S^2LD: Semi-Supervised Landmark Detection in Low Resolution Images and Impact on Face Verification

Amit Kumar, Rama Chellappa
Center for Automation Research, UMIACS, University of Maryland, College Park
{akumar14, rama}@umiacs.umd.edu

Abstract

Landmark detection algorithms trained on high resolution images perform poorly on datasets containing low resolution images. This degrades the performance of facial verification, recognition and modeling that rely on accurate detection of landmarks. To the best of our knowledge, there is no dataset consisting of low resolution face images along with their annotated landmarks, making supervised training infeasible. In this paper, we present a semi-supervised approach to predict landmarks on low resolution images by learning them from labeled high resolution images. The objective of this work is to show that predicting landmarks directly on low resolution images is more effective than the current practice of aligning images after rescaling or super-resolution. In a two-step process, the proposed approach first learns to generate low resolution images by modeling the distribution of target low resolution images. In the second stage, the model learns to predict landmarks for target low resolution images from generated low resolution images. With extensive experimentation, we study the impact of the various design choices and also show that prediction of landmarks directly in low resolution, improves performance on the critical task of face verification in low resolution images. As a byproduct, the proposed method also achieves competitive land mark detection results for high resolution images, with a single U-Net.

1. Introduction

Convolution Neural Networks (CNNs) have revolutionized the computer vision field, to the point that current systems can recognize faces with more than 99.7% accuracy or achieve detection, segmentation and pose estimation results up to sub-pixel accuracy. However, CNN-based methods assume access to good quality images. ImageNet [23], CASIA [31] or 300W[24] datasets all consist of high resolution images. As a result of domain shift much lower performance is observed when networks trained on these datasets

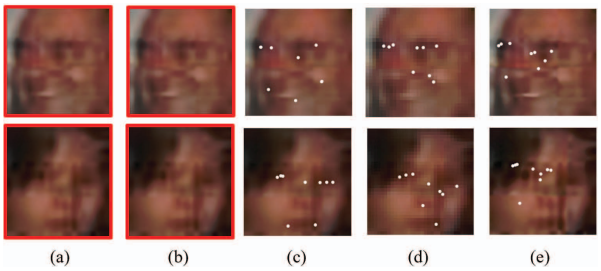

Figure 1: Inaccurate landmark detections on LR images. We show landmark predicted by different systems. (a) MTCNN and (b) 2D-FAN are not able to detect any face in the LR image. (c) Bilinear upsampling. (d) Output from a network trained on downsampled version of HR images. (e) Landmark detection using super-resolved images. $*$ Results from the proposed method in Figure 6.

are applied to images which have suffered degradation due to intrinsic or extrinsic factors. In this work, we address landmark localization in low resolution images and evaluate its impact on face verification. Although, we use face images in our case, the proposed method is also applicable to other tasks, such as human pose estimation. Throughout this paper we use **HR and LR to denote high and low resolutions** respectively.

Facial landmark localization, also known as key-point or fiducial detection, refers to the task of detecting specific points such as eye corners and nose tip on a face image. The detected key-points are used to align images to canonical coordinates for further processing. It has been experimentally shown in [2], that accurate face alignment leads to improved performance in face verification. Though great strides have been made in this direction, mainly addressing large-pose face alignment, landmark localization for LR images, still remains an understudied problem, mostly because of the absence of large scale labeled dataset(s). To the best of our knowledge, Semi-Supervised Landmark Detection (S^2LD) is the first attempt to address landmark localization directly on LR images.

Main motivation: In Figure 1, we examine possible sce-

2160-7516/20/$31.00 ©2020 IEEE
DOI 10.1109/CVPRW50498.2020.00387

narios which are currently practiced for LR images. Figure 1 shows the predicted landmarks when the input image size is less than 32×32 pixels. Typically, landmark detection networks are trained with 224×224 crops of HR images taken from AFLW [13] and 300W datasets. During inference, irrespective of resolution, the input image is re-scaled to 224×224. We deploy two methods: MTCNN [33] and 2D-FAN [4], which have detection and localization built in a single system. In Figures 1(a) and (b) we see these networks fail to detect face in the given image. Figure 1(c), shows the output when a Landmark Detector network trained on HR images (HR-LD) is applied to a re-scaled LR one; It is important to note that this network achieves state of the art performance on AFLW and 300W test sets. A possible solution is to train this network on sub-sampled images as a substitute for LR images. Figure 1(d) shows the output of one such network. It is evident from these experiments that networks trained with HR images or sub-sampled images are not effective for genuine LR images. It can also be concluded that sub-sampled images are unable to capture the distribution of real LR images.

Super-resolution is widely used to resolve LR images to reveal more details. Significant developments have been made in this field and methods based on encoder-decoder architectures and GANs [10] have been proposed. We employ two recent deep learning based methods, SRGAN and ESRGAN [27] to resolve a given LR image. Figure 1(e) shows the result when super-resolved image is passed through HR-LD. It can be hypothesized that possibly, the super-resolved images do not lie in the same space of images using which HR-LD was trained and this result can be generalized to other state of the art methods. Super resolution networks are trained using synthetic LR images obtained by down-sampling the image after applying Gaussian smoothing. In some cases, the training data for super-resolution networks consists of paired LR and HR images. Neither of the mentioned scenarios is effective in real world situations.

Contribution: Different from these approaches, S^2LD is based on the concept of 'learning from synthetic data'. This work aims to show that landmark localization in LR can not only be achieved, but it also improves the performance over the current practice. To this end, we first train a deep network which generates LR from HR images and tries to model the distribution of target LR images. *Since, there is no publicly available dataset, containing LR images along with landmark annotations*, we take a semi-supervised approach and train an adversarial landmark localization network on the generated LR. We design a Heatmap confidence discriminator (with three sets of inputs) in a way that to be fooled, the heatmap generator learns the structure of the face in the target unlabeled LR dataset. We perform extensive set of experiments explaining all the design choices. In addition, we also propose a new state of the art landmark detector (HR-LD) for HR images.

2. Related Work

Being one of the most important pre-processing steps in face analysis tasks, facial landmark detection has been a topic of immense interest among computer vision researchers. MTCNN [34] and KEPLER [14] proposed methods based on direct regression. The CNNs in MTCNN and KEPLER act as non-linear regressors and learn to directly predict the landmarks. Both works are designed to predict other attributes along with key-points such as 2D pose, visibility of key-points, gender and many others. Hyperface [22] has shown that learning multiple tasks using a single network does in fact, improves the performance of individual tasks. Recently, architectures based on Encoder-Decoder paradigm have become popular and are used extensively for tasks that require per-pixel labeling such as semantic segmentation [21] and key-point detection [15, 1, 32]. Despite making significant progress in this field, predicting landmarks on LR faces still remains a relatively unexplored topic. All of the works mentioned above are trained on high quality images and their performance degrades on LR images.

One of the closely related works, is Super-FAN [5] which predicts landmarks on LR images using a super-resolution approach. However, as shown in experiments section, face verification performance degrades even on super-resolved images. This necessitates that super-resolution, face-alignment and face verification be learned in a single model, trained end to end, making it not only slow in inference stage but also limited by the GPU memory constraints. The proposed work is different from Super-FAN in many aspects as it needs labeled data only in HR and learns to predict landmarks in LR images in an unsupervised way. Due to adversarial training, S^2LD not only acts as a facial parts detector but also learns the inherent structure of the facial parts. The proposed method makes the pre-processing task faster and independent of face verification training. During inference, only the heatmap generator network is used which is based on the fully convolutional architecture of U-Net and works at the spatial resolution of 32×32 making the alignment process real time.

3. Proposed Method

S^2LD predicts landmarks directly on a LR image of spatial size less than 32×32 pixels. We show that predicting landmarks directly in LR is more effective than the current practices of rescaling or super-resolution. The entire pipeline can be divided into two stages: (a) Generation of LR images in an unpaired manner (b) Generating heatmaps for target LR images in a semi-supervised fashion. An

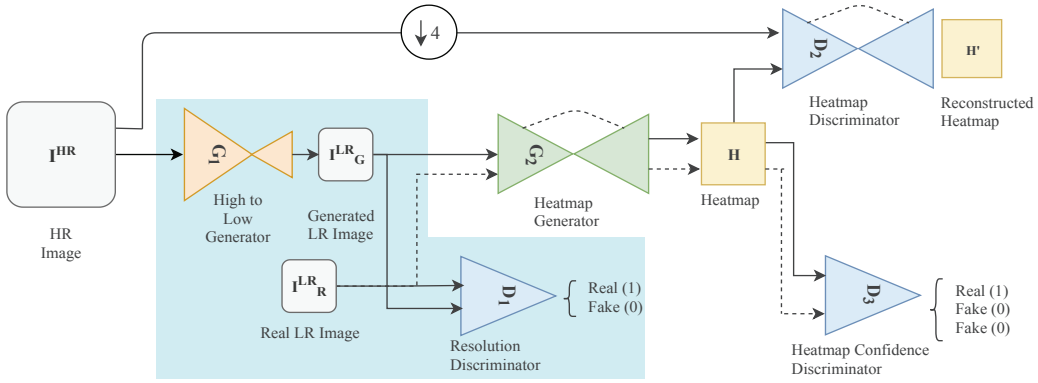

Figure 2: Overview of the proposed approach. HR input is passed through High-to-Low generator G_1 (shown in cyan colored block). The discriminator D_1 learns to distinguish generated LR images vs. real LR images in an unpaired fashion. This generated image is fed to heatmap generator G_2. Heatmap discriminator D_2 distinguishes generated heatmap vs. groundtruth heatmaps. The pair G_2, D_2 is inspired by BEGAN [3]. In addition to generated and groundtruth heatmaps, the discriminator D_3 also receives predicted heatmaps for real LR images. This enables the generator G_2 to generate realistic heatmaps for unannotated LR images.

overview of the proposed approach is shown in Figure 2. Being a semi-supervised method, it is important to first describe the datasets chosen for the experiments.

High Resolution Dataset: We construct the HR dataset by combining the $20,000$ training images from AFLW and the entire 300W dataset. We divide the Widerface dataset [30] into two groups based on their spatial size. The first group consists of images with spatial size between 20×20 and 40×40, whereas the second group consists of images with more than 100×100 pixels. We combine the second group in HR training set, resulting in a total of $35,543$ HR faces. The remaining $4,386$ images from AFLW are used as validation images for the ablative study and test set for the landmark localization task.

Low Resolution Datasets:

- The first group from Widerface dataset consists of $47,046$ faces is used as real LR images for ablative study.
- For face verification experiments, we use recently published TinyFace dataset [7] as the target LR dataset.
- Due to the unavailability of an LR annotated dataset, we create a real LR landmark detection dataset which we call Annotated LR Faces (ALRF https://sites.google.com/view/amitumd) by manually annotating 700 LR images of TinyFace dataset. The details of ALRF creation are discussed in the supplementary.

3.1. High to Low Generator and Discriminator

High to low generator G_1, shown in Figure 3, is designed following the Encoder-Decoder architecture, where both encoder and decoder consists of multiple residual blocks. The input to the first convolution layer is the HR image concatenated with the noise vector which has been projected using a fully connected layer and reshaped to match the input size. Similar architectures have also been used in [6, 16]. The encoder in the generator consists of eight residual blocks each followed by a convolution layer to increase the dimensionality. Max-pooling is used after every 2 residual block to decrease the spatial resolution to 4×4, for HR image of 128×128 pixels. The decoder is composed of six residual units followed by up-sampling and convolution layers. Finally, one convolution layer is added in order to output a three channel image. BatchNorm is used after every convolution layer.

The discriminator D_1, shown in Figure 3 is also constructed in a similar way, except that due to low spatial resolution of the input image, max-pooling is only used in the last three layers. In Figure 2, we use I^{HR} for HR input images of size 128×128, I^{LR}_G for generated LR images of size 32×32 and I^{LR}_R for target LR images of the same size. Spectral Normalization [20] is also used in the convolutional layers of D_1 to satisfy the Lipschitz constraint $\sigma(\boldsymbol{W}) = 1$, where the weights $\boldsymbol{W}$ presented in Equation 1:

$$\boldsymbol{W}_{SN}(\boldsymbol{W}) = \frac{\boldsymbol{W}}{\sigma(\boldsymbol{W})} \tag{1}$$

We train G_1 using a weighted combination of GAN loss; L_2 pixel loss to encourage convergence in initial training iterations and perceptual loss [12] back-propagated from a pre-trained VGG network. The final loss is summarized in Equation 2.

$$l_{G_1} = \alpha l^G_{GAN} + \beta l_{pixel} + \gamma l_{perceptual} \tag{2}$$

where α, β and γ are hyperpameters which are empirically set. Following recent developments in GANs we ex-

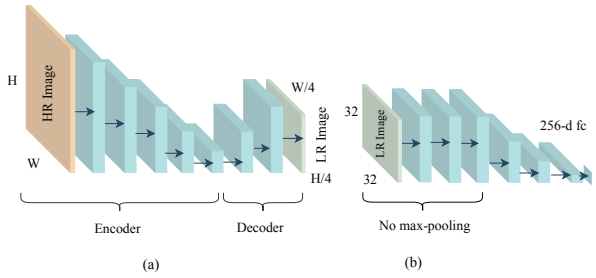

Figure 3: (a) High to low generator G_1. Each $\rightarrow$ represents two residual blocks followed by a convolution layer. (b) Discriminator used in D_1 and D_2. Each $\rightarrow$ represents one residual block followed by a convolution layer.

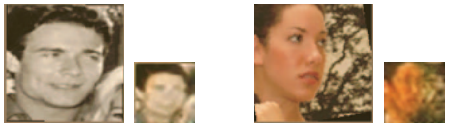

Figure 4: Sample outputs of High to Low generation of AFLW dataset. For more results please refer to the supplementary material.

perimented with different loss functions. However, we settled on the hinge loss. In Equation 2, l^G_{GAN} is computed as:

$$l^G_{GAN} = E_{\hat{x}\in P_g}[\min(0, -1 + D_1(\hat{x}))] \tag{3}$$

where P_g is the distribution of generated images I^{LR}_G. Also L_2 pixel loss, l_{pixel}, is derived from the following expression:

$$l_{pixel} = \frac{1}{H\times W}\sum_{i=1}^{W}\sum_{i=1}^{H}(F(I^{HR}) - I^{LR}_G)^2 \tag{4}$$

where W and H represent the generated image width and height respectively; also the operation F is implemented as a sub-sampling operation obtained by passing I^{HR} through four average pooling layers. This loss is used to minimize the distance between the generated and sub-sampled images which ensures that the content is not lost during the generation process. To train discriminator D_1 we use hinge loss with gradient penalty and Spectral Normalization for faster training. The discriminator D_1 loss can be defined as:

$$l_{D_1} = l^D_{GAN} + GP \tag{5}$$

where

$$l^D_{GAN} = E_{x\in P_r}[\min(0, -1 + D_1(x))] + E_{\hat{x}\in P_g}[\min(0, -1 - D_1(\hat{x}))] \tag{6}$$

and P_r is the distribution of real LR images I^{LR}_R from Widerface dataset. GP in Equation 5 represents the gradient penalty term. Figure 4 shows some sample LR images generated from the network G_1.

3.2. Semi-Supervised Landmark Localization

3.2.1 Heatmap Generator G_2

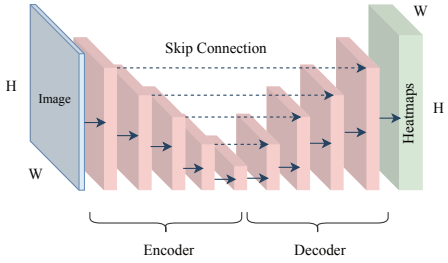

Figure 5: Architecture of the heatmap generator G_2. Architecture of this network is based on U-Net. Each $\rightarrow$ represents two residual blocks. $\dashrightarrow$ represents skip connections between the encoder and decoder.

The key-point heatmap generator, G_2 in Figure 5 produces heatmaps corresponding to N (in our case 19 or 68) key-points in a given image. As mentioned earlier, the objective of this paper is to show that landmark prediction directly on LR images is feasible even in the absence of labeled LR data. To this end, we choose a simple network based on the U-Net architecture as the heatmap generator. The network consists of 16 residual blocks where both encoder and decoder have eight residual blocks. In the last layer, G_2 outputs (N+1) feature maps corresponding to N key-points and 1 background channel. After experimentation, this design for landmark detection has proven to be very effective and results in state of the art resutls for HR landmark predictions. Further architectural details are presented in the supplementary materials.

3.2.2 Heatmap Discriminator D_2

The heatmap discriminator D_2 follows the same architecture as the heatmap generator G_2 with different number of input channels, *i.e.*, input to the discriminator is a set of heatmaps concatenated with their respective color images. D_2 receives two sets of inputs: *generated LR image with down-sampled groundtruth heatmaps* and *generated LR images with predicted heatmaps*. This discriminator predicts another set of heatmaps and learns whether the key-points described by the input heatmaps are correct and correspond to the input face image. The quality of the output heatmaps is determined by their similarity to the input heatmaps, following the notion of an autoencoder. The loss is computed as the error between the input and the reconstructed heatmaps.

3.2.3 Heatmap Confidence Discriminator D_3

The architecture of D_3 is identical to D_1 except for the number of input channels. This discriminator receives three inputs: *generated LR image with corresponding groundtruth heatmaps*, *generated LR image with predicted heatmaps* and *target LR image with predicted heatmaps*. D_3 learns to distinguish between the groundtruth and predicted heatmaps. To fool this discriminator, G_2 should learn to: (a) generate heatmaps for generated LR images similar to their respective groundtruth, (b) generate heatmaps for unlabeled target LR images with similar statistical properties to the groundtruth heatmap, *i.e.*, G_2 should understand the inherent structure of the face in LR images and generate accurate and realistic heatmaps.

3.3. Semi-supervised Learning

The learning process of this setup is inspired by the seminal works BEGAN [3] and [35] called Energy-based GANs. It is worth recalling that HR images have annotations associated with them and we assume key-point locations in a generated LR image stay relatively the same as its downsampled version. Therefore, while training G_2, the downsampled annotations are considered to be groundtruth for the generated LR images.

The discriminator D_2, when the input consists of groundtruth heatmaps, is trained to recognize it and reconstruct a similar one to minimize the error between the groundtruth and reconstructed heatmaps. On the other hand, if the input consists of generated heatmaps, the discriminator is trained to reconstruct different heatmaps to drive the error as large as possible. The losses are expressed as

$$l_D^{real} = \sum_{i=1}^{N+1} (H_i - D_2(H_i, I_G^{LR}))^2 \tag{7}$$

$$l_D^{fake} = \sum_{i=1}^{N+1} (\hat{H}_i - D_2(\hat{H}_i, I_G^{LR}))^2 \tag{8}$$

$$l_D^{kp} = l_D^{real} - k_t l_D^{fake} \tag{9}$$

where H_i and $\hat{H}_i$ represent the i^{th} key-point groundtruth and generated heatmap of the generated LR image I_G^{LR}. Inspired by BEGAN, we use a variable k_t to control the balance between heatmap generator and discriminator. The variable is updated every t iterations. The adaptive term k_t is defined by:

$$k_{t+1} = k_t + \lambda_k(\gamma l_D^{real} - l_D^{fake}) \tag{10}$$

where k_t is bounded between 0 and 1, and λ_k is a hyper-parameter. As in Equation 9, k_t controls the emphasis on l_D^{fake}. When the generator is able to fool the discriminator, l_D^{fake} becomes smaller than γl_D^{real}. As a result of this k_t increases, making the term l_D^{fake} dominant. The amount of acceleration to train on l_D^{fake} is adjusted proportional to

Figure 6: Sample key-point detections on TinyFace images.

$\gamma l_D^{real} - l_D^{fake}$, *i.e* the distance the discriminator falls behind the generator. Similarly, when the discriminator gets better than the generator, k_t decreases, to slow down the training on l_D^{fake} making the generator and the discriminator train together.

The discriminator D_3 is trained using the loss function from Least squares GAN [19] as shown in Equation 11. This loss function was chosen to be consistent with the losses computed by D_2.

$$l_D^{conf} = \mathbb{E}_{x \in \mathbb{P}_r}[(D_3(x) - 1)^2] + \mathbb{E}_{\hat{x} \in \mathbb{P}_g}[D_3(\hat{x})^2] + \mathbb{E}_{\hat{y} \in \mathbb{P}_g}[D_3(\hat{y})^2] \tag{11}$$

It is noteworthy to mention in this case $\mathbb{P}_r$ represents the groundtruth heatmaps distribution on generated LR images, while $\mathbb{P}_g$ represents the distribution on generated heatmaps of generated LR images and real LR images.

The generator G_2 is trained using a weighted combination of losses from the discriminators D_2 and D_3 and l_{MSE} heatmap loss. The loss functions for the generator G_2 are described in the following equations:

$$l_G^{MSE} = \sum_{i=1}^{N+1} (H_i - G_2(I_G^{LR}))^2 \tag{12}$$

$$l_G^{kp} = \sum_{i=1}^{N+1} (\hat{H}_i - D_2(\hat{H}_i, I_g^{LR}))^2 \tag{13}$$

$$l_G^{conf} = \mathbb{E}_{x \in \mathbb{P}_g}[(D_3(x) - 1)^2] \tag{14}$$

$$l_G = a l_G^{MSE} + b l_G^{kp} + c l_G^{conf} \tag{15}$$

where a, b and c are hyper parameters set empirically obeying $a l_G^{MSE} > b l_G^{kp} > c l_G^{conf}$. We put more emphasis on l_G^{MSE} to encourage convergence of the model in initial iterations. Some target LR images with key-points predicted from the G_2 are shown in Figure 6.

4. Experiments and Results

4.1. Ablation Experiments

We qualitatively demonstrated in Figure 1 that networks trained on HR images perform poorly on LR. Moreover, as there are no LR image datasets with landmark annotations available, we propose semi-supervised learning as an alternative. Given the above mentioned networks and loss functions; it is important to understand the implication of each component. This section examines each of the design choices quantitatively. To this end, we first train the high

to LR network, G_1, on WiderFace dataset and then generate LR version of AFLW testset. In the absence of real LR images with annotated landmarks, this is done to create a substitute for LR dataset with annotations on which localization performance can be evaluated. Data augmentation techniques such as random scaling $(0.9, 1.1)$, random rotation $(-30°, 30°)$ and random translation up to 20 pixels are used.

Evaluation Metric: Following prior works, we use the Normalized Root Mean Square Error (NRMSE) to measure key-point localization performance.

Training Details: All the networks are trained in Pytorch using the Adam optimizer with an initial learning rate of 2e-4 and β_1, β_2 values of $0.5, 0.9$. We train the networks with a batch size of 32 for 200 epochs, while dropping the learning rates by 0.5 after 80 and 160 epochs. Performance is evaluated on generated LR AFLW test images and our manually annotated ALRF dataset.

Setting S1: *Train networks on sub-sampled images?* We train network G_2 in a supervised manner with the sub-sampled AFLW training images using the loss function in Equation 12.

Setting S2: *Train networks on generated LR images?* In this experiment, we train the network G_2 using the generated LR images, in a supervised manner using the loss function from Equation 12.

Observation: From the results summarized in Table 1, it is evident that there is a significant reduction in the localization error when G_2 is trained on the generated LR images validating our hypothesis that sub-sampled images on which many super-resolution networks are trained may not represent real LR images. Hence, we need to train the networks on real LR images.

Setting S3: *Does adversarial training help?* This question is relevant to understanding the importance of training G_2 in an adversarial way. In this experiment, we train D_2 and G_2 using the losses in Equations 7, 8, 12, 13.

Setting S4: *Does G_2 trained in adversarial manner scale to real LR images?* In this experiment, we wish to examine if training networks G_2, D_2 and D_3 jointly, improves the performance on real LR images and whether D_3 can help G_2 to generate heatmaps that can characterize a face image in LR.

Observation: From Table 1 we observe that the network trained with setting S4 performs comparable to setting S3 for the AFLW dataset which is expected since G_2, D_2 and D_3 are only trained on the AFLW training dataset and G_2 can learn the inherent LR face characteristics in AFLW using only D_2. However; when we compare settings for the ALRF dataset there is a significant boost from S3 to S4 which substantiate the knowledge that D_3 provides to G_2 once there is a domain shift between LR generated images of AFLW and the target LR domain. Since D_3 sees data from target LR images domain, it enforces G_2 to learn the structure of faces corresponding to target LR images and generates accurate heatmaps for face alignment. We highly encourage the readers to refer to the supplementary material for more detailed explanations of *S settings*.

Dataset	NRMSE			
	Settings			
	S1	S2	S3	S4
AFLW Testset	11.33	4.23	4.12	**4.12**
ALRF	0.71	0.70	0.65	**0.37**

Table 1: Ablation experiment results under settings S1-S4 on synthesized LR images.

Method	NRMSE (all)	NRMSE (479)	Time (s)
MTCNN	-	0.9736	0.388
HRNet	0.4055	0.3107	0.076
SAN	0.3901	0.3141	0.0178
S^2LD	**0.257**	**0.1803**	**0.0105**

Table 2: Numerical comparison with recent state of the arts key-point Detection methods on the ALRF dataset. NRMSE(479) corresponds to images that MTCNN detected.

4.2. Comparison with State of the Art Methods

Here using the ALRF dataset, we perform a numerical comparison with respect to recent state of the art methods namely MTCNN, HRNet [25] and SAN [9]. Table 2 summarizes the result of this comparison. Note that here we use TinyFace dataset as real LR images. We also calculate inference time per face image in a single gtx1080. Note that MTCNN which has detection and alignment in a single system, was able to detect only 479 faces out of 700 test images, therefore we add another column to have a fair comparison.

4.3. Face Verification experiments

In the previous section we studied the generalization of the S^2LD to landmark detection in LR face images. Therefore we choose to evaluate models from setting S3 and setting S4 in previous section, by comparing the statistics obtained by applying the two models to align the face images for facial verification task. The reason stems from the fact that performance of a face verification system is directly impacted by the accuracy of face alignment.

We use the TinyFace dataset [7] in the following experiments. It is one of the very few datasets aimed towards understanding LR face verification and consists of $5,139$ labeled facial identities with an average of three face images per identity, giving a total of $15,975$ LR face images. $5,139$ known identities is divided into two splits: $2,570$ for training and the remaining $2,569$ for test.

Setting	L1	L2	L3	L4	L5
top-1	31.17	35.11	39.03	39.87	43.82

(a)

Setting	top-1	top-5	top-10	top-20	mAP
Baseline	34.71	44.82	49.01	53.70	0.32
I1	34.01	41.98	45.36	49.22	0.29
I2	45.04	56.30	60.11	63.71	0.43
I3	**51.10**	**61.05**	**64.38**	**67.89**	**0.47**

(b)

Table 3: Verification performance on TinyFace dataset under different settings (a) LightCNN trained from scratch (b) Using Inception-ResNet pre-trained on MsCeleb-1M

Evaluation Protocol: To compare model performance, we adopt the closed-set face identification (1:N matching) protocol. Specifically, the task is to match a given probe face against a gallery set of face images with true match from the gallery at top-1 of the ranking list. For each test class, half of the face images are randomly assigned to the probe set, and the remaining to the gallery set. For face verification, we report statistics on Top-k (k=1,5,10,20) and mean average precision (mAP).

Experiments with face verification network trained from scratch: Since the number of images in TinyFace dataset is much smaller compared to larger datasets such as CASIA or MsCeleb-1M [11], we observed that training a very deep model like Inception-ResNet [26], quickly leads to over-fitting. Therefore, we adopt a CNN with fewer parameters, specifically, LightCNN [29]. Since inputs to the network are images of size 32×32, we disable first two max-pooling layers to keep the spatial resolution. After detecting the landmarks, training and testing images are aligned to the canonical coordinates using affine transformation. We train LightCNN with training split of TinyFace dataset under the following settings:

Setting L1: *Train networks using generated LR images?* Here, G_2 from setting S2 of previous section is used to extract key-points and align face images to be used for LightCNN training.

Setting L2: *Does adversarial training help?* We use the G_2 trained from setting S3 to align the faces in training and testing sets.

Setting L3: *Does G_2 trained in adversarial manner scale to real LR images?* Here G_2 trained from setting S4 with the TinyFace train set as real LR images, is used for key-point detection and image alignment in LightCNN.

Setting L4: *End-to-end training?* Here, we also train the High to Low networks G_1 and D_1, using the TinyFace train dataset as real LR images. With the obtained S^2LD model, landmarks are extracted and images are aligned for training LightCNN.

Setting L5: *End-to-end training with pre-trained weights?* This setting is similar to the setting L4 above, except instead of training a LightCNN model from scratch we initialize the weights from a pre-trained model, trained with CASIA-Webface dataset.

Observation: Table 3a summarizes the results of the experiments done under settings L1 to L5. We observe that there is a significant gap in rank-1 performance between setting L2 and L3. This indicates that with semi-supervised learning G_2 generalizes well to real LR data, and hence validates our hypothesis of training G_2, D_2 and D_3 together. Unsurprisingly, insignificant difference is seen between settings L3 and L4. More details about *L settings* is provided in supplementary materials

Experiments with pre-trained network: Next, to further understand the implications of joint semi-supervised learning, we design another set of experiments. In these experiments, we use a pre-trained Inception-ResNet model, trained on MsCeleb-1M using ArcFace [8] and Focal Loss [17]. This model expects an input of size 112×112 pixels, hence the images are re-sized after alignment in LR. Using this network, we perform the following experiments:

Baseline: We follow the usual practice of re-scaling images to a fixed size irrespective of resolution. We trained our own HR landmark detector (HR-LD) on $20,000$ AFLW images. TinyFace gallery and probe images are re-sized to 128×128 and are fed to HR-LD and aligned similar to ArcFace. Baseline performance was obtained by computing cosine similarity between gallery and probe features extracted from the network

Setting I1: *Does adversarial training help?* The model trained for S3 is used to align the images directly in LR. Features for gallery and probe images are extracted after rescaling and cosine distance is measured.

Setting I2: *Does G_2 trained in adversarial manner scale to real LR images?* For this experiment, the model from setting L3 is used for landmark detection in LR and face verification is done on aligned and re-sized images.

Setting I3: *End-to-end training?* In this case, we align images using the model from setting L4, re-size images and measure face verification metrics.

Observation: As expected, in Table 3b we observe training the landmark detector in a semi-supervised manner and aligning the images directly in LR, improves performance of any face verification system trained on HR images.

4.4. Additional Experiments

Setting A1: *Does super-resolution help?* The aim of this experiment is to understand if super-resolution can be used to enhance the image quality for landmark detection. We use SRGAN to super-resolve the images before using 2D-FAN face alignment method.

Setting A2: *Does super-resolution help?* In this case, we use ESRGAN to super-resolve the images before using

Setting	top-1	top-5	top-10	top-20	mAP
A1	11.75	14.58	24.57	30.47	0.10
A2	26.21	34.76	39.03	43.99	0.24

Table 4: Face verification performance using super-resolution prior to face-alignment

HR-LD to align.

Observation: It is evident from Table 4 that face verification performance obtained after aligning super-resolved images is not at par even with the baseline. It can be hypothesized that possibly super-resolved images do not represent HR images using which 2D-FAN or HR-LD are trained.

HR Landmark Detector (HR-LD) Here, we train G_2 on HR images of size 128×128 of AFLW and 300W using MSE loss in Equation 12. We evaluate the performance of this network on AFLW-Full and 300W test sets, shown in Table 5. We would like to make a note that LAB [28] and SAN either uses extra data or extra annotations or larger spatial resolution to train deep networks. A few sample outputs of HR-LD are shown in Figure 7. For details on HR-LD architecture please refer to supplementary materials.

	300W			AFLW
Method	Common	Challenge	Full	Full
LBF	4.95	11.98	6.32	4.25
CFSS	4.73	9.98	5.76	3.92
TCDCN	4.80	8.60	5.54	-
MDM	4.83	10.14	5.88	-
PCD-CNN	3.67	7.62	4.44	2.36
SAN*	3.34	6.60	3.98	1.91
LAB*	**2.57**	**4.72**	**2.99**	1.85
HR-LD	3.60	7.301	4.325	**1.753**

Table 5: Comparison of the HR-LD with state of the art methods on AFLW and 300-W testsets. NMSEs on 300W dataset are taken from the Table 3 of [18]. * uses extra annotation or data.

Figure 7: Sample outputs of HR-LD. First and second row show samples from AFLW and 300W test sets respectively.

5. Conclusion

In this paper, we first present an analysis of landmark detection methods when applied to LR images, and the implications on face verification. We also discuss the proposed method for predicting landmarks directly on LR images. We show that the proposed method improves key-point detection as well as face recognition performance over commonly used practices of rescaling and super-resolution. As a by-product, we also developed a simple but state of the art landmark detector. Although, LR is chosen as the source of degradation, the proposed method can trivially be extended to capture other degradation in the imaging process, such as motion blur or atmospheric turbulence. In addition, the proposed method can be applied to detect human pose in LR to improve skeletal action recognition. In the era of deep learning, LR landmark detection and face recognition is a relatively understudied topic; however, we believe this work will open new avenues in this direction.

6. Acknowledgment

This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA RD Contract No. 2019-022600002. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

References

[1] A recurrent autoencoder-decoder for sequential face alignment. http://arxiv.org/abs/1608.05477. Accessed: 2016-08-16. 2

[2] Ankan Bansal, Carlos Castillo, Rajeev Ranjan, and Rama Chellappa. The Do's and Don'ts for CNN-based face verification. *arXiv preprint arXiv:1705.07426*, 2017. 1

[3] David Berthelot, Tom Schumm, and Luke Metz. BEGAN: Boundary equilibrium generative adversarial networks. *CoRR*, abs/1703.10717, 2017. 3, 5

[4] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2D & 3D face alignment problem?(and a dataset of 230,000 3D facial landmarks). In *International Conference on Computer Vision*, volume 1, page 8, 2017. 2

[5] 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. *CoRR*, abs/1712.02765, 2017. 2

[6] Adrian Bulat, Jing Yang, and Georgios Tzimiropoulos. To learn image super-resolution, use a GAN to learn how to do image degradation first. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 185–200, 2018. 3

[7] Zhiyi Cheng, Xiatian Zhu, and Shaogang Gong. Low-resolution face recognition. *CoRR*, abs/1811.08965, 2018. 3, 6

[8] Jiankang Deng, Jia Guo, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *CoRR*, abs/1801.07698, 2018. 7

[9] Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Style aggregated network for facial landmark detection. In *CVPR*, pages 379–388, 2018. 6

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014. 2

[11] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. MS-celeb-1M: A dataset and benchmark for large-scale face recognition. *CoRR*, abs/1607.08221, 2016. 7

[12] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, 2016. 3

[13] Martin Koestinger, Paul Wohlhart, Peter M. Roth, and Horst Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies*, 2011. 2

[14] A. Kumar, A. Alavi, and R. Chellappa. KEPLER: Keypoint and pose estimation of unconstrained faces by learning efficient h-cnn regressors. In *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, pages 258–265, May 2017. 2

[15] Amit Kumar and Rama Chellappa. Disentangling 3d pose in A dendritic CNN for unconstrained 2d face alignment. *CoRR*, abs/1802.06713, 2018. 2

[16] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew P. Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. *CoRR*, abs/1609.04802, 2016. 3

[17] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *CoRR*, abs/1708.02002, 2017. 7

[18] Jiangjing Lv, Xiaohu Shao, Junliang Xing, Cheng Cheng, and Xi Zhou. A deep regression architecture with two-stage re-initialization for high performance facial landmark detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 8

[19] Xudong Mao, Qing Li, Haoran Xie, Raymond Y. K. Lau, and Zhen Wang. Multi-class generative adversarial networks with the L2 loss function. *CoRR*, abs/1611.04076, 2016. 5

[20] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *CoRR*, abs/1802.05957, 2018. 3

[21] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. *arXiv preprint arXiv:1505.04366*, 2015. 2

[22] Rajeev Ranjan, Vishal M. Patel, and Rama Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *CoRR*, abs/1603.01249, 2016. 2

[23] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015. 1

[24] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *2013 IEEE International Conference on Computer Vision Workshops*, pages 397–403, Dec 2013. 1

[25] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 6

[26] Christian Szegedy, Sergey Ioffe, and Vincent Vanhoucke. Inception-v4, inception-resnet and the impact of residual connections on learning. *CoRR*, abs/1602.07261, 2016. 7

[27] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Chen Change Loy, Yu Qiao, and Xiaoou Tang. ESRGAN: enhanced super-resolution generative adversarial networks. *CoRR*, abs/1809.00219, 2018. 2

[28] Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou. Look at boundary: A boundary-aware face alignment algorithm. In *CVPR*, 2018. 8

[29] Xiang Wu, Ran He, and Zhenan Sun. A lightened CNN for deep face representation. *CoRR*, abs/1511.02683, 2015. 7

[30] 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 (CVPR)*, 2016. 3

[31] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z. Li. Learning face representation from scratch. *CoRR*, abs/1411.7923, 2014. 1

[32] Jie Zhang, Shiguang Shan, Meina Kan, and Xilin Chen. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *ECCV*, volume 8690 of *Lecture Notes in Computer Science*, pages 1–16. Springer International Publishing, 2014. 2

[33] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, Oct 2016. 2

[34] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *ECCV*, pages 94–108, 2014. 2

[35] Junbo Jake Zhao, Michaël Mathieu, and Yann LeCun. Energy-based generative adversarial network. *CoRR*, abs/1609.03126, 2016. 5