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

Amit Kumar, Rama Chellappa
Department of Electrical and Computer Engineering, CFAR and UMIACS
University of Maryland-College Park,USA
akumar14@umiacs.umd.edu, rama@umiacs.umd.edu

1. Creation of ALRF

This dataset was collected in order to evaluate the performance of landmark localization on real low resolution face dataset. We randomly selected 700 identities from the TinyFace dataset, out of which one LR image (of spatial size less than 32×32 pixels and more than 15×15 pixels) per identity was again randomly selected, resulting in a total of 700 LR images. Next, three individuals were asked to manually annotated all the images with 5 landmarks (two eye centers, nose tip and mouth corners) in MTCNN style, where invisible points were annotated with -1. The mean of the points obtained from the three users was taken to be the groundtruth. Note that ALRF is excluded from TinyFace for the face verification experiments. Downloadable link for ALRF https://sites.google.com/view/amitumd

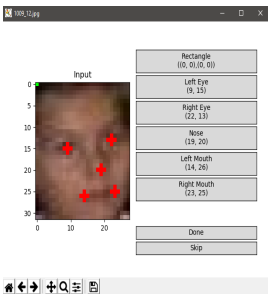


Figure 1: Snippet of the annotation tool used to label landmarks in low resolution images.

2. Architecture of HR-LD

High resolution landmark detector and heatmap generator use similar architectures based on UNet with difference being in the input sizes. The encoder of the UNet consists of 4 layers, each of which consists of two residual blocks. The feature maps are down-sampled between every layer. Two dilated convolutions are applied at the end of encoding stage. The decoder is structured in exactly similar fashion, where it is divided into 4 layers. Each layer again consists of two residual blocks. Between each layer, the feature maps are up-sampled and the skip connections from the encoder are added.

The input to HR-LD is images of size $128 \times 128 \times 3$, whereas the output heatmaps are of dimensions $128 \times 128 \times 20$; 19 channels for each key-point and 1 channel for background.

It is worth mentioning that many previously existing algorithms like SBR and CPM [2, 5] cannot be trained on generated images. These methods expect an input of larger spatial size while regressing key-points at 56×56. The generated LR images from G_1 are of spatial size 32×32, which will output a feature map of size 8×8, rendering training of SBR or CPM on generated images infeasible.

Distinction with Adversarial PoseNet [1]: The proposed heatmap confidence discriminator is evidently distinguishable from the one in Adversarial PoseNet. D_3 in the proposed work expects three inputs corresponding to predicted and groundtruth heatmaps of generated images and predicted heatmaps of target images. In contrast to this, heatmap confidence discriminator in Adversarial PoseNet takes two inputs and makes decisions based on the visibility confidence. We choose to use the same name as, in essence they are responsible for ensuring that the heatmap generator predicts feasible heatmaps.

3. Explanation of different settings in experiments

In this section, we elaborate different settings in which experiments were performed.

Setting S: Experiments under this setting were performed to understand the significance of training G_2, D_2 and D_3. The weights of G_1 and D_1 were frozen after training, considering AFLW+300W as high resolution images and small Widerface images as real low resolution images. After this we systematically proceed towards training G_2

which can be trained either in supervised way or adversarial manner. Supervised training can also be performed either with sub-sampled images (Setting S1) or generated images (Setting S2). G_2 and D_2 are trained following adversarial learning in setting S3. We observe that D_3 is a three way discriminator and expects target LR images with predicted heatmaps as the third input. This is quite unconventional when training generative adversarial networks. Target low resolution images can be taken from either Widerface or TinyFace, giving rise to setting S4. Table 1 in main paper clearly demonstrates the effect of three way discriminator in generalizing to TinyFace dataset, even when G_1 and D_1 were trained with Widerface as target dataset.

Setting L: In these experiments key-points are obtained on target low resolution images from TinyFace, which are then used to align images to canonical coordinates. Subsequently we train LightCNN from scratch to understand the impact of each training strategy on face verification. G_2 trained under different settings in 'Setting S' are used to extract key-points. We do not use setting S1 as it is quite evident that sub-sampled images are not representative of real low resolution images. G_2 from setting S2 (G2 trained in supervised manner with generated LR images) gives rise to setting L1. G_2 from setting S3 (G_2 and D_2 trained in adversarial BEGAN manner) gives the verification numbers corresponding to setting L2. In setting L3 we train the verification network after aligning the faces from key-points obtained from network in setting S4, second row (where the third input to D_3 was real LR images from TinyFace with predicted heatmaps) in Table 1 in the main paper.

Next question we try to answer is the effect of training G_1 and D_1 also, by considering TinyFace as real LR dataset as opposed to Widerface in networks in 'setting S'. For setting L4 G_1 and D_1 are trained with AFLW+300W as HR images and TinyFace as real LR images. Later G_2, D_2 and D_3 are also trained and key-points are extracted, after which we train LightCNN from scratch for verification. In setting L5, LightCNN is initialized with pre-trained weights.

Setting I: In these experiments, neither key-point networks nor image generators are trained. The key-points obtained from G_2 trained under different settings in 'setting S' are used to extract key-points and align images. The aligned images are fed through a pre-trained Inception ResNet from ArcFace. No training of Inception ResNet is done, as we observed training Inception ResNet on a small dataset leads to prompt overfitting.

Setting A: These represent additional experiments performed to gain insight as to the outcome of face verification when super-resolved images from famous deep learning based methods are used for key-point extraction.

	Crystal Loss	**Semi-Supervised**
Rank 1	23.65	28.88
Rank 2	26.03	32.42
Rank 3	27.58	33.57
Rank 4	28.14	34.46
Rank 5	28.64	35.05
Rank 7	29.54	36.61
Rank 10	30.42	37.46
Rank 20	32.58	39.95
Rank 30	34.38	42.05
Rank 40	35.79	43.34
Rank 50	36.69	44.61

Table 1: Retrieval rates at different ranks(Higher is better)

FPIR/Method	**Crystal Loss**	**Semi-Supervised**
1e2	0.9450	0.8959
1e3	0.9081	0.8767
1e4	0.8808	0.8485
1e5	0.8114	0.7720

Table 2: False negative rates at different false positive rates. (Lower is better)

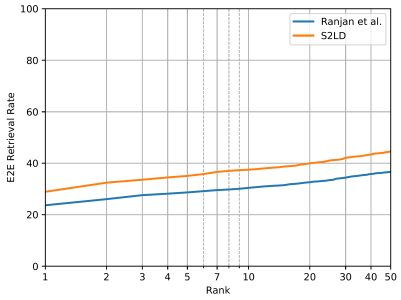
4. Evaluation on the IJB-S dataset

Along with the method to predict landmarks in low resolution images, the paper presents a rather counter-intuitive result that performing landmark detection directly in low resolution leads to higher face recognition performance. To understand this further we performed experiments on recently released IJB-S dataset [3]. IJB-S dataset is one of the most challenging dataset available, and consists of several videos collected with surveillance cameras. The subjects in this dataset are extremely challenging to verify because of the distance from the camera and low resolution. We randomly selected 10 videos from the dataset which contained at least 5 subjects from the two galleries the dataset provides. We used surveillance to booking protocol for the purpose of this experiment. Only 10 videos were chosen attributing to the fact that IJB-S is an extremely large dataset and experimenting on the entire dataset takes more than a month on a single GPU machine. Tables 1 and 2 shows retrieval rates at different ranks and false negative rates vs false positives. We compare with [4].

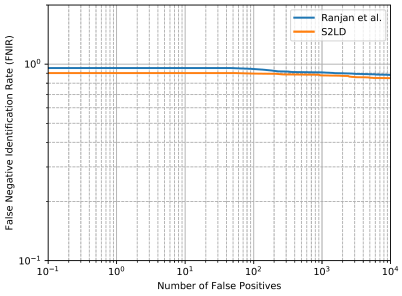
References

[1] Yu Chen, Chunhua Shen, Xiu-Shen Wei, Lingqiao Liu, and Jian Yang. Adversarial posenet: A structure-aware convolutional network for human pose estimation. *CoRR*, abs/1705.00389, 2017. 1

[2] Xuanyi Dong, Shoou-I Yu, Xinshuo Weng, Shih-En Wei, Yi Yang, and Yaser Sheikh. Supervision-by-registration: An un-



(a)



(b)

Figure 2: (a) Retrieval rates at different ranks. (b) False negatives at different false positive rates.

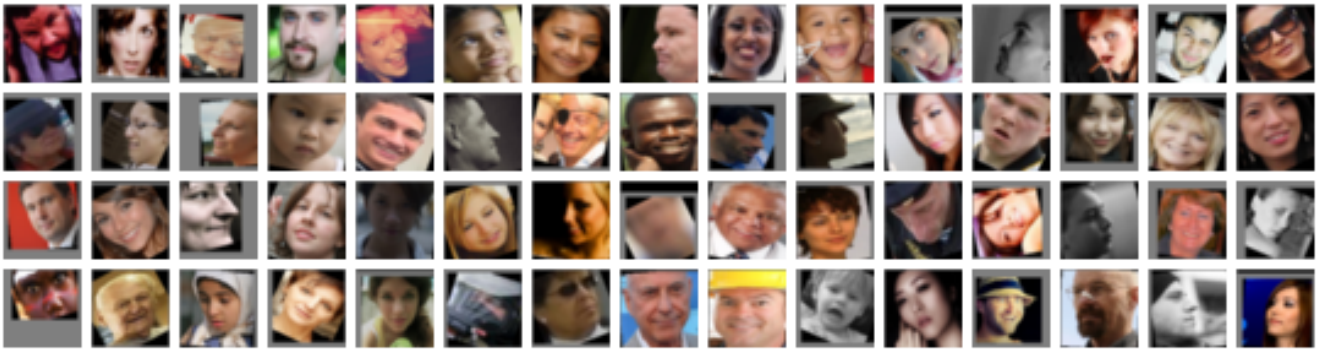
supervised approach to improve the precision of facial landmark detectors. *CoRR*, abs/1807.00966, 2018. 1

[3] Nathan D. Kalka, Brianna Maze, James A. Duncan, Kevin OrConnor, Stephen Elliott, Kaleb Hebert, Julia Bryan, and Anil K. Jain. Ijb–s: Iarpa janus surveillance video benchmark. *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–9, 2018. 2

[4] R. Ranjan, A. Bansal, J. Zheng, H. Xu, J. Gleason, B. Lu, A. Nanduri, J. Chen, C. D. Castillo, and R. Chellappa. A fast and accurate system for face detection, identification, and verification. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 1(2):82–96, April 2019. 2

[5] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. *CoRR*, abs/1602.00134, 2016. 1

(a)



(b)

(c)

Figure 3: (a) LR images from IJB-S dataset. (b) Subsampled AFLW images. (c) Generated low resolution images from G_1. Bluish tinge and pixelated effect can easily be observed in the generated image. Change in color scheme is evident.