Reduced Residual Nets (Red-Nets): Low Powered Adversarial Outlier Detectors

Saad Sadiq, Marina Zmieva, Mei-Ling Shyu
Department of Electrical and Computer Engineering
University of Miami
Coral Gables, FL, USA
s.sadiq@miami.edu, mxz391@miami.edu, shyu@miami.edu

Shu-Ching Chen
School of Computing and Information Sciences
Florida International University
Miami, FL, USA
chens@cs.fiu.edu

***Abstract*—The evolution of information science has seen an immense growth in multimedia data, specially in the case of CCTV live stream capture. The tremendously large volumes of multimedia data give rise to a particularly challenging problem called the outlier events of interest detection. In the wake of growing school shootings in the United States, there needs to be a rethinking of our security strategies regarding the safety of children at school utilizing multimedia data mining research. This paper proposes a novel method to identify faces of interest using live stream CCTV data. By integrating the adversarial information, the proposed framework can help imbalance facial recognition and enhance rare class mining even with trivial scores from the minority class. Experimental results on the Faces in the Wile (FIW) dataset demonstrate the effectiveness of the proposed framework with promising performance. The proposed method was implemented on a low powered Nvidia TX2 for real-time face recognition. The proposed framework was benchmarked against several existing state-of-the-art methods for accuracy, computational complexity, and real-time power measurement. The proposed method performs very well under the power and complexity constraints.**

***Keywords*-Face recognition; Rare class mining; Low powered; Realtime multimedia.**

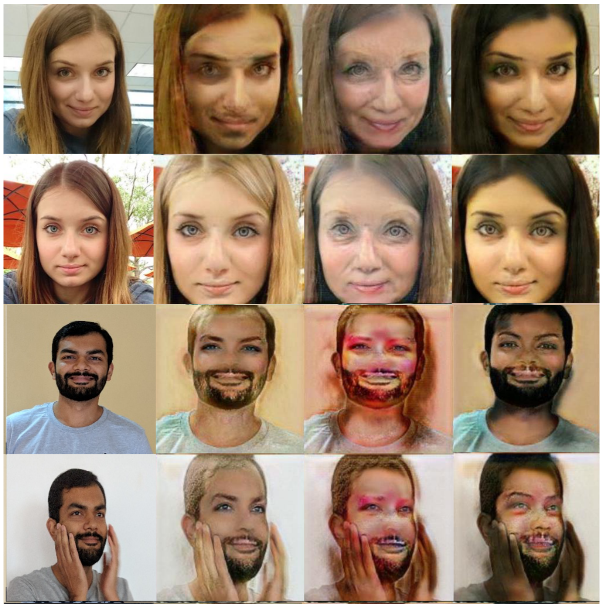

Figure 1: Adversarially generated images to capture key markers of the target faces

I. Introduction

The evolution of information science has seen an immense growth of data, which has attracted many researchers to the field of multimedia data mining. In many real-world multimedia applications, massive quantities of data are highly imbalanced. Large volumes of multimedia data are generated everyday, which gives rise to a particularly challenging problem called the outlier events of interest detection. These outlier events occur very infrequently and the detection of these events is an interesting research problem. Despite rigorous research endeavors, outlier detection remains one of the most challenging problems in information science, particularly for multimedia data.

Among them, outlier event mining from the imbalanced data has gained more attentions as lots of applications do not have uniform class distributions [1]. That is, the majority of the cases belong to some classes (i.e., the majority classes) and far fewer data instances belong to the minority classes. The minority classes, however, represent the outlier cases of interest, like unusual events in surveillance, disaster events, intrusion detection, etc. Most classifiers are modeled by exploring data statistics. As a result, they may be biased towards the majority classes and show very poor classification accuracy results on the minority classes, which is one of the centric research tasks in content-based information retrieval [2], [3]. To overcome this challenge, a lot of effort has been put into Scale Invariant Feature Transform (SIFT) and Histogram of Oriented Gradients (HOG) based feature detectors [4]–[7]. Other methods try to increase the ratio of positive and negative data (for example, video frames) to improve the classification accuracy for automatic labeling and to build the correlations between the labeled concepts to utilize the underlying predictors [8]–[10].

In the wake of recent school shootings, there has been a rethinking of our strategies regarding the safety of children at school. Deep learning is at the forefront of face recognition, however, deep learning models require a substantial amount of computing power in order to perform face recognition. Moreover, they require a large number of training samples for each suspect in order to classify them correctly. Thus the need is to develop a face recognition system that can

1) Work with a minimal number of annotated photos
2) Cheap and compact to be deployed at scale
3) Low powered so can use 9volts
4) Able to work online and offline

The rest of this paper is organized as follows. In Section 2, previous work on facial recognition and outlier target of interest classification are discussed. Section 3 describes a novel idea of reducing the method complexity in outlier or rare face detection using the proposed Reduced RESNets. Section 4 shows the experiment and compares the results of the proposed system on the Labeled Faces in the Wild (LFW) data set. Finally, Section 5 draws the conclusions and lays out directions for future research.

II. Previous Work

All face recognition methods depend upon retrieving and comparing available faces in the databases to seek a possible suspect match. However, with the tremendous increase in the data size, the complexity and cost of the data storage and retrieval for multimedia research and applications have also increased tremendously [1], [11], [12]. How to store and index multimedia data in various media types including video, audio, image, text, etc. for efficient and effective data retrieval has drawn a lot of attention [13]–[15]. To solve this problem, multimedia data is labeled with respect to their real high-level semantic meanings such as “Person”, “Boat”, and “Football”. These labels are often referred to as “concepts” or “semantic concepts” [16], [17]. The foremost challenge in this research domain is to reduce the gap between the low-level features [18], [19] and high-level semantic concepts [19]–[22], i.e., to build a connection between the different meanings and conceptions formed by different representation systems.

Previous work on adversarial training at scale has produced encouraging results, showing strong robustness to (single-step) adversarial examples (Goodfellow et al., 2015; Kurakin et al., 2017b). Yet, these results are misleading, as the adversarially trained models remain vulnerable to simple black-box and white-box attacks. The results by Kurakin et al. [23] suggest that adversarial training can be improved by decoupling the generation of adversarial examples from the model being trained. Compared to IcGAN [24], their model demonstrates an advantage in preserving the facial identity feature of an input. This is because their method maintains the spatial information by using activation maps from the convolutional layer as a latent representation, rather than just a low-dimensional latent vector as in IcGAN.

It can be seen that DeepFool [25] estimates small perturbations in their generators. On the ILSVRC2012 challenge dataset, the average perturbation is one order of magnitude smaller compared to the fast gradient method. Adversarial training caused a slight (less than 1%) decrease of accuracy on clean examples in our ImageNet experiments. This differs from the results of adversarial training reported previously, where adversarial training increased accuracy on the test set (Goodfellow et al., 2014; Miyato et al., 2016b;a). One possible explanation is that adversarial training acts as a regularizer. For datasets with few labeled examples where overfitting is the primary concern, adversarial training reduces the test error. It is noteworthy to see that when using Jacobian Clamping on MNIST, CIFAR-10, and STL-10 samples, reducing the number of discriminator steps does not reduce the score, but it more than halves the wallclock time [26]. Therefore, all zero-risk linear classifiers are not robust to adversarial perturbations. Unlike linear classifiers, a more flexible classifier that correctly captures the orientation of the lines in the images will be robust to adversarial perturbation, unless this perturbation significantly alters the image and modifies the direction of the line [27].

The authors in [28] trained a robust network against PGD adversaries, that will be robust against a wide range of attacks. This classic attack model allows the adversary to only solve problems that require at most polynomial computation time. They employ an optimization-based view on the power of the adversary as it is more suitable in the context of machine learning. A multi-step adversarial training [29] shows an improvement on white box accuracies from the previous state-of-the-art, from 1.5% to 3.9%. They further improve the white box accuracy from the adversarial training baseline showing an improvement from 3.9% to 27.9%. Adversarial logit pairing also improves black box accuracy from the MPGD baseline, going from 36.5% to 47.1% . Adversarial logit pairings in [30] forces the explanations of a clean example and the corresponding adversarial example to be similar. This is essentially a prior encouraging the model to learn logits that are a function of the truly meaningful features in the image (position of cat ears, etc.) and ignore the features that are spurious (off-manifold directions introduced by adversarial perturbations).

III. The Proposed Framework

Conventional real-time facial recognition systems compute and match all faces in the camera view sequentially. But sometimes we are only interested in identifying a few labeled suspects. In these situations, we don’t need to compare and match everyone so these systems can be implemented at scale. Our proposed method performs Adversarial training for preemptive identification of suspects entering schools. The basic framework design is shown in Figure 2. The framework has 4 novel characteristics

1) Reduced Complexity Residual Network
2) Adversarial training samples
3) Low power Implementation on Nvidia Tx2

The Reduced ResNet network proposed in this paper is inspired from the Deep Residual Learning ResNet from

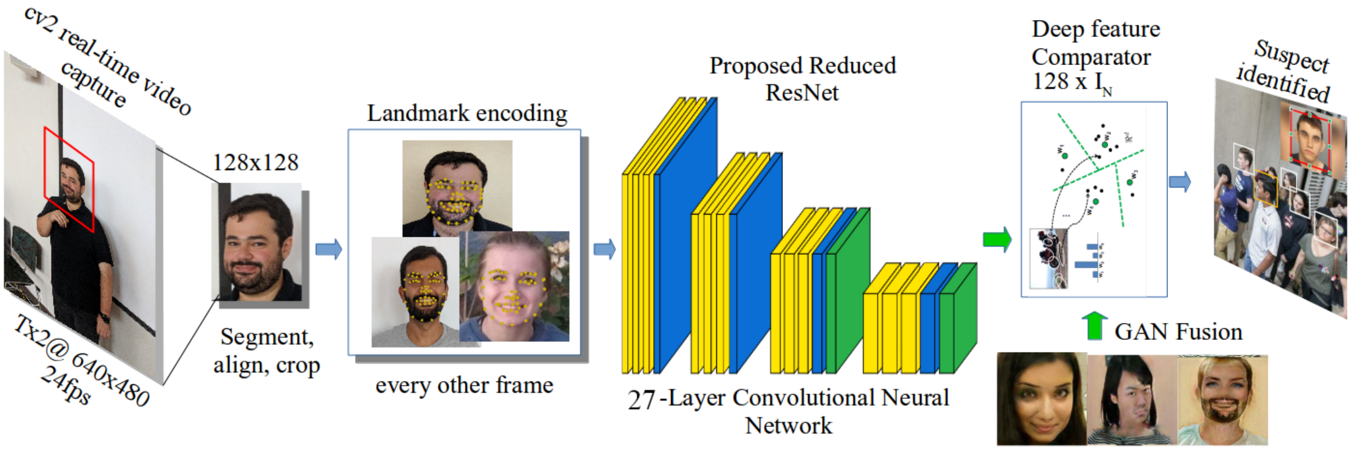

Figure 2: Framework diagram of the proposed low powere facial recognition system

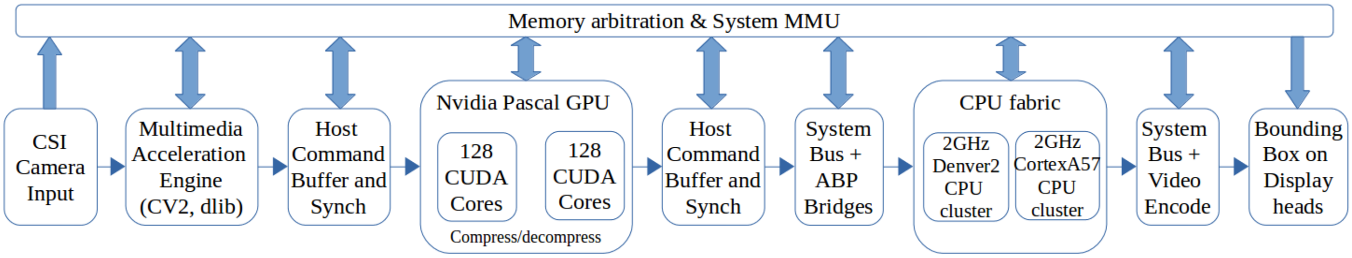
Memory arbitration & System MMU
CSI Camera Input
Multimedia Acceleration Engine (CV2, dlib)
Host Command Buffer and Synch
Nvidia Pascal GPU
128 CUDA Cores
128 CUDA Cores
Compress/decompress
Host Command Buffer and Synch
System Bus + ABP Bridges
CPU fabric
2GHz Denver2 CPU cluster
2GHz CortexA57 CPU cluster
System Bus + Video Encode
Bounding Box on Display heads

Figure 3: Pipeline diagram of the hardware implementation of the proposed framework, on the Nvidia Jetson TX2

Micrososft [31]. Our empirical results indicate that the residual network is well suited for facial recognition, and with proper data conditioning and establishing a pipeline around the network can significantly reduced the filters and number of layers. More details are presented in Section IV.A.

The training dataset was derived from three independent datasets, i.e. the LFW dataset [32] and the VGG dataset [33], with a total of 2 million images. Let us define each person in the dataset as $\mathcal{P} = \{1, 2, ..., K\}$. We are working with a training set $\{(X_i, Y_i) | X_i \in \mathbb{R}^{H \times W \times 3} Y_i \in \mathcal{P}^{H \times W}$ for all $i = 1, ..., \mathcal{I}\}$ consisting of $\mathcal{I}$ three-channel RGB images (X_i), with uniform sizes $H \times W$ along with the true labels (Y_i) for each person. Then, for the k^{th} query i.e. the k^{th} individual in the repository, we have a given subset of images $\mathcal{I}^{(k)}$ that fit the response to that query, and a set $\mathcal{F}^{(k)}$ that represents a list of faces in those images. Let the final set of descriptors for the subset of faces in $\mathcal{F}^{(i)}$ as $\mathcal{D}^{(i)}$, and individual descriptors as $d^{(ij)} \in \mathcal{D}^{(i)}, j = 1, ..., |\mathcal{D}^{(i)}|$. $|\mathcal{D}^{(i)}|$ is a subset of the set $\mathcal{D}^{(i)}$. $\mathcal{D} = \bigcup_{i=1}^{K} \mathcal{D}^{(i)}$ represents the set of descriptors from all K=7695 individuals in our dataset.

One of our primary assumptions is that the pixels of each image are independent and identically distributed that follow a categorical distribution. Using this assumption we can define a Residual CNN network that performs multinomial logistic regression. The network determines the features from the input layer and consequently provides classification score using filtering for each label. We can thus model the network as associative functions corresponding to L layers with parameters denoted by $W = [w^{(1)}, w^{(2)}, ..., w^{(L)}$, that is

$$f(x; W) = g^{(L)}(g^{(L-1)}(...g^{(2)}(g^{(1)}(x; w^{(1)}); w^{(2)})...; w^(L-1); w^(L)) \tag{1}$$

The classification score of a pixel x for a given class c is obtained from the function $f_c(x; W)$, which is the cth component of $f(x; W)$. Using the *softmax* function, we can map this score to a probability distribution:

$$p(c|x, W) = \frac{exp(f_c(x; W))}{\sum_{k=1}^{K} exp(f_k(x; W))} \tag{2}$$

For the training of the network, i.e. learning the optimal parameters W*, the cross-entropy loss is used, which minimizes the KL-divergence between the predicted and the true class distribution:

$$W* = \underset{w}{\operatorname{argmin}} \frac{1}{2} ||W||^2 - \frac{\lambda}{MHW} \sum_{i=1}^{M} \sum_{j=1}^{HW} logp(y_{ij} | x_{ij}, W), \tag{3}$$

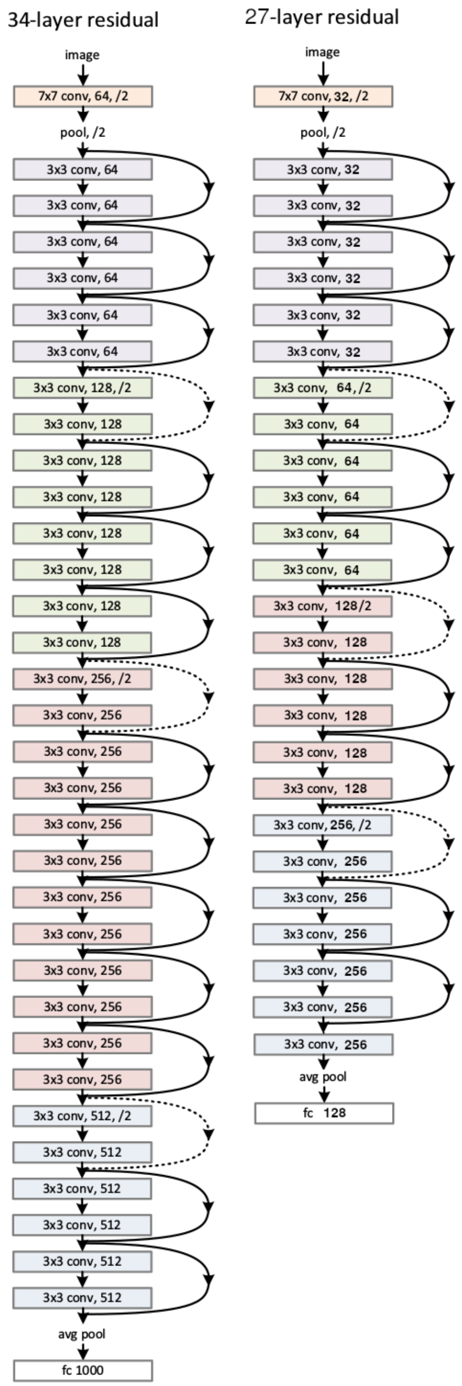

Figure 4: Network design comparison of the proposed Reduced ResNet. Left: the residual network with 34 parameter layers (3.6 billion FLOPs) as a reference. Right: the proposed network with 28 parameter layers (2.8 billion FLOPs). Table 2 shows more details and other variants.

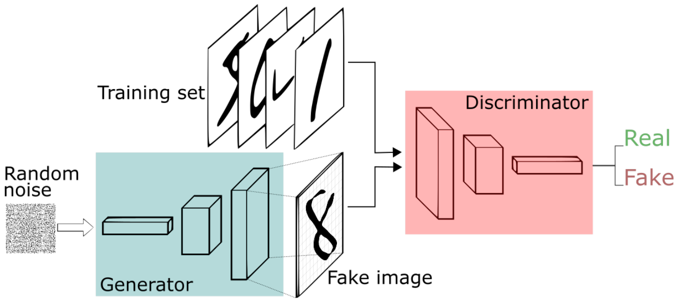

Figure 5: Basic architecture of the Generative Adversarial Networks used to generate adversarial images for facial recognition

where $x_{ij} \in \mathbb{R}^4$ stands for the jth pixel of the ith training image and $y_{ij} \in \mathcal{P}$ is its ground-truth label for an individual person. The hyper-parameter $\lambda > 0$ is chosen to apply weighting for the regularization of the parameters (i.e., L_2-norm of W. At inference, a probability distribution is predicted for each pixel via softmax normalization, defined in Equation (2), and the labeling is calculated based on the highest class probability. The final list of faces in our dataset contains a 128-bit vector converted from each face image.

A. Generative Adversarial Networks

Generative adversarial networks (GANs) are a type of artificial intelligence that makes use of the generator that synthesizes data, and the discriminator that uses the same data to determine whether the input is real. A target face can mislead the conventional classifiers by trivial non-random perturbations to only a few features of the face that otherwise looks very similar to the original face. This degrading process is caused by max-pooling [34] which applies a filter to the input volume and outputs the maximum number in every sub-region that the filter convolves around.

We use a multi-domain feature translation that takes an input image X_i and generates an output image Y_i conditioned on the target domain label $c, G(X,c) \rightarrow Y$. The discriminator produces probability distributions over both sources and domain labels, $S : X \rightarrow \{S_{src}(x, S_{cls}(X))\}$. To make the generated images indistinguishable from the real images, we adopt an adversarial loss as follows.

$$\begin{aligned}\mathcal{L}_{adv} = & \mathbb{E}_X[log S_{src}(X)] + \\ & \mathbb{E}_{X,c}[log(1 - S_{src}(G(X,c))],\end{aligned} \tag{4}$$

where G generates an image $G(X,c)$ conditioned on both the input image X_i and the target domain label c, while the discriminator S tries to distinguish between real and fake images. In this paper, we refer to the term $S_{src}(X)$ as a probability distribution over sources given by S. The generator G tries to minimize this objective, while the discriminator S tries to maximize it.

Another auxiliary classifier on top of S imposes the domain classification loss when optimizing both S and G. This loss is defined as

$$\mathcal{L}^r_{cls} = \mathbb{E}_{X,c'}[-log S_{cls}(c'|X)], \tag{5}$$

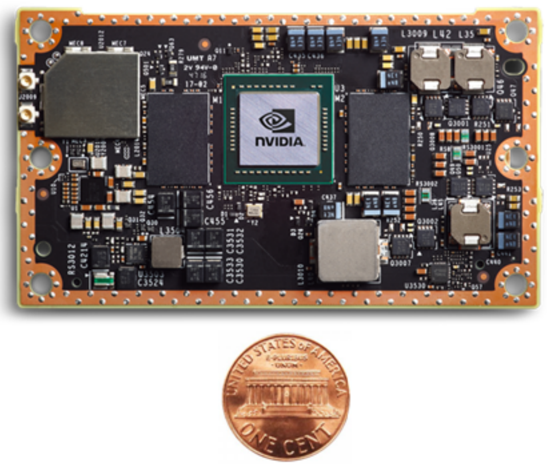

Figure 6: Portability comparison of the TX2 computing package

where the term $S_{cls}(c'|X)$ represents a probability distribution over domain labels computed by S. On the other hand, the loss function for the domain classification of fake images is defined as

$$\mathcal{L}^f_{cls} = \mathbb{E}_{X,c}[-logS_{cls}(c|G(X,c))], \tag{6}$$

In order to preserve the content of ths input images while changing only the domain-related part of the inputs, we apply a cycle consistency loss [24] to the generator

$$\mathcal{L}_{rec} = \mathbb{E}_{X,c,c'}[||X - G(G(X,c),c')||_1], \tag{7}$$

where G takes in the translated image $G(X,c)$ and the original domain label c' as input and tries to reconstruct the original image X. Finally, the objective functions to optimize G and S are written as

$$\mathcal{L}_D = -\mathcal{L}_{adv} + \lambda_{cls}\mathcal{L}^r_{cls} \tag{8}$$

$$\mathcal{L}_G = -\mathcal{L}_{adv} + \lambda_{cls}\mathcal{L}^f_{cls} + \lambda_{rec}\mathcal{L}_{rec} \tag{9}$$

where λ_{cls} and λ_{rec} are hyper-parameters for domain classification and reconstruction losses, respectively. We use $\lambda_{cls} = 1$ and $\lambda_{rec} = 10$ in all of our experiments.

B. Low Power Deep Learning

Conventional deep learning methods are computationally very demanding and require significant efforts to cut down their demands while keeping the accuracy reasonably high. We utilize the Nvidia Jetson TX2 platform for the realization because it is made for manufacturing, industrial, and retail product, ranging from commercial drones to hardware for artificially intelligent cities. The size comparison of the TX2 GPU is illustrated in Figure 6 while its basic specification are listed in Table I.

Table I: The basic components of TX2 hardware/software kit versus the previous TX1

Module	Description
GPU	Pascal
CPU	64-bit Denver 2 and A57 CPUs
Memory	8 GB 128 bit LPDDR4 58.4 GB/s
Storage	32 GB eMMC
Wi-Fi/BT	802.11 2x2 ac/BT support
Video Encode	2160p @ 60
Video Decode	2160p @ 60 12 bit support for H.265, VP9
Camera	1.4 Gpix/s up to 2.5 Gbps per lane
Mechanical	50mm x 87mm, 400-pin Compaitible Board to Board connector

The basic design pipeline is shown in Figure 3 where all processing components and device controllers are made to ubiquitously share the 128-bit wide memory bus arbitration and system memory. Since the TX2 kit consists of several components, the pipeline was designed to minimize the memory overhead by turning off all unused ports and devices. The onboard 640x480 24fps camera continuously captures streaming video that is passed to the OpenCV2 Multimedia acceleration Engine to be decoded and later buffered. The video stream is directly passed to the dual 128 CUDA core Pascal GPU where our Reduced Resnet model aligns, crops and encodes all faces in the given frame. The landmarks are converted to 128-bit vectors for each face-encoding and stored in the buffer. The dual 2GHz CPUs then fuse the adversarial data samples to the deep feature comparator and detect the suspects. The final video is encoded with the bounding boxes of the classified faces and displayed using the display head port.

IV. Experiment & Results

The proposed framework was implemented in Python using parallel and distributed (CPU+GPU) design practices specific to Nvidia TX2 SDK. Our implementation was GPU optimized and utilized CPU for specific tasks. We also performed data augmentation in the form of adversarial data generation that transfers styles of target domains while keeping the facial integrity in place. Rather than training a model to perform a fixed translation (e.g., brown-to-blond hair), which is prone to over fitting, we train our model to flexibly translate images according to the labels of the target domain. This allows our model to learn reliable features universally applicable to multiple domains of images with different facial attribute values. Our experiments with proposed framework shows that the robustness attained to adversarial data augmentation improves the recognition accuracy while keeping the overall model computationally reasonable for small and low powered devices. Obtaining real time performance of facial recognition systems is an important consideration for mobility. The following sections talk about different aspects of the experiment design and

Table II: Comparing the various depths of the Residual Networks and how they compare to the original network

layer name	output size	27-layer	34-layer	50-layer
conv1	112x112	7x7,32, stride 2	7x7,64,stride 2	
conv2x	56x56	3x3 max pool, stride 2		
		[3x3,32] x6	[3x3,64] x6	[1x1,64] [3x3,64] [1x1x256] x3
conv3x	28x28	[3x3,64] x6	[3x3,128] x8	[1x1,128] [3x3,128] [1x1x512] x4
conv4x	14x14	[3x3,128] x6	[3x3,256] x12	[1x1,256] [3x3,256] [1x1x1024] x6
conv5x	7x7	[3x3,256] x7	[3x3,512] x6	[1x1,512] [3x3,512] [1x1x2048] x3
	1x1	acg pool, 128-d fc, softmax	acg pool, 1000-d fc, softmax	
FLOPS		2.8x10^9	3.6x10^9	3.8x10^9

Table III: Nvidia TX2 various NVPModels for fine tuning of the power consumption

Mode	Mode Name	Denver 2	ARM A57	GPU Freq
0	Max-N	2@2GHz	4@2GHz	1.30 Ghz
1	Max-Q	0	4@1.2GHz	0.85 Ghz
2	Max-P Core-All	2@1.4GHz	4@1.4GHz	1.12 Ghz
3	Max-P ARM	0	4@2GHz	1.12 Ghz
4	Max-P Denver	2	0	1.12 Ghz

Table IV: Power Comparison for real-time face detection

Method	Power (Watt-Hour)
Eigen Faces [35]	231
HAAR Cascades [36]	255
Deep Face [37]	414
Open Face [38]	476
Face Net [39]	511
Red-ResNet (Proposed)	17.6

compare results with other state-of-the-art face recognition models.

A. Architecture Comparison

We evaluate our proposed 27 layer Reduced Resenet against the 34-layer and 50-layer residual networks proposed in [31]. See Table 1 for detailed architectures. The two 27-layer and 34-layer residual network as shown in Figure 4 shows similar baseline architectures, but Table II indicates almost 10 GFlops computation reduction in performance measures. We replaced the loss layer with *loss metric* and made the network considerably smaller by limiting the number of filters per layer by one-half as shown in Figure 4. Moreover, the batches were increased from 5x5 to 35x15 and the number of iterations were chosen to be 10000. The pooling layer was reduced from 1000 dimensions to 128 dimensions softmax. The Reduced Resnet network starts training with random weights and employs a cross entropy loss that assigns all the target individuals to mutually exclusive regions. This allows us to run the face recognition on a cheap low powered Nvidia Tx2 that can be mounted on CCTV cameras for real time facial threat alert systems.

B. Power Comparison

Power was compared using industry standard Watt-Hour meters with precision upto milli-Watts. The Nvidia TX2 system ships with a 19V 4.74A power supply with 90W max throughput. The device's input voltage ranges from +9V to +15V DC while the TX2 Module's power consumption is between 6.5W to 18W, dependending on the NVPModel configuration. The various power models supported by TX2 are listed in Table III. The arbitrary carrier consumption, i.e., draw of peripheral ports, were limited to 5W max. Our testbed was running on Max-N mode at full clock modes. The PC used in the comparison was running an Intel Xeon(R) CPU E5-2687W v4 @ 3.00GHz 48 core 64bit with 64GB of memory and two Titan Xp PCIe/SSE2 GPUs. Some face recognition models did not consume any GPU capability while others ran completely on GPUs thus ramping up their power requirements. The final power consumption comparison is shown in Table IV. We can observe from the comparison that the TX2 board achieve several times less power consumption due to its purpose built architecture and low powered processing units.

C. Accuracy comparison

We evaluate our proposed framework under the LFW face verification protocols where around 6000 validation face pairs are computed to tell if they are from the same person. The LFW dataset is a standard benchmark in face recognition research. The identities in our neural network training data does not overlap with the LFW identities. The LFW has 13,233 images from 5,750 people and this experiment provides 6,000 pairs broken into ten folds. The accuracy in the restricted protocol is obtained by averaging the accuracy of ten experiments. The data is separated into ten equally-sized folds and each experiment trains on nine folds and computes the accuracy on remaining testing fold. We achieve a mean accuracy of 97.52% under this protocol. Comparisons with previous works on mean accuracy shown in Table V. With the usage of DIL, we are the model has an accuracy of 99.38% on the standard Labeled Faces in the Wild benchmark. This is comparable to other state-of-the-art models and means that, given two face images, it correctly predicts if the images are of the same person 99.38% of the time.

Table V: Face Verification Accuracy on LFW Dataset

Face Recognition Technique	Model Accuracy
Eigen Faces [35]	60.19%
KNN + GPCA [40]	76.12 %
HAAR Cascades [36]	80.52%
Open Face [38]	92.92%
TL Joint Bayesian [41]	96.33%
GaussianFace [42]	95.5%
Deep Face [37]	97.35%
DeepID [43]	99.15%
Face Net [39]	99.64%
Red-ResNet (Proposed)	97.52%

V. Conclusion

CCTV cameras are being used for facial detection systems that doesn't consider adversarial cases i.e. to trick machine vision algorithms into perceiving them as something completely different from who they are. The proposed model shows how a simple inclusion of self generated adversarial images can overcome this challenge. GANs are very good at creating realistic adversarial examples, which end up being a very good way to train AI systems to develop a robust defense against suspects attempting to fool facial recognition systems. The framework uses common face style transfer elements to generate variations of the same person's face to enhance the predictions. It is challenging to obtain reasonable classification accuracy when the target is rare, since the data instances in the majority classes usually overshadow those in the minority classes. In this paper, a novel framework is proposed to enhance facial recognition. The experimental results on a multimedia big data set clearly show the effectiveness of the proposed framework and how it can successfully enhance the prediction scores of the chosen rare faces. In the future, the co-existence of multiple tree hierarchies when building them will be considered.

References

[1] M.-L. Shyu, K. Sarinnapakorn, I. Kuruppu-Appuhamilage, S.-C. Chen, L. Chang, and T. Goldring, "Handling nominal features in anomaly intrusion detection problems," in *15th International Workshop on Research Issues in Data Engineering: Stream Data Mining and Applications (RIDE-SDMA 2005)*, 2005, pp. 55–62.

[2] M.-L. Shyu, T. Quirino, Z. Xie, S.-C. Chen, and L. Chang, "Network intrusion detection through adaptive sub-eigenspace modeling in multiagent systems," *ACM Trans. Auton. Adapt. Syst.*, vol. 2, no. 3, Sep. 2007.

[3] Y. Yan, M. Chen, S. Sadiq, and M.-L. Shyu, "Efficient imbalanced multimedia concept retrieval by deep learning on spark clusters," *International Journal of Multimedia Data Engineering and Management (IJMDEM)*, vol. 8, no. 1, pp. 1–20, 2017.

[4] S.-C. Chen, S. Rubin, M.-L. Shyu, and C. Zhang, "A dynamic user concept pattern learning framework for content-based image retrieval," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 36, no. 6, pp. 772–783, Nov 2006.

[5] S. Sadiq, Y. Yan, M.-L. Shyu, S.-C. Chen, and H. Ishwaran, "Enhancing multimedia imbalanced concept detection using vimp in random forests," in *2016 IEEE 17th International Conference on Information Reuse and Integration (IRI)*, July 2016, pp. 601–608.

[6] S. Sadiq, Y. Tao, Y. Yan, and M.-L. Shyu, "Mining anomalies in medicare big data using patient rule induction method," in *Multimedia Big Data (BigMM), 2017 IEEE Third International Conference on*. IEEE, 2017, pp. 185–192.

[7] S. Sadiq, Y. Yan, A. Taylor, M.-L. Shyu, S.-C. Chen, and D. Feaster, "Aafa: Associative affinity factor analysis for bot detection and stance classification in twitter," in *2017 IEEE International Conference on Information Reuse and Integration (IRI)*. IEEE, 2017, pp. 356–365.

[8] Q. Zhu, L. Lin, M.-L. Shyu, and D. Liu, "Utilizing context information to enhance content-based image classification," *International Journal of Multimedia Data Engineering and Management*, vol. 2, no. 3, pp. 34–51, 2011.

[9] Q. Zhu and M.-L. Shyu, "Sparse linear integration of content and context modalities for semantic concept retrieval," *IEEE Transactions on Emerging Topics in Computing*, vol. 3, no. 2, pp. 152–160, June 2015.

[10] Q. Zhu, L. Lin, M.-L. Shyu, and S.-C. Chen, "Feature selection using correlation and reliability based scoring metric for video semantic detection," in *Proceedings of the Fourth IEEE International Conference on Semantic Computing*, 2010, pp. 462–469.

[11] X. Li, S.-C. Chen, M.-L. Shyu, and B. Furht, "An effective content-based visual image retrieval system," in *Proceedings of the Computer Software and Applications Conference*, 2002, pp. 914–919.

[12] X. Huang, S.-C. Chen, M.-L. Shyu, and C. Zhang, "User concept pattern discovery using relevance feedback and multiple instance learning for content-based image retrieval," in *Proceedings of the Third International Workshop on Multimedia Data Mining, in conjunction with the 8th ACM International Conference on Knowledge Discovery & Data Mining*, July 2002, pp. 100–108.

[13] X. Chen, C. Zhang, S.-C. Chen, and S. Rubin, "A human-centered multiple instance learning framework for semantic video retrieval," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 39, no. 2, pp. 228–233, 2009.

[14] M.-L. Shyu, S.-C. Chen, and C. Haruechaiyasak, "Mining user access behavior on the www," in *Systems, Man, and Cybernetics, 2001 IEEE International Conference on*, vol. 3. IEEE, 2001, pp. 1717–1722.

[15] L. Lin, M.-L. Shyu, G. Ravitz, and S.-C. Chen, "Video semantic concept detection via associative classification," in *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*. IEEE, 2009, pp. 418–421.

[16] D. Liu, Y. Yan, M.-L. Shyu, G. Zhao, and M. Chen, "Spatio-temporal analysis for human action detection and recognition in uncontrolled environments," *International Journal of Multimedia Data Engineering and Management*, vol. 6, no. 1, pp. 1–18, Jan. 2015.

[17] M.-L. Shyu, S.-C. Chen, M. Chen, C. Zhang, and K. Sarinnapakorn, "Image database retrieval utilizing affinity relationships," in *Proceedings of the 1st ACM International Workshop on Multimedia Databases*, ser. MMDB '03. New York, NY, USA: ACM, 2003, pp. 78–85.

[18] J. Fan, H. Luo, and A. K. Elmagarmid, "Concept-oriented indexing of video databases: toward semantic sensitive retrieval and browsing," *Image Processing, IEEE Transactions on*, vol. 13, no. 7, pp. 974–992, 2004.

[19] L. Lin, G. Ravitz, M.-L. Shyu, and S.-C. Chen, "Effective feature space reduction with imbalanced data for semantic concept detection," in *Proceedings of the IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing*, 2008, pp. 262–269.

[20] M. L. Shyu, Z. Xie, M. Chen, and S. C. Chen, "Video semantic event/concept detection using a subspace-based multimedia data mining framework," *IEEE Transactions on Multimedia*, vol. 10, no. 2, pp. 252–259, Feb 2008.

[21] S.-C. Chen, A. Ghafoor, and R. L. Kashyap, *Semantic Models for Multimedia Database Searching and Browsing*. Norwell, MA, USA: Kluwer Academic Publishers, 2000.

[22] X. Chen, C. Zhang, S.-C. Chen, and M. Chen, "A latent semantic indexing based method for solving multiple instance learning problem in region-based image retrieval," in *Proceedings of the 7th IEEE International Symposium on Multimedia*, Dec 2005, pp. 37–44.

[23] A. Kurakin, D. Boneh, F. Tramèr, I. Goodfellow, N. Papernot, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses," 2018.

[24] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," *arXiv preprint arXiv:1711.09020*, 2017.

[25] S. M. Moosavi Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," in *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, no. EPFL-CONF-218057, 2016.

[26] A. Odena, J. Buckman, C. Olsson, T. B. Brown, C. Olah, C. Raffel, and I. Goodfellow, "Is generator conditioning causally related to gan performance?" *arXiv preprint arXiv:1802.08768*, 2018.

[27] A. Fawzi, O. Fawzi, and P. Frossard, "Analysis of classifiers robustness to adversarial perturbations," *Machine Learning*, vol. 107, no. 3, pp. 481–508, 2018.

[28] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[29] R. Huang, B. Xu, D. Schuurmans, and C. Szepesvári, "Learning with a strong adversary," *arXiv preprint arXiv:1511.03034*, 2015.

[30] H. Kannan, A. Kurakin, and I. Goodfellow, "Adversarial logit pairing," *arXiv preprint arXiv:1803.06373*, 2018.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[32] 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 07-49, University of Massachusetts, Amherst, Tech. Rep., 2007.

[33] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conference*, 2015.

[34] Y. LeCun, Y. Bengio *et al.*, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, no. 10, p. 1995, 1995.

[35] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of cognitive neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[36] P. D. Wadkar and M. Wankhade, "Face recognition using discrete wavelet transforms," *International Journal of Advanced Engineering Technology*, vol. 3, no. 1, pp. 239–242, 2012.

[37] 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*, 2014, pp. 1701–1708.

[38] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, "Openface: A general-purpose face recognition library with mobile applications," *CMU School of Computer Science*, 2016.

[39] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[40] P. Parveen and B. Thuraisingham, "Face recognition using multiple classifiers," in *Tools with Artificial Intelligence, 2006. ICTAI'06. 18th IEEE International Conference on*. IEEE, 2006, pp. 179–186.

[41] X. Cao, D. Wipf, F. Wen, G. Duan, and J. Sun, "A practical transfer learning algorithm for face verification," in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 3208–3215.

[42] C. Lu and X. Tang, "Surpassing human-level face verification performance on lfw with gaussianface." in *AAAI*, 2015, pp. 3811–3819.

[43] Y. Sun, D. Liang, X. Wang, and X. Tang, "Deepid3: Face recognition with very deep neural networks," *arXiv preprint arXiv:1502.00873*, 2015.