Supplementary Material: Unsupervised Domain Adaptation for Face Recognition in Unlabeled Videos

Kihyuk Sohn[1] Sifei Liu[2] Guangyu Zhong[3] Xiang Yu[1] Ming-Hsuan Yang[2] Manmohan Chandraker[1,4]
[1]NEC Labs America [2]UC Merced [3]Dalian University of Technology [4]UC San Diego

S1. Implementation Details

In this section, we present detailed network architecture as well as implementation details, for reproducible research. The network architecture for the reference network (RFNet) and video domain-adapted network (VDNet) are equivalent, as described in Table S1. The network architecture is mainly motivated from [8], but we replace ReLU with maxout units at every other convolution layer to further improve the performance while maintaining the same number of neurons at each layer.

The RFNet is pretrained on the CASIA webface database [8], which includes 494414 images from 10575 identities from the Internet, using the same training setup [7]. The implementation is based on Torch [1] with $N = 1080$ (that is, number of examples per batch is set to 2160) for N-pair loss. The N-pair loss, which pushes (N-1) negative examples at the same time while pulling a single positive example, is used on 8 GPUs for training. The VDNet is initialized with the RFNet followed by a discriminator composed of two fully connected layers (160, 3) followed by a ReLU on top of 320-dimensional output feature of VDNet. The VDNet is trained with the following objective function:

$$\mathcal{L} = \mathcal{L}_{\text{FM}} + \alpha\mathcal{L}_{\text{FR}} + \beta\mathcal{L}_{\text{IC}} + \gamma\mathcal{L}_{\text{Adv}} \quad \text{(S1)}$$

where the forms of the loss functions are described in the main paper and we set $\alpha = \beta = \gamma = 1$ for all our experiments. The learning rate is set to 0.0003 with a default setting of the Adam optimizer [5] (for example, $\beta_1 = 0.9$, $\beta_2 = 0.999$). The network is trained for 1500 iterations with batch size of 512, where we allocate 256 examples from the image domain and remaining 256 examples from the video domain for each mini batch.

S2. Ablation Study

In this section, we present further results for different design choices of the proposed algorithm. In particular, we consider the alternative of pixel-space image restoration and study the effect of scale of unlabeled video training data.

Table S1. Network architecture for RFNet and VDNet. The network is composed of 10 layers of 3×3 convolution layers followed by either ReLU or maxout units [2]. The volumetric max pooling (Vmax pooling) extends (spatial) max pooling to input channels and can be used to model maxout units.

operation	kernel	output size
Conv1-1 + ReLU	3×3	100×100×32
Conv1-2	3×3	100×100×128
Vmax pooling	2×2×2	50×50×64
Conv2-1 + ReLU	3×3	50×50×64
Conv2-2	3×3	50×50×256
Vmax pooling	2×2×2	25×25×128
Conv3-1 + ReLU	3×3	25×25×96
Conv3-2	3×3	25×25×384
Vmax pooling	2×2×2	13×13×192
Conv4-1 + ReLU	3×3	13×13×128
Conv4-2	3×3	13×13×512
Vmax pooling	2×2×2	7×7×256
Conv5-1 + ReLU	3×3	7×7×160
Conv5-2	3×3	7×7×320
Average pooling	7×7	1×1×320

Table S2. Architecture for image or video pixel-space restoration network. The network is composed of 8 residual blocks and few more convolution layers before and after a series of residual blocks. There also exists a shortcut connection where the output of C1 and C2 are added before fed into C3.

name	operation	kernel	output size
C1	Conv + ReLU	3×3	100×100×32
Res1–8	Conv + BN + ReLU + Conv + BN	3×3	100×100×32
C2	Conv + BN	3×3	100×100×32
C3	Conv + ReLU	3×3	100×100×32
C4	Conv + ReLU	3×3	100×100×32
C5	Conv + Tanh	3×3	100×100×1

S2.1. Pixel-space Restoration and Adaptation

The focus of our paper is feature-level domain adaptation. In Section 3.3 of the main paper, lines 368–371, we state that this is preferable over pixel-level alternatives. To illustrate this further, we now compare the performance of our proposed method to several baselines on pixel-space restoration

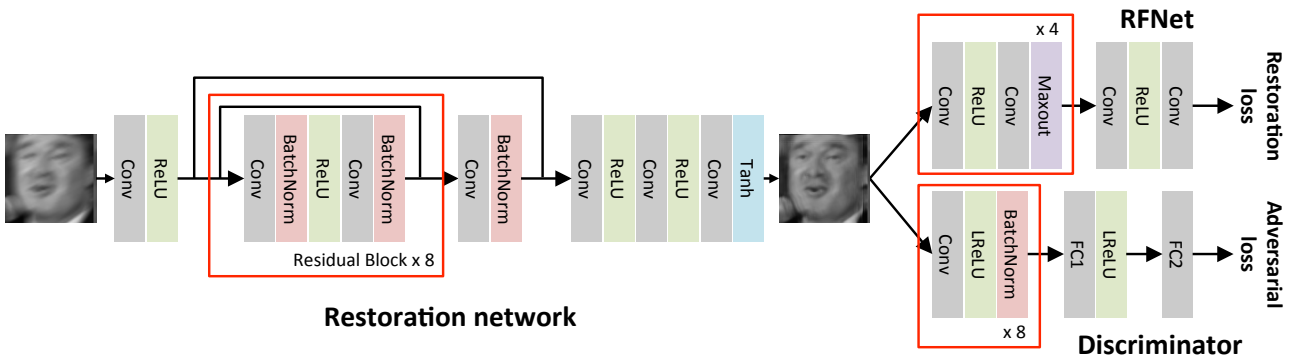


Figure S1. Illustration of pixel-space restoration and adaptation framework. The restoration network is composed of 8 residual blocks and a shortcut connection as described in Table S2. The network architectures of RFNet and discriminator are provided in Table S1 and S3, respectively. The input for restoration network is either synthetically degraded images (for IRResNet or IRGAN) or video frames (for VRGAN) and the output is fed into either RFNet for feature restoration loss with restored images or discriminator for adversarial loss with both restored images and video frames.

Table S3. Network architecture for discriminator. The network is composed of 8 layers of 3×3 convolution layers followed by by Leaky ReLU (LReLU) and Batch Normalization (BN) layers and 2 fully-connected (FC) layers whose final output is either 2 for IRGAN or 3 for VRGAN.

operation	kernel, stride	output size
Conv1-1 + LReLU	3×3, 1	100×100×64
Conv1-2 + LReLU + BN	3×3, 2	50×50×64
Conv2-1 + LReLU + BN	3×3, 1	50×50×128
Conv2-2 + LReLU + BN	3×3, 2	25×25×128
Conv3-1 + LReLU + BN	3×3, 1	25×25×256
Conv3-2 + LReLU + BN	3×3, 2	13×13×256
Conv4-1 + LReLU + BN	3×3, 1	13×13×512
Conv4-2 + LReLU + BN	3×3, 2	7×7×512
FC1 + LReLU	–	1024
FC2	–	2 (IRGAN) 3 (VRGAN)

Table S4. Face verification accuracy on the degraded LFW dataset. The baseline network (RFNet) is evaluated on both degraded and original (†) test set. We run evaluation on degraded test images for 10 times with different random seeds and report the mean accuracy.

Model	baseline	IRResNet	IRGAN	VRGAN	VDNet-F
VRF	88.68	92.69	92.38	92.41	93.72
VRF†	98.85				

and domain adaptation.

The pixel-space restoration applies a similar combination of loss strategies, but shifts the restored domain from feature to image pixels. Instead of training VDNet, we use RFNet for feature restoration loss and a discriminator for domain adversarial loss on top of a "restored" input image, with an additionally trained image restoration network. Based on the single-image super-resolution generative adversarial network (SRGAN) [6], we train several image restoration networks for face images, which we call image-restoration ResNet (IRResNet), image-restoration GAN (IRGAN) and video-restoration GAN (VRGAN), as follows:

- **IRResNet**: An image restoration network based on very deep residual network [3], trained with feature restoration loss guided by pretrained face recognition engine (RFNet) on synthetically degraded images.
- **IRGAN**: An image restoration network based on very deep residual network trained with feature restoration loss as well as discriminator loss on synthetically degraded images.
- **VRGAN**: A video restoration network based on very deep residual network trained with feature restoration loss as well as discriminator loss on both synthetically degraded images and video frames.

The pixel-space restoration models are illustrated in Figure S1. The network architectures for the pixel-space restoration network and discriminator are summarized in Tables S2 and S3, respectively. We use grayscale images for restoration networks as our RFNet accepts grayscaled images as input.[1] We use Adam optimizer with a learning rate of 0.0003. Compared to the training of VDNet, we reduce the batch size to 96, where we allocate 48 examples from the image domain and another 48 examples from the video domain for each mini batch. This is due to additional CNNs such as restoration network and discriminator. We increase the number of iterations to 30000 due to slower convergence.

For fair comparisons, we apply the same set of random noise processes to generate synthetically degraded images

[1] Although we use grayscale images as input and output of the restoration network, one can construct the restoration network for RGB images with an additional differentiable layer that converts RGB images into grayscale images before feeding into RFNet.

Table S5. Video face verification accuracy and standard error on the YTF database with different image- and video-restoration networks. The verification accuracy averaged over 10 folds and corresponding standard error are reported. The best performer and those with overlapping standard error are boldfaced.

Model	fusion	1 (fr/vid)	5 (fr/vid)	20 (fr/vid)	50 (fr/vid)	all
baseline	–	91.12±0.318	93.17±0.371	93.62±0.430	93.74±0.443	93.78±0.498
IRResNet	–	90.40±0.366	92.59±0.388	93.13±0.405	93.15±0.416	93.26±0.400
IRGAN	–	90.67±0.314	92.88±0.369	93.25±0.389	93.24±0.368	93.22±0.383
VRGAN	–	90.77±0.346	92.93±0.402	93.41±0.429	93.46±0.410	93.62±0.439
F (ours)	–	92.17±0.353	94.44±0.343	**94.90**±0.345	**94.98**±0.354	**95.00**±0.415
	✓	–	94.52±0.356	**95.01**±0.352	**95.15**±0.370	**95.38**±0.310

Table S6. Video face verification accuracy and standard error on the YTF database with different number of unlabeled videos for domain-adversarial training. The verification accuracy averaged over 10 folds and corresponding standard error are reported. VDNet model F is used for experiments, where all four losses including feature matching, feature restoration, image classification, as well as domain adversarial losses, are used. The best performer and those with overlapping standard error are boldfaced.

# videos	fusion	1 (fr/vid)	5 (fr/vid)	20 (fr/vid)	50 (fr/vid)	all
10	–	91.54±0.339	93.62±0.338	94.05±0.350	94.17±0.390	94.16±0.369
	✓	–	93.63±0.365	94.12±0.377	94.22±0.386	94.22±0.381
25	–	91.80±0.337	93.84±0.320	94.24±0.328	94.38±0.321	94.46±0.323
	✓	–	93.90±0.331	94.29±0.338	94.40±0.334	94.56±0.333
100	–	92.34±0.289	94.42±0.348	**94.78**±0.379	**94.82**±0.385	**94.78**±0.363
	✓	–	94.50±0.331	**94.85**±0.388	**94.87**±0.403	**95.00**±0.417
250	–	92.03±0.348	94.21±0.342	94.70±0.314	94.73±0.324	**94.90**±0.291
	✓	–	94.23±0.348	94.70±0.322	**94.81**±0.313	**94.98**±0.323
all (∼2780)	–	92.17±0.353	94.44±0.343	**94.90**±0.345	**94.98**±0.354	**95.00**±0.415
	✓	–	94.52±0.356	**95.01**±0.352	**95.15**±0.370	**95.38**±0.310

for training, as described in Section 3.2. We note that other approaches might be used for image restoration, including ones that rely on further supervision through specification of a restoration target within a video sequence. But our baselines based on synthetically degraded images are reasonable in a setting comparable to VDNet that uses only unlabeled videos and produce visually good results. In particular, we visualize the image restoration results on synthetically degraded images of the LFW test set in Figure S2 and note that each baseline produces qualitatively reasonable outputs.

For quantitative validation of IRResNet, IRGAN and VRGAN, we evaluate the face verification performance on synthetically degraded Labeled Faces in the Wild (LFW) [4] dataset, where we apply the same set of image degradation kernels as used in the training with images of the LFW test set. The results are summarized in Table S4. The image restoration methods effectively improve the performance from 88.68% to 92.69%, 92.38%, and 92.41% with IRResNet, IRGAN, and VRGAN, respectively. On this test set, adversarial training does not improve the verification performance since it aims to perform global distribution adaptation, thereby losing discriminative information that can be directly learned from the feature restoration loss defined between corresponding synthetic and original images. However, none of the image restoration models are as effective as the feature-level restoration model (VDNet model F in the main paper), which achieves 93.72%.

We further evaluate the verification performance on the YouTube Face database (YTF) whose results are summarized in Table S5. Different from the results on the synthetically degraded LFW dataset, there is no performance gain when evaluated on YTF dataset with image and video restoration networks. In addition, we observe significant performance drops when trained only on the synthetic image database. Training VRGAN with an additional domain adversarial loss for video data improves the performance on video face verification over restoration models trained only with synthetically degraded images, but the improvement is not as significant as we have observed from the VDNet experiments.

In summary, it is evident that aligning distributions in pixel space is more difficult and there is a clear advantage of feature-level domain adaptation, especially when our ultimate goal is to improve the performance of high-level tasks such as classification.

S2.2. Number of Unlabeled Videos

We perform controlled experiments with different number of unlabeled videos at training. Specifically, we utilize randomly selected 10, 25, 100 and 250 videos of YTF database for training. As shown in Table S6, we observe a general trend that the more video data is used for training, the higher verification accuracy we obtain. For example, if we use only 10 videos for training, we obtain 94.22% accuracy which is far lower than the best accuracy of 95.38% achieved using all videos for training, which is approximately 2780 unlabeled videos for each training fold. It is also worth noting that even using a very small number of unlabeled videos for training already improves the performance over other

Figure S2. Image restoration on synthetically degraded LFW test set. From top to bottom, we visualize ground truth images, synthetically degraded images, and restored images with IRResNet, IRGAN, and VRGAN, respectively.

models that do not use domain adversarial training, such as model B (93.94%) or model C (93.82%). But at the same time, the margin of improvement gets smaller as we include more unlabeled videos for training.

Our paper demonstrates initial success with utilizing unlabeled video data for video face recognition, but interesting further problems remain. A promising direction of future is to explore ways to better utilize the fact that unlabeled videos are easier to acquire than labeled ones, translating to steady improvements in recognition performance as progressively larger scales of unlabeled data become available.

S3. Qualitative Visualization of Guided Fusion

We present further examples for the qualitative visualization of three-way domain discriminator. Similar to Figure 3 in the paper, we sort and show ten additional video clips with the confidence score of discriminator in Figure S3. As in the main paper, we observe that the discriminator ranks the video frames in a reasonable order, encompassing variations along several meaningful factors.

References

[1] R. Collobert, K. Kavukcuoglu, and C. Farabet. Torch7: A Matlab-like environment for machine learning. In *BigLearn, NIPS Workshop*, 2011. 1

[2] I. Goodfellow, D. Warde-farley, M. Mirza, A. Courville, and Y. Bengio. Maxout networks. In *ICML*, 2013. 1

[3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2

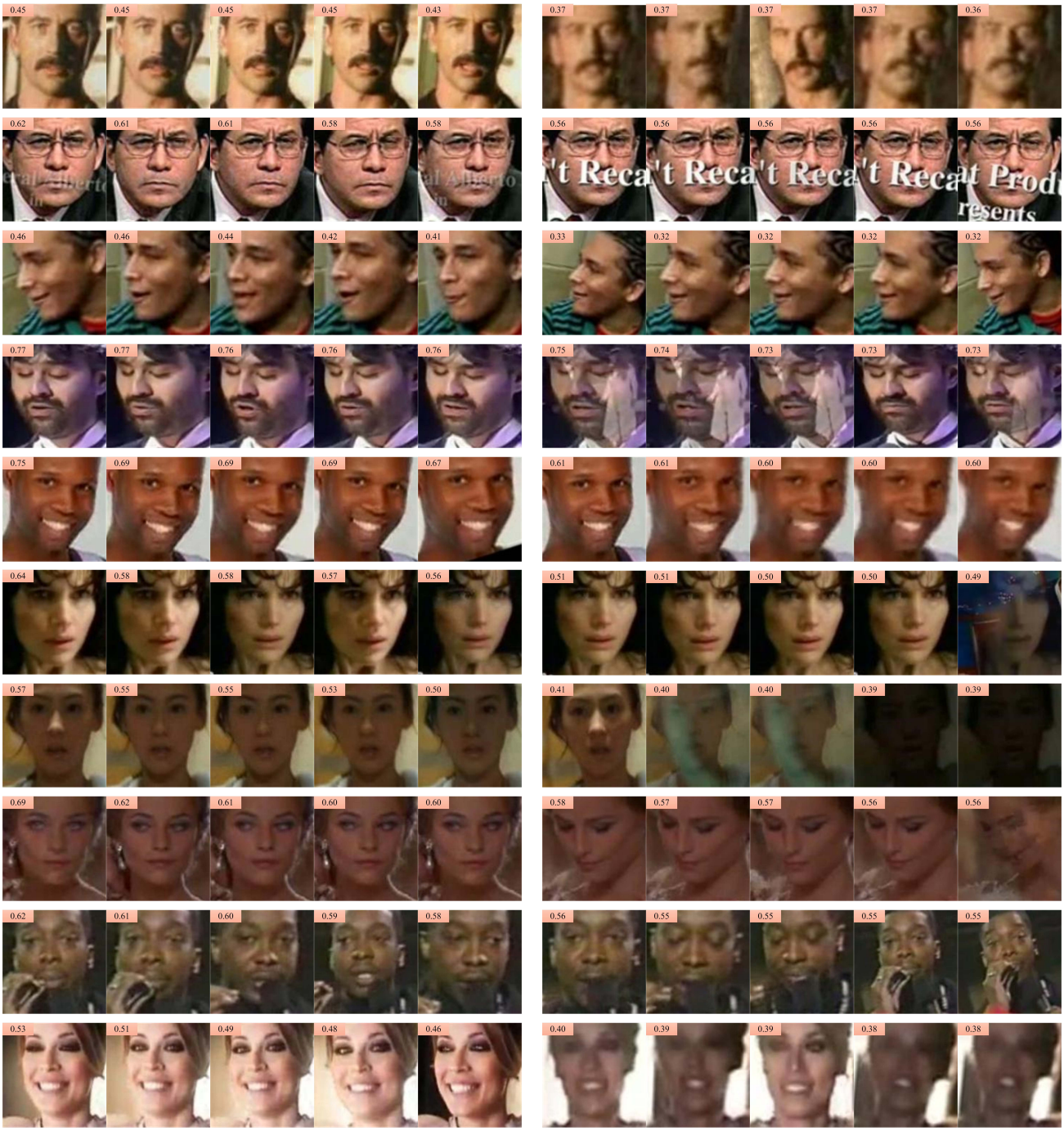
[4] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, University of Massachusetts, Amherst, 2007. 3

[5] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 1

[6] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 2017. 2

[7] K. Sohn. Improved deep metric learning with multi-class N-pair loss objective. In *NIPS*. 2016. 1

[8] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *CoRR*, abs/1411.7923, 2014. 1



(a) top-5 (b) bottom-5

Figure S3. We sort the frames within a sequence in a descending order with respect to the confidence score of three-way discriminator ($\mathcal{D}(y = 1|v)$), and display them by showing the top-5 and bottom-5 instances, respectively. The weights are shown in the upper-left corner of each frame.