Supplementary Material for Generate to Adapt: Resolution Adaption Network for Surveillance Face Recognition

Han Fang, Weihong Deng*, Yaoyao Zhong, and Jiani Hu

Beijing University of Posts and Telecommunications
{fanghan, whdeng, zhongyaoyao, jnhu}@bupt.edu.cn

Abstract. In this supplementary material, we present detailed information including: 1. ablation study; 2. sample face of MR-GAN; 3. details of constructed IJB-C TinyFace; 4. learning algorithm of Multi-Resolution Generative Adversarial Networks (MR-GAN); 5. detailed network architectures; 6. learning algorithm of feature adaption network.

Fig. 1. Visualization for effects of local networks. **MR-GAN (1)** represents the MR-GAN without local generators, attention networks and local discriminators, which is the simplest baseline. **MR-GAN (2)** represents the MR-GAN without local generators and attention networks. And **MR-GAN(3)** shows the results of MR-GAN without local generators. Besides, the last row shows the compared faces generated by MR-GAN.

1 Ablation Study

Effects of attention networks on MR-GAN. To confirm the advances of our attention part, we investigate the effect of attention networks and attention

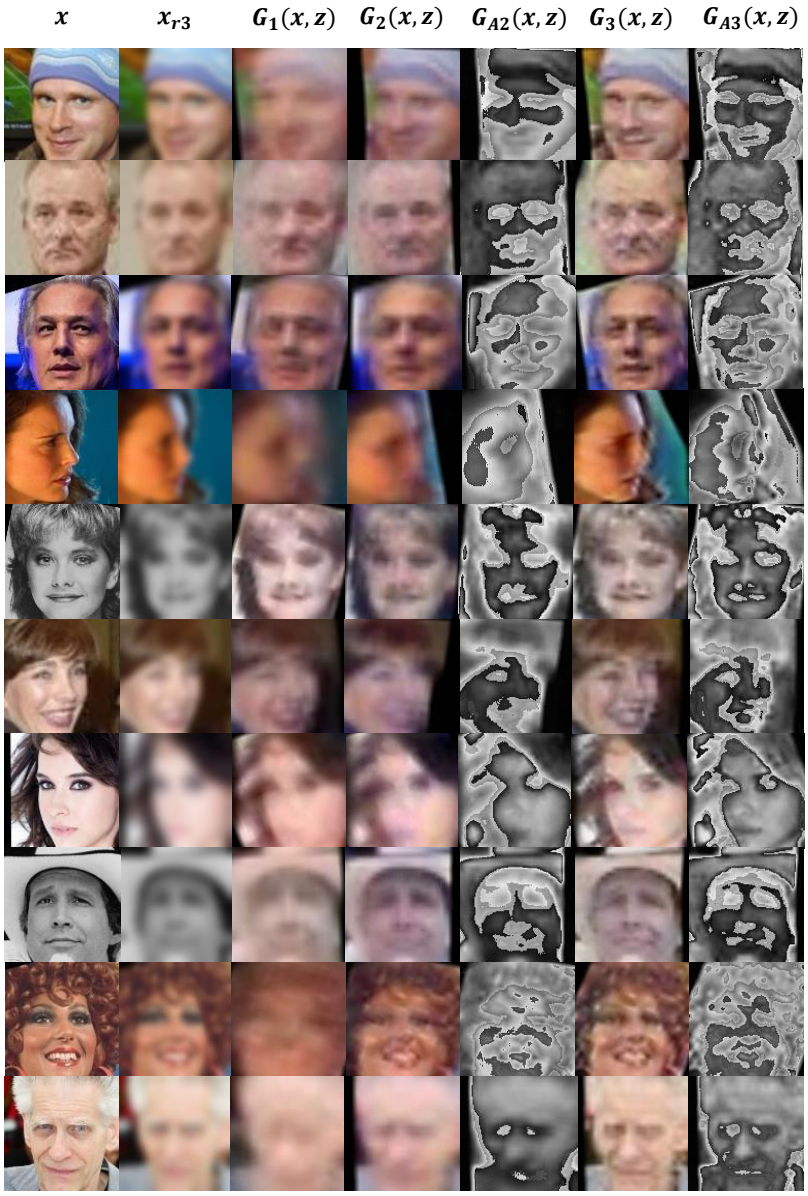

Fig. 2. Visualization for spatial attention in different cases. The leftmost face x is the input face. And x_{r3} is the down-sampled face of the lowest resolution. $G_1(x,z)$ is the face generated by MR-GAN without spatial attention mechanism. $G_2(x,z)$ is the face generated by MR-GAN without attention activation loss and $G_{A2}(x,z)$ is the corresponding attention map. Besides, $G_3(x,z)$ is the face generated by MR-GAN and $G_{A3}(x,z)$ is the corresponding attention map.

Table 1. Evaluation results on IJB-C TinyFace 1:1 covariate protocol.

Method	10^{-7}	10^{-6}	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}
MR-GAN + Adaption (RAN)	**0.0699**	**0.1031**	**0.1616**	**0.2287**	**0.3273**	**0.4817**	**0.7095**
MR-GAN + Adaption(*Cosine* loss)	0.0479	0.0837	0.1253	0.1814	0.2722	0.4181	0.6488
MR-GAN + Adaption(*Euclidean* loss)	0.0553	0.0861	0.1277	0.1848	0.2728	0.4213	0.6525
MR-GAN (w/o attention loss) + Adaption	0.0812	0.1065	0.1568	0.2209	0.3160	0.4666	0.6911
MR-GAN (w/o spatial attention)+ Adaption	0.0664	0.1011	0.1421	0.2015	0.2915	0.4411	0.6810
MR-GAN (w/o local generators)+ Adaption	0.0771	0.1086	0.1564	0.2188	0.3101	0.4545	0.6839
MR-GAN (w/o local generators and spatial attention)+ Adaption	0.0631	0.0924	0.1406	0.2024	0.2888	0.4281	0.6613
MR-GAN (w/o local generators, spatial attention and local discriminators)+ Adaption	0.0709	0.1014	0.1430	0.1999	0.2871	0.4257	0.6556

activation loss qualitatively and quantitatively. We train the same MR-GAN without activation loss and attention networks respectively and report the results of IJB-C TinyFace in Table 1. And we visualize some sample faces and their attention maps in Figure 2. Without spatial attention and activation loss, MR-GAN focuses more on reducing the resolution of global information and ignores the modification around the details. Compared with $G_{A2}(x, z)$, the attention maps of $G_{A3}(x, z)$, which supervised by attention activation loss, can learn more identity-relevant regions and help global generator to pay attention to the learned masks. And the masks learned by $G_{A2}(x, z)$ are over-smoothed and lost some important regions, which can deteriorate recognition performance to some extent.

Effects of local networks on MR-GAN. To investigate the effectiveness of local networks, we train MR-GAN without different parts respectively. The quantitative results are shown in Table 1. Row 6 shows the results of MR-GAN without local generators. Furthermore, Row 7 adopts global generator without the spatial attention networks. And Row 8 depicts results which reduces local generators, attention networks and local discriminators to show a simple baseline. As reported in Table 1, the results of IJB-C TinyFace decrease in turn. To better compare with the faces, we also visualize the sample faces and show in Figure 1. Without paying attention to facial details, MR-GAN can still generate low-resolution faces. However, the aim of MR-GAN is to generate the LR faces which can be used for data augmentation. The preservation of realism around the facial details can help adaption networks to achieve the better performance.

Effects of different loss functions on feature adaption networks. To avoid deteriorating the HR feature space by directly minimizing HR and LR features, we propose the novel translation gate to translate the HR feature into LR feature and minimize the distance between translated LR features and realistic LR features. Only minimizing the distance in the low-resolution domain will make the network to generate more realistic $f_{LR}^{Translate}(x_{HR})$ without affecting the HR feature space. Besides, f_{HR} can be gradually translated into f_{MR} by generating $f_{LR}^{Translate}(x_{HR})$ and preserving enough LR representations. To effectively minimize the distance, we explore the influence of L_1 loss, *Cosine* loss and *Euclidean* loss in IJB-C TinyFace and report the results in Table 1. In our paper, we have empirically chosen the best way to minimize the distance.

Effects of W in translation gate. To demonstrate the change of W, we visualize the whole process and depict in Figure 3. As we mentioned in paper, due to the limited LR representations in f_{HR} during the early stage of training, translation gate adopts translator to amplify the LR representations to obtain

$T_{LR}(f_{HR})$ and $T_{LR}(f_{HR})$ plays the dominant role in the feature and probabilistic supervision. So $T_{LR}(f_{HR})$ can be more close to the LR domain and W increases rapidly. Then as HR features can preserve and provide more realistic LR representations gradually without amplification, f_{HR} can be translated into f_{MR} directly and weight of f_{HR} will increase in weighted architecture. Besides, because we set the small value in hyper-parameters of low-resolution adversarial networks, W will decrease and maintain within a stable range to balance two sources of low-resolution representations. With this weighted translation, f_{HR} can retain enough LR representations to construct resolution-robust embedding.

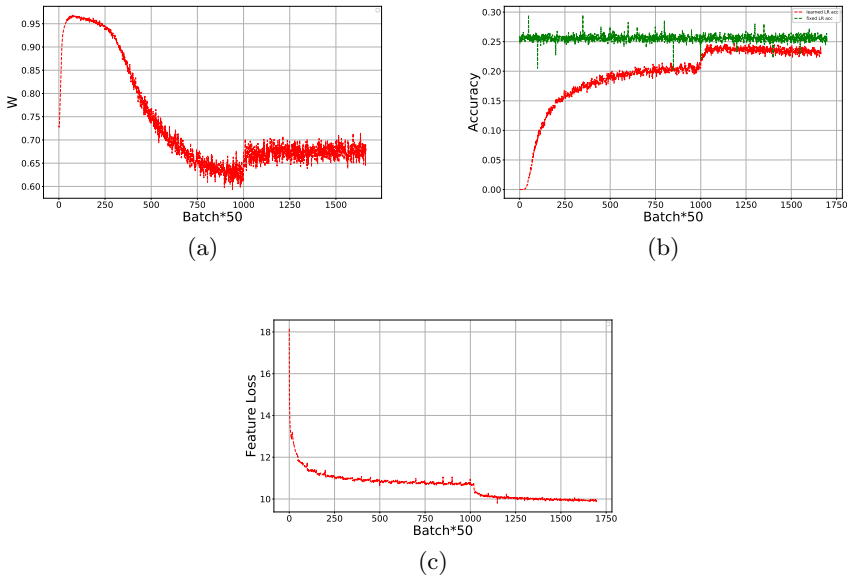

Fig. 3. (a): Visualization for change of W in translation gate. (b): Visualization for change of accuracy in LR domain. The red curve represents the learned LR accuracy by translation gate and the green curve depicts the fixed LR accuracy obtained by LR model. (c): Visualization for L_f in feature supervision.

2 Sample Face of MR-GAN

We visualize the low-resolution face synthesis of MR-GAN on CASIA-WebFace [4] in Figure 4. Our MR-GAN presents a great identity preserved but blurred enough faces. In our MR-GAN, we observe the facial regions of low resolution can also provide a part of discriminative information, which can improve the performance in low-resolution face recognition.

Fig. 4. The sample results on CASIA-WebFace [4]. The leftmost face in each group is the input face, and the rest 3 faces are the synthesized visualizations from down-sampling, Cycle-GAN [5] and MR-GAN.

3 Details of IJB-C TinyFace

The IARPA Janus Benchmark-C face challenge (IJB-C) [3] is the challenging dataset collected from unconstrained environment with resolution and illumination variations. To focus on studying the problem of low resolution, we utilize the distance between center point of eyes and mouth center as the criterion to define small face and analyze it. The distribution is depicted in Figure 5. Most distances are concentrated below 50 which reveals that IJB-C contains a lot of low-resolution faces. Faces whose distances less than 30 and more than 10 are selected as realistic LR faces, which can avoid the artifacts and large pose variations and provide enough LR representations. We totally get 7,985 LR faces and select the corresponding pairs under the same subject for each selected LR face to construct the positive pairs. For each selected LR face, we randomly select

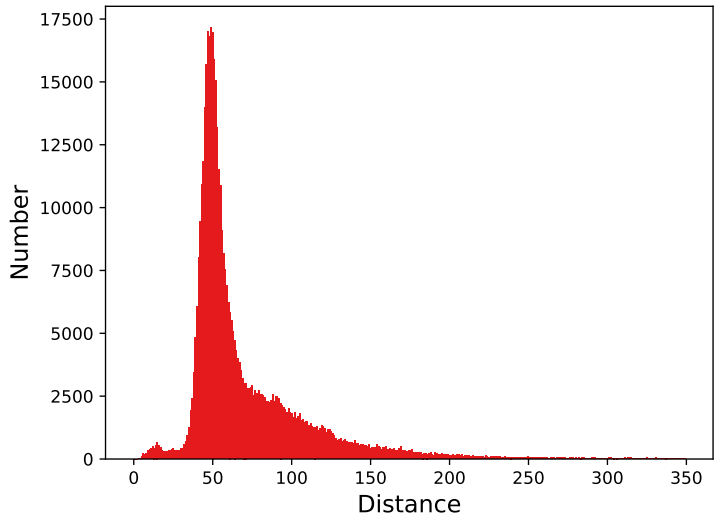

Fig. 5. The distribution of distances between center point of eyes and mouth center in IJB-C [3] .

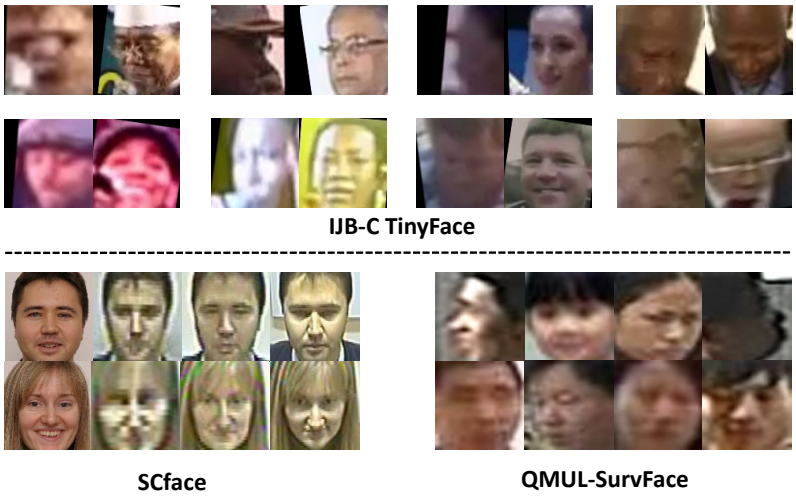

Fig. 6. Top: Example positive pairs from IJB-C TinyFace; **Bottom:** Example face images from SCface [2] and QMUL-SurvFace [1].

20 non-repeating faces with the same subject to construct pairs, and for the subject whose number of face is less than 20, all the faces are selected. Finally 158,338 positive pairs are determined. Following IJB-C 1:1 covariate verification protocol, selected 158,338 positive pairs and the same 39,584,639 negative pairs are used to construct IJB-C TinyFace. More sample faces including SCface [2] and QMUL-SurvFace [1] can be shown in Figure 6.

4 Learning Algorithm of MR-GAN

Algorithm 1 Learning algorithm of MR-GAN

Require:
High-resolution face images x.
Realistic low-resolution face images y.
Max number of epochs E.
Number of network updates per epoch B.
Ensure:
Resolution-aggregated generator G.
Global-local focused discriminator D.

for each $e \in [1, E]$ **do**
  for each $b \in [1, B]$ **do**
    if $b \bmod 5 == 0$ **then**
      Optimize G by employing pixel loss;
    else
      Optimize G without pixel loss;
    end if
    Optimize D;
  end for
  Save G and D in epoch e;
end for

5 Detailed Network Architecture

Architecture of global and local discriminator are shown in Figure 8. In global discriminator network, We use several convolutional layers, instance norm and leaky ReLU to learn and activate the feature maps. PatchGANs [5] are utilized to classify whether the image patches are real or fake. For the local discriminator networks, the sizes of local patch are $20\times72\times3$, $32\times20\times3$ and $24\times52\times3$ for region of eyes, nose and mouth respectively. An to reduce parameters, we only adopt three convolutional layers to focus on real or fake region to help global discriminator to pay attention to discriminating the realism of facial details.

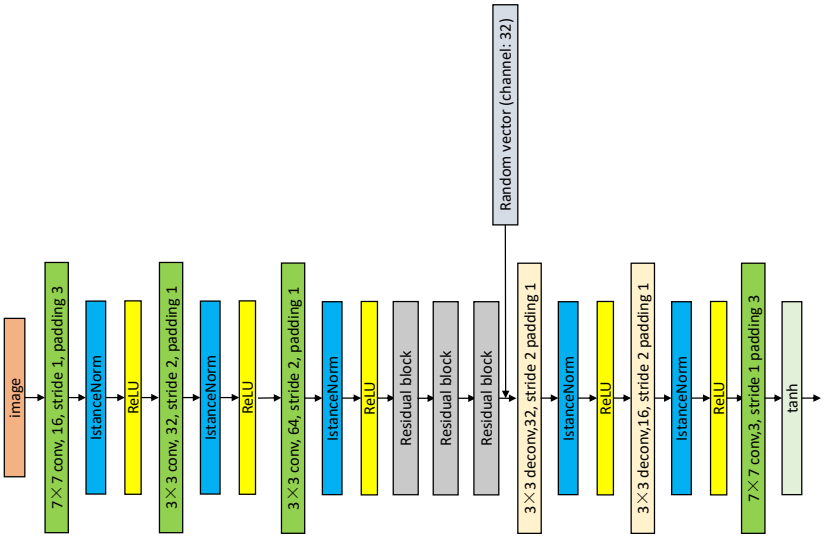

Fig. 7. Local generator consists of several convolutional layers to encode images into feature maps, residual blocks to deepen the network and transposed convolutional layers to decode the local regions. And the random vector is connected with the feature after residual blocks, deepening the channel.

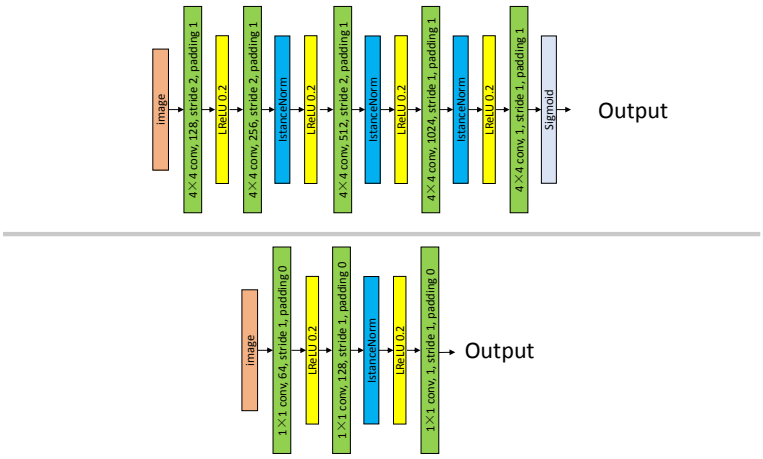

Fig. 8. Architecture of global and local discriminator. **Top:** Global discriminator; **Bottom:** local discriminator. Input of global discriminator is the entire face. And input face are cropped into eyes, nose and mouth, feeding into local discriminator.

The architecture of local generator is depicted in Figure 7. With injection of random vector, local generator aims to blur the local region randomly, avoiding the artifacts. In Figure 9, multi-resolution fusions are adopted to aggregate streams of different resolutions. The information of higher-resolution is repeatedly combined with the feature maps of lower-resolution and finally stream of the lowest resolution contains multiple resolutions which can be randomly selected to refine the LR faces.

6 Learning Algorithm of Feature Adaption Network

Algorithm 2 Learning algorithm of feature adaption network

Require:
High-resolution face images x_{HR}.
Synthesized low-resolution face images x_{LR}.
Pre-trained HR model ($Model_{HR}$) by using x_{HR}.
Pre-trained LR model ($Model_{LR}$) by using x_{LR}.
Max number of epochs E.
Number of network updates per epoch B.

Ensure:
Multi-resolution model $Model_{MR}$.
Low-resolution discriminator D.
for each $e \in [1, E]$ **do**
  for each $b \in [1, B]$ **do**
    1. Forward propagating $Model_{LR}$ to obtain $f_{LR}^{Real}(x_{LR})$;
    2. Forward propagating $Model_{MR}$ to obtain $f_{LR}^{Translate}(x_{HR})$ (input W=1);
    3. Forward propagating D and optimize D;
    4. Forward propagating D and obtain weight W;
    5. Forward propagating $Model_{MR}$ with W;
    if $b \bmod 4 == 0$ **then**
      Optimize $Model_{MR}$ by employing $L_{feature}^{G}$;
    else
      Optimize $Model_{MR}$ without $L_{feature}^{G}$;
    end if
  end for
end for

References

1. Cheng, Z., Zhu, X., Gong, S.: Surveillance face recognition challenge. arXiv preprint arXiv:1804.09691 (2018)
2. Grgic, M., Delac, K., Grgic, S.: Scface–surveillance cameras face database. Multimedia tools and applications **51**(3), 863–879 (2011)

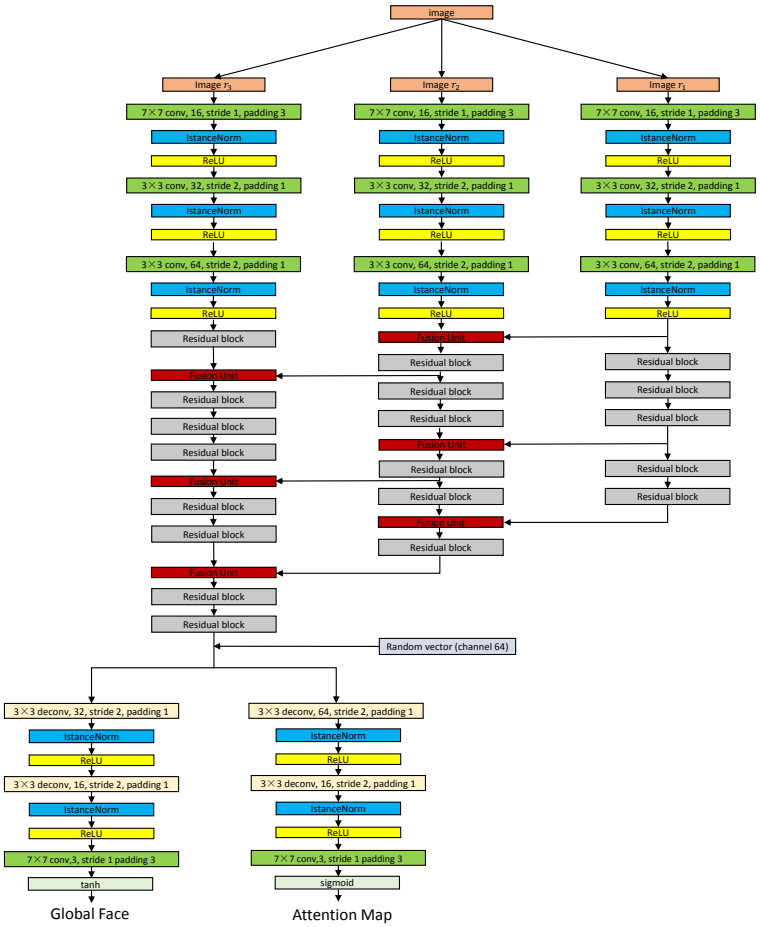

Fig. 9. The architecture of global generator.

3. Maze, B., Adams, J., Duncan, J.A., Kalka, N., Miller, T., Otto, C., Jain, A.K., Niggel, W.T., Anderson, J., Cheney, J., et al.: Iarpa janus benchmark-c: Face dataset and protocol. In: 2018 International Conference on Biometrics (ICB). pp. 158–165. IEEE (2018)
4. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Learning face representation from scratch. arXiv preprint arXiv:1411.7923 (2014)
5. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)