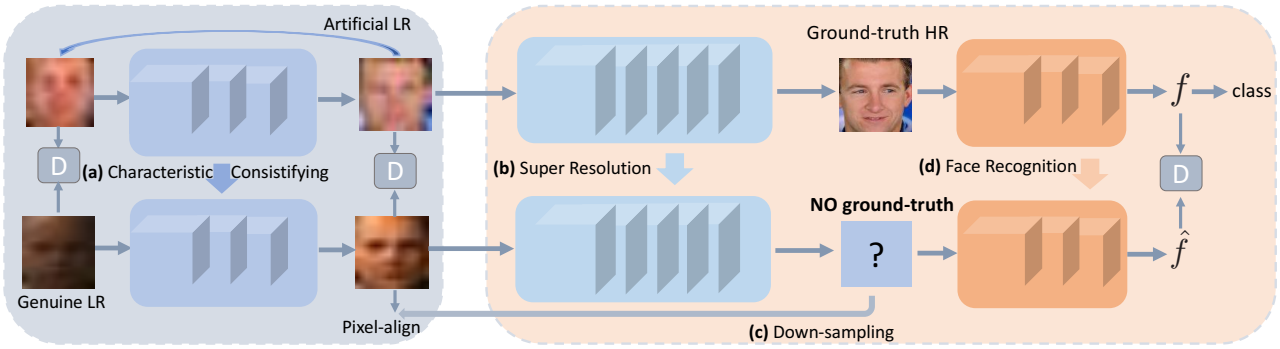

Figure 2. An overview of the proposed *Characteristics Regularisation* (CR) approach for super-resolving genuine LR facial imagery data. The CR model performs multi-task learning. **(a)** The auxiliary task is *characteristic consistifying* in order to transform genuine LR images into the artificial 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, we propose to align 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, we formulate an unsupervised semantic adaptation loss term in the adversarial learning spirit, w.r.t. a supervised face recognition model trained on auxiliary HR images.

problem. The LR images are *synthesised* by pre-defined down-sampling processes, with different imaging characteristics to *genuine* LR images. When applied to genuine LR images, this acquisition difference causes the domain shift problem for the existing supervised SR models trained by artificial LR images. Learning to super-resolve genuine LR imagery is much harder due to no paired HR ground-truth available for model training.

Super-resolution for genuine imagery. There are a few recent attempts on resolving genuine image SR [4, 5, 9, 10, 28]. In particular, Shocher et al. [28] 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. Bulat and Tzimiropoulos [4] 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. [5] 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 [40, 36, 18] 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. [9, 10] 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, we formulate a unified method that enjoy the strengths of both conventional SR and UDA methods in a principled manner. In particular, we separate 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.

3. Method

We aim 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, we have 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 $I^{\text{hr}}_{\text{aux}}$. We first give an overview of existing image SR models before introducing the proposed characteristic regularisation method.

3.1. Facial Image Super-Resolution

Given auxiliary HR facial images $I^{\text{hr}}_{\text{aux}}$, we can easily generate corresponding LR images $I^{\text{lr}}_{\text{aux}}$ by down-sampling. With such paired data, we can train a common supervised image SR CNN model optimised by some pixel alignment loss constraint such as the Mean-Squared Error (MSE) between the resolved and ground-truth images [21]:

$$\mathcal{L}_{\text{sr}} = \|I^{\text{hr}}_{\text{aux}} - \phi_{\text{sr}}(I^{\text{lr}}_{\text{aux}}))\|_2^2 \quad (1)$$

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

$$I^{\text{sr}} = \phi_{\text{sr}}(I^{\text{lr}}_{\text{aux}}) \tag{2}$$

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.

3.2. Characteristics Regularisation

To address the domain gap in SR, we take a divide-and-conquer strategy: first characteristic consistifying, then image super-resolving. 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, we exploit the unsupervised GAN learning framework [12]. 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:

$$\begin{aligned}\mathcal{L}_{\text{gan}} = \mathbb{E}_{I^{\text{lr}}_{\text{aux}}}[\log D(I^{\text{lr}}_{\text{aux}})] + \\ \mathbb{E}_{I^{\text{lr}}}[\log\left(1 - D(\phi_{\text{cr}}(I^{\text{lr}}))\right)],\end{aligned} \tag{3}$$

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:

$$G^* = \arg\min_{\phi_{\text{cr}}} \max_{D} \mathcal{L}_{\text{gan}}. \tag{4}$$

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

$$\begin{aligned}\arg\min_{\tilde{\phi}_{\text{cr}}} \max_{\tilde{D}} \mathbb{E}_{I^{\text{lr}}}[\log \tilde{D}(I^{\text{lr}})] + \\ \mathbb{E}_{I^{\text{lr}}_{\text{aux}}}[\log\left(1 - \tilde{D}(\tilde{\phi}_{\text{cr}}(I^{\text{lr}}_{\text{aux}}))\right)],\end{aligned} \tag{5}$$

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

$$\mathcal{L}_{\text{cr}} = ||I^{\text{lr}}_{\text{aux}} - \phi_{\text{cr}}(\tilde{\phi}_{\text{cr}}(I^{\text{lr}}_{\text{aux}}))||_2^2 + \lambda\mathcal{L}_{\text{gan}} \tag{6}$$

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

3.3. 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, we need to address the problem of lacking HR supervision. Instead of leveraging the conventional cycle consistency idea, we adopt 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, we do not need to access the unknown HR ground-truth. Formally, we design this SR loss function for regulated LR images as:

$$\mathcal{L}_{\text{cr-sr}} = ||f_{\text{DS}}\Big(\phi_{\text{sr}}\big(\phi_{\text{cr}}(I^{\text{lr}})\big)\Big) - \phi_{\text{cr}}(I^{\text{lr}})||_2^2 \tag{7}$$

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

3.4. Unsupervised Semantic Adaptation

Apart from visual fidelity, the SR output is also required to be semantically meaningful with good identity information. To this end, we form an unsupervised semantic adaptation loss term 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 $I^{\text{hr}}_{\text{aux}}$. It is formally written as:

$$\begin{aligned}\mathcal{L}_{\text{cr-gan}} = \mathbb{E}_{I^{\text{hr}}_{\text{aux}}}[\log D'(\phi_{\text{fr}}(I^{\text{hr}}_{\text{aux}}))] + \\ \mathbb{E}_{\phi_{\text{sr}}(\phi_{\text{cr}}(I^{\text{lr}}))}\Big[\log\Big(1 - D'\big(\phi_{\text{fr}}\big(\phi_{\text{sr}}(\phi_{\text{cr}}(I^{\text{lr}}))\big)\big)\Big)\Big]\end{aligned} \tag{8}$$

where ϕ_{fr} is a CentreFace [34] based feature extractor pre-trained with $I^{\text{hr}}_{\text{aux}}$. 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.

3.5. Model Training and Inference

Due to introduction of characteristic regularisation in the middle of our full model, more effective model training is enabled. It facilitates a two-staged training strategy. In the first stage, we pre-train the CNN for image SR 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. We train the image SR model ϕ_{sr} as [21] by deploying the pixel-wise MSE loss

Figure 3. Examples of genuine facial images randomly sampled from LR-PIPA (**top**) and LR-DukeMTMC (**bottom**).

function (Eq (1)). This model training benefits from the normal adversarial loss for achieving better perceptual quality. Other existing image SR methods [17, 11] can be readily considered in our framework.

CNN for characteristic regularisation. We train the CNN for characteristic regularisation ϕ_{cr} by an adversarial loss and a pixel-wise loss jointly (Eq (6)).

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

$$\mathcal{L} = \mathcal{L}_{\text{sr}} + \lambda_{\text{cr}}\mathcal{L}_{\text{cr}} + \lambda_{\text{cr-sr}}\mathcal{L}_{\text{cr-sr}} + \lambda_{\text{cr-gan}}\mathcal{L}_{\text{cr-gan}} \quad (9)$$

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

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

4. Experiments

Datasets. For model performance evaluation, we created two new real-world genuine LR facial image datasets sampled from web social-media imagery and surveillance videos. Following [5], we define LR faces as those with an average size of $\leq$16$\times$16 pixels. In particular, we constructed the web social-media based real-world face image dataset by assembling LR faces from the People In Photo Albums (PIPA) benchmark [39], called ***LR-PIPA***. Similarly, we collected LR face images (small faces) from a multi-target multi-camera tracking benchmark DukeMTMC [27] and built our surveillance video real-world face dataset, called ***LR-DukeMTMC***. There are 8,641 and 7,044 face images in LR-PIPA and LR-DukeMTMC, respectively. All the face images were obtained by deploying the TinyFace detector [16]. We manually filtered out non-face images. 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. We will release both datasets publicly. We show some randomly selected examples in Fig 3.

Training and test data. To effectively train a competing model, we need both real-world genuine LR images and web auxiliary HR facial images. For the former, we used 153,440 LR face images collected from the Wider Face benchmark [35]. 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, we selected the standard CelebA benchmark with 202,599 HR web facial images [29]. Such a training set design ensures that each model can be trained with sufficiently diverse data to minimise the learning bias. For model test, we utilised the entire LR-PIPA and LR-DukeMTMC. 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. We utilise the Frechet Inception Distance (***FID***) [14] to assess the quality of resolved face images, similar to the state-of-the-art method [5]. Specifically, ***FID*** is measured by the Frechet Distance between two multivariate Gaussian distributions.

Implementation details. We performed all the following experiments in Tensorflow. We used the residual blocks [13] as the backbone unit of our network. In particular, we used 3 residual blocks in the net for the characteristics regularisation module ϕ_{cr} and $\tilde{\phi}_{\text{cr}}$, and we further adapted the SRGAN (3 groups containing 12/3/2 residual blocks, respectively. Resolution is increased 2 times across each group) [21] for our facial SR module ϕ_{sr}. The adversarial discriminator for ϕ_{cr} and $\tilde{\phi}_{\text{cr}}$ 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. We set the learning rate to 10^{-4}, the batch size to 16. The SR module (ϕ_{sr} in Fig. 2) 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	LR-PIPA	LR-DukeMTMC
VDSR [17]	94.49	229.56
SRGAN [21]	103.85	232.38
FSRNet [8]	117.19	218.30
SICNN [38]	129.23	223.08
CycleGAN [40]	33.62	42.41
CSRI [9]	104.68	240.99
LRGAN [5]	29.80	31.20
CR (Ours)	**23.09**	**25.56**

Table 1. Comparing the image quality on *genuine* LR facial image super-resolution. Metric: FID. **Lower is better.**

4.1. Test Genuine Low-Resolution Facial Images

Competitors. To evaluate the effectiveness of our CR model for genuine facial image SR, we compared with four *groups* of the state-of-the-art methods including, two generic image SR models (VDSR [17], SRGAN [21]), one image-to-image translation model (CycleGAN [40]), one *non*-genuine face SR model (FSRNet [8]), one UDA-based genuine face SR model (LRGAN [5]), and one facial identity-guided genuine SR model (CSRI [9]). Same as our 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 1. We have the following observations: **(1)** The proposed CR model achieves the best FID score among all the competitors, suggesting the overall performance advantage of our 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 no 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 our method.

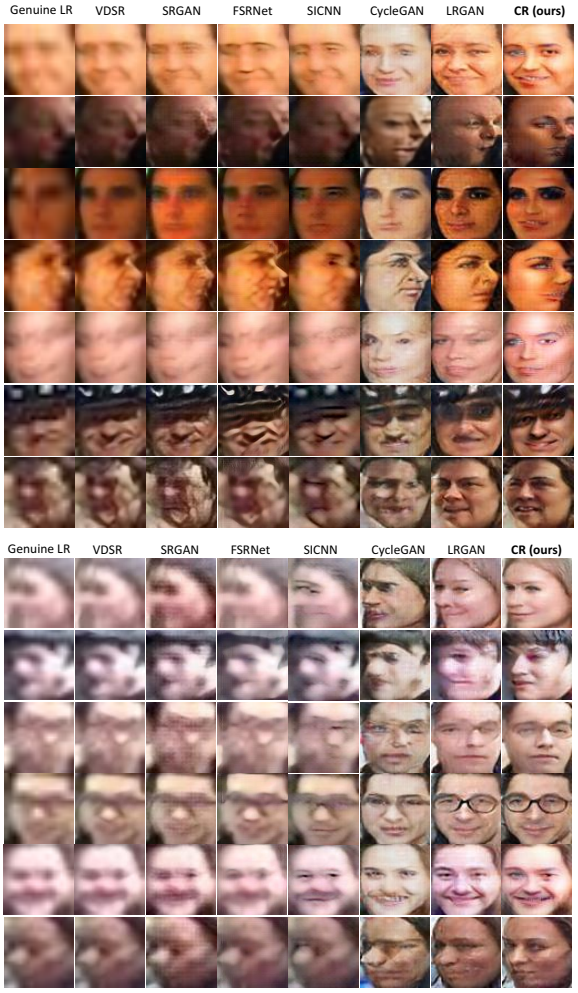

Figure 4. Examples of genuine LR image super-resolution on (**top**) LR-PIPA and (**bottom**) LR-DukeMTMC.

Qualitative evaluation. To conduct visual comparisons between different alternative methods, we provided SR results of random genuine LR facial images in Fig 4 . Overall, the visual examination is largely consistent with the nu-

merical 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 our method in super-resolving genuine LR facial images.

Model complexity. We compared the top-3 models (CR, LRGAN [5], and CycleGAN [40]) 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, our model is the most compact and most efficient.

4.2. Face Recognition on Genuine LR Face Imagery

We tested the benefit of image SR on a downstream task, *face recognition*, on the LR-PIPA dataset. We used the CentreFace model trained on the auxiliary HR images and the CMC rank metrics. The results in Table 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.

Dataset	Rank-1 (%)
VDSR [17]	25.45
SRGAN [21]	27.00
FSRNet [8]	26.50
SICNN [38]	28.85
CycleGAN [40]	25.12
CSRI [9]	29.59
LRGAN [5]	21.99
CR (Ours)	**30.53**
Raw LR input	24.83

Table 2. Face recognition performance on super-resolved *genuine* LR images from LR-PIPA. Metric: Rank-1. **Higher is better.**

4.3. Test Artificial Low-Resolution Facial Images

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

Model deployment. By design, our 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 we selected the Helen face dataset [20] with 2,330 images. We produced the artificial LR test images by bicubic down-sampling, as the conventional SR evaluation setting.

Metrics. For performance evaluation, we used the common Peak Signal-to-Noise Ratio (PSNR) and structure similarity index (SSIM) [33]. This is because, we have the ground-truth HR images for pixel-level assessment in this case.

Metric	PSNR	SSIM
VDSR [17]	**26.31**	0.7918
SRGAN [21]	25.10	0.7873
FSRNet [8]	25.10	0.7234
SICNN [38]	26.10	0.7986
CycleGAN [40]	18.85	0.6061
CSRI [9]	25.40	0.7388
LRGAN [5]	21.88	0.6869
CR (Ours)	25.50	**0.8184**

Table 3. Comparison of state-of-the-art methods on *artificial* LR facial image super-resolution. Dataset: Helen. Metric: PSNR & SSIM. **Higher is better.**

Results. Table 3 compares the performances on normal Helen LR facial images of our CR and state-of-the-art SR methods. It is observed that our method can generate better results than all the competitors except VDSR for the PSNR metric. Interestingly, CR outperforms SRGAN which is actually our SR module. This implies that the model generalisation for conventional SR tasks can be improved by the proposed unsupervised SR learning objective (Eq (7)).

4.4. Component Analysis and Discussion

We conducted a series of model component analysis for giving insights to our model performance.

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

Table 4. Effect of characteristics regularisation (CR). Metric: FID.

Characteristic regularisation. We evaluated the effect and benefits of characteristic regularisation (CR) on model performance. We 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 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

We further examined the result of characteristic consistifying, i.e. the regulated LR images. To this end, we measured the FID between artificial and regulated LR images, in comparison to that between artificial and genuine LR images. Table 5 shows that although regulated LR images match significantly better to artificial LR data than their

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

Table 5. Evaluate regulated LR images (R-LR). G-LR: Genuine LR images; A-LR: Artificial LR images.

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).

We showed qualitative results Fig. 5. 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.

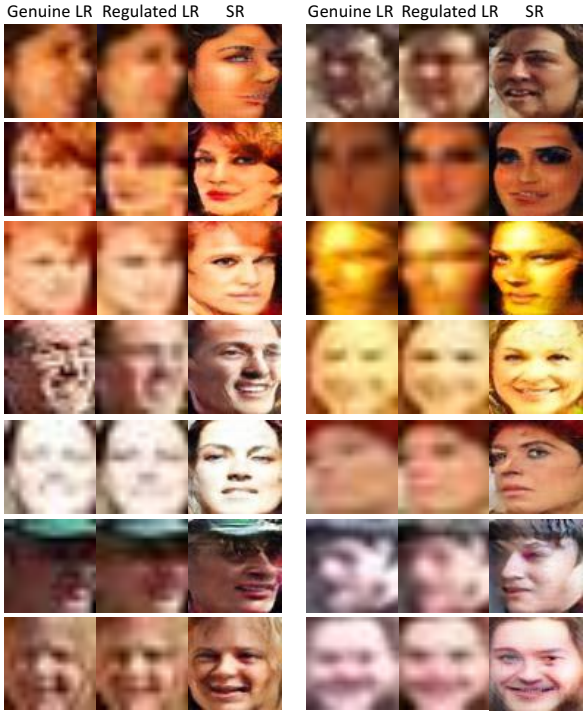

Figure 5. Genuine LR *vs.* regulated LR *vs.* resolved face images.

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

Table 6. Effect of super-resolving regulated images (SR-RI). Metric: FID. **Lower is better.**

Super-resolving regulated images. In the second training stage, we fine-tune the SR module for better super-resolving regulated LR images. We evaluated the effect of this design. Table 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 5.

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

Table 7. Effect of unsupervised loss (UL) for super-resolving regulated images. Metric: FID. **Lower is better.**

Figure 6. Visual examination: W/O *vs.* W/ unsupervised loss (UL) in super-resolving regulated images.

Recall that we introduce an unsupervised SR loss (Eq (7)) for regulated LR images due to no HR ground-truth. We consider pixel-wise alignment in LR image space, without the need for cycle consistency. We tested its impact on the model performance. Table 7 shows that applying this loss can clearly boost the fidelity quality of resolved faces. Also, we found that it makes the model training more stable. Further qualitative evaluation in Fig 6 shows that the unsupervised SR loss can help reduce the noise and distortion in SR, leading to visually more appealing results.

5. Conclusion

We present a Characteristic Regularised (CR) method for super-resolving genuine LR facial imagery. 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. We conduct extensive comparative experiments on both genuine and artificial LR facial images. The results show the performance and generalisation advantages of our method over a variety of state-of-the-art image SR and UDA models. We carry out detailed model component analysis for revealing the model formulation insights.

Acknowledgement

This work was partially supported by the Alan Turing Institute Turing Fellowship, the Innovate UK Industrial Challenge Project on Developing and Commercialising Intelligent Video Analytics Solutions for Public Safety (98111-571149), Vision Semantics Ltd, and SeeQuestor Ltd.

References

[1] S. Baker and T. Kanade. Hallucinating faces. *FG*, 2000:83–88, 2000.

[2] M. S. Bartlett, G. Littlewort, I. Fasel, and J. R. Movellan. Real time face detection and facial expression recognition: Development and applications to human computer interaction. In *2003 Conference on computer vision and pattern recognition workshop*, volume 5, pages 53–53, 2003.

[3] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. 2017.

[4] A. Bulat and G. Tzimiropoulos. Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans. 2018.

[5] A. Bulat, J. Yang, and G. Tzimiropoulos. To learn image super-resolution, use a gan to learn how to do image degradation first. 2018.

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

[7] S. Chandar, C. Sankar, E. Vorontsov, S. E. Kahou, and Y. Bengio. Towards non-saturating recurrent units for modelling long-term dependencies. In *AAAI*, 2019.

[8] Y. Chen, Y. Tai, X. Liu, C. Shen, and J. Yang. Fsrnet: End-to-end learning face super-resolution with facial priors. 2018.

[9] Z. Cheng, X. Zhu, and S. Gong. Low-resolution face recognition. In *Asian Conference on Computer Vision*, pages 605–621. Springer, 2018.

[10] Z. Cheng, X. Zhu, and S. Gong. Surveillance face recognition challenge. *arXiv preprint arXiv:1804.09691*, 2018.

[11] C. Dong, C. C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. 38(2):295–307, 2016.

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. 2014.

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. 2016.

[14] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, G. Klambauer, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a nash equilibrium. 2017.

[15] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell. CyCADA: Cycle-consistent adversarial domain adaptation. 2018.

[16] P. Hu and D. Ramanan. Finding tiny faces. pages 1522–1530, 2017.

[17] J. Kim, J. Kwon Lee, and K. Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654, 2016.

[18] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim. Learning to discover cross-domain relations with generative adversarial networks. 2017.

[19] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang. Deep laplacian pyramid networks for fast and accurate superresolution. volume 2, page 5, 2017.

[20] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang. Interactive facial feature localization. In *European conference on computer vision*, pages 679–692. Springer, 2012.

[21] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. volume 2, page 4, 2017.

[22] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu. Deeply-supervised nets. In *Artificial Intelligence and Statistics*, pages 562–570, 2015.

[23] M.-Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. 2017.

[24] X. Mao, C. Shen, and Y.-B. Yang. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. pages 2802–2810, 2016.

[25] Z. Murez, S. Kolouri, D. Kriegman, R. Ramamoorthi, and K. Kim. Image to image translation for domain adaptation. 2018.

[26] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

[27] E. Ristani and C. Tomasi. Tracking multiple people online and in real time. In *Asian Conference on Computer Vision*, pages 444–459. Springer, 2014.

[28] A. Shocher, N. Cohen, and M. Irani. Zero-shot? super-resolution using deep internal learning. In *Conference on computer vision and pattern recognition (CVPR)*, 2018.

[29] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. 2014.

[30] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[31] Y. Taigman, A. Polyak, and L. Wolf. Unsupervised cross-domain image generation. 2017.

[32] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014.

[33] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[34] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, pages 499–515. Springer, 2016.

[35] S. Yang, P. Luo, C. C. Loy, and X. Tang. Wider face: A face detection benchmark. 2016.

[36] Z. Yi, H. R. Zhang, P. Tan, and M. Gong. Dualgan: Unsupervised dual learning for image-to-image translation. 2017.

[37] X. Yu, B. Fernando, R. Hartley, and F. Porikli. Super-resolving very low-resolution face images with supplementary attributes. 2018.

[38] K. Zhang, Z. Zhang, C.-W. Cheng, W. H. Hsu, Y. Qiao, W. Liu, and T. Zhang. Super-identity convolutional neural

network for face hallucination. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 183–198, 2018.

[39] N. Zhang, M. Paluri, Y. Taigman, R. Fergus, and L. Bourdev. Beyond frontal faces: Improving person recognition using multiple cues. 2015.

[40] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. 2017.

[41] S. Zhu, S. Liu, C. C. Loy, and X. Tang. Deep cascaded bi-network for face hallucination. In *European Conference on Computer Vision*, pages 614–630. Springer, 2016.