This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2908802, IEEE Transactions on Image Processing

Enhance Visual Recognition under Adverse Conditions via Deep Networks

Ding Liu, *Student Member, IEEE,* Bowen Cheng, Zhangyang Wang, *Member, IEEE,* Haichao Zhang, *Member, IEEE,* and Thomas S. Huang, *Life Fellow, IEEE*

***Abstract*—Visual recognition under adverse conditions is a very important and challenging problem of high practical value, due to the ubiquitous existence of quality distortions during image acquisition, transmission, or storage. While deep neural networks have been extensively exploited in the techniques of low-quality image restoration and high-quality image recognition tasks respectively, few studies have been done on the important problem of recognition from very low-quality images. This paper proposes a deep learning based framework for improving the performance of image and video recognition models under adverse conditions, using robust adverse pre-training or its aggressive variant. The robust adverse pre-training algorithms leverage the power of pre-training and generalizes conventional unsupervised pre-training and data augmentation methods. We further develop a transfer learning approach to cope with real-world datasets of unknown adverse conditions. The proposed framework is comprehensively evaluated on a number of image and video recognition benchmarks, and obtains significant performance improvements under various single or mixed adverse conditions. Our visualization and analysis further add to the explainability of results.**

***Index Terms*—deep learning, neural network, image recognition.**

I. Introduction

While the visual recognition research has made tremendous progress in recent years, most models are trained, applied, and evaluated on high-quality (HQ) visual data, such as the LFW [1] and ImageNet [2] benchmarks. However, in many emerging applications such as autonomous driving, intelligent video surveillance and robotics, the performances of visual sensing and analytics can be seriously endangered by different adverse conditions [3] in complex unconstrained scenarios, such as limited resolution, noise, occlusion and motion blur. For example, video surveillance systems have to rely on cameras of limited definitions, due to the prohibitive costs of installing high-definition cameras everywhere, leading to the practical need to recognize faces reliably from very low-resolution images [4]. Other quality factors, such as occlusion and motion blur, are also known as critical concerns for commercial face recognition systems. As similar problems are ubiquitous for recognition tasks in the wild, it becomes highly desirable to investigate and improve the robustness of visual recognition systems to low-quality (LQ) image data.

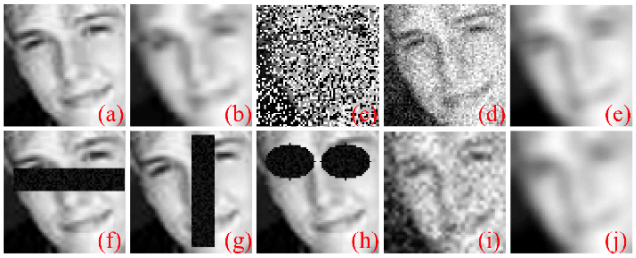


Figure 1. The original high-quality image from the MSRA-CFW dataset in (a), and (b) - (j) list various low-quality images generated from (a), that are all correctly recognized by our proposed models: (b) downsampled by a factor of 4; (c) 50% salt & pepper noise; (d) Gaussian noise (std $= 25$); (e) Gaussian blur (std $= 5$); (f)-(h) random synthetic occlusions; (i) downsampled by 4 followed by adding Gaussian noise (std $= 25$); (j) downsampled by 4 followed by adding Gaussian blur (std $= 5$).

Unfortunately, exiting studies demonstrate that most state-of-the-art models appear fragile when applied on low-quality data. The literature [5], [6] has confirmed the significant effects of quality factors such as low-resolution, contrast, brightness, sharpness, focus, and illumination on commercial face recognition systems. The recent work [7] revealed that common degradations can even dramatically lower face recognition accuracy of the latest deep learning based face recognition models [2], [8], [9]. In particular, blur, noise, and occlusion cause the most significant performance deterioration. Besides face recognition, the low-quality data is also found to adversely affect other recognition applications, such as handwritten digit recognition [10] and style recognition [11].

This paper targets this important but less explored problem of visual recognition under adverse conditions. We study how and to what extent such adverse visual conditions can be coped with, aiming to improve the robustness of visual recognition systems on low-quality data. We carry out a comprehensive study on improving deep learning models for both image and video recognition tasks. We generalize conventional unsupervised pre-training and data augmentation methods, and propose the *robust adverse pre-training* algorithms. The algorithms are generally applicable to various adverse conditions, and are jointly optimized with the target task. Figure 1 (b)-(j) depict a series of heavily corrupted, low-quality images. They are all correctly recognized by our proposed models, though challenging even for human to recognize.

The first two authors contributed equally to this work.

This work was supported in part by US Army Research Office grant W911NF-15-1-0317.

D. Liu, B. Cheng and T. S. Huang are with the Department of Electrical and Computer Engineering and Beckman Institute, Univerisity of Illinois at Urbana-Champaign, Urbana, IL, 61801 USA e-mail: (dingliu2@illinois.edu; bcheng9@illinois.edu; t-huang1@illinois.edu).

Z. Wang is with the Department of Computer Science and Engineering, Texas A&M University, TX 77843 USA (e-mail: atlaswang@tamu.edu).

H. Zhang is with Baidu Research, Sunnyvale, CA 94089 USA (e-mail: hczhang1@gmail.com).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2908802, IEEE Transactions on Image Processing

The major technical innovations are summarized in three aspects:

- We present a framework for visual recognition under adverse conditions, that improves deep learning based models via robust pre-training and its aggressive variant. The framework is extensively evaluated on various datasets, settings and tasks. Our visualization and analysis further add to the explainability of results.
- We extend the framework to video recognition, and discuss how the temporal fusion strategy should be adjusted under different adverse conditions.
- We develop a transfer learning approach for real-world datasets of unknown adverse conditions without synthetic LQ-HQ pairs directly available. We empirically demonstrates that our approach also improves the recognition on the original benchmark dataset.

In the following, we will first review related work in Section II. Our proposed robust adverse pre-training algorithm and its variant, as well as the corresponding image based experiments are introduced in Section III. Video based experiments are reported with implementation details in Section IV. The transfer learning approach for dealing with real-world datasets is described in Section V. Finally, conclusions and discussions are provided in Section VI.

II. Related Work

A. Visual Recognition under Adverse Conditions

In a real-world visual recognition problem, there is indeed no absolute boundary between LQ and HQ images. Yet as commonly observed, while some mild degradations may have negligible impact on the recognition performance, the impact turns much notable once the level of adverse conditions passes some empirical threshold. The object and scene recognition literature reports a significant performance drop when the image resolution is decreased below 32×32 pixels [12]. In [4], the authors find the face recognition performance to be notably deteriorated when face regions become smaller than 16×16 pixels. [7] reports a rapid decline of face recognition accuracies, with Gaussian noise of standard deviation (std) between 10 and 20. [5], [6] reveal more impacts of contrast, brightness, sharpness, and out-of-focus on image based face recognition.

To resolve that, the conventional approach first resorts to image restoration and then feeds the restored image into a classifier [13], [14], [15]. Such a straightforward approach yields the sub-optimal performance: the artifacts introduced by the reconstruction process will undermine the final recognition. [4], [16] incorporate class-specific features in the restoration as a prior. A domain adaptation scheme of pre-training is utilized for sentiment classification [17] and font recognition [18], where a sub-model is first trained in a unsupervised manner from both LQ and HQ data, aiming to minimize the low-level feature difference between them. [19] presents a joint image restoration and recognition method, based on the assumption that the degraded images, if correctly restored, also have a good identifiability. A similar approach is adopted for jointly dealing with image dehazing and object detection in [20]. Those "close-the-loop" ideas achieve superior performance over the traditional two-stage pipelines.

Compared to single image object recognition, the impact of adverse conditions on video recognition is as profound and significant, with many attentions paid to tasks such as video face recognition and tracking [21], license plate recognition [22], and facial expression recognition [23]. [24] introduces robust hand-crafted features to low-resolution and head motion blur. [25] combines a shape-illumination manifold framework with implicit super-resolution. [26] adopts a residual neural network trained with synthetic LQ samples, which are generated by a controlled corruption process such as adding motion blur or compression artifacts.

B. Deep Networks under Adverse Conditions

Convolutional neural networks (CNNs) have gained explosive popularity in recent years for visual recognition tasks [2], [27]. However, their robustness to adverse conditions remain unsatisfactory [7]. Deep networks are shown to be susceptible to adversarial samples [28], generated by introducing carefully chosen perturbations to the input. Besides that, the common adverse conditions, stemming from artifacts during image acquisition, transmission, or storage, still easily mislead deep networks in practice [29]. [7] confirms the fragility of the state-of-the-art deep face recognition models [2], [8], [9], to various adverse conditions, in particular blur, noise, and periocular region occlusion. Besides face recognition, the adverse conditions are also found to negatively affect other recognition tasks, such as hand-written digit recognition [10] and style recognition [11].

While data augmentation has become a standard tool [2], the primary goal is to artificially increase the training data volume and improve the model generalization. The augmentation methods are moderate in practice, by adding small noise or pixel translations, etc. The learned model is then to be applied on clean HQ images for testing. Those methods are thus not dedicated to handling specific types of severe degradation. Unsupervised pre-training [30], [31] also effectively regularizes the training process, especially when labeled data is insufficient. Classical pre-training methods reconstruct the input data from itself [31] or its slightly transformed versions [30], [32] from auto-encoders. However, these methods are focused more on improving model generalization on standard (HQ) data, but not the LQ data that suffers a certain type of severe degradation. The recent work [11] describes an approach of pre-training a deep network model for image recognition under the low-resolution case. However, it neither considers any other type of adverse conditions or mixed degradations,[1] nor takes into account any video based problem setting. Most crucially, [11] required pairs of synthetic training samples before and after degradation. While the degradation process is unknown in real-world data, the applicability of the proposed algorithm is severely limited.

[1]The solutions to low-resolution cases cannot be straightforwardly extended to other adverse conditions. For example, we tried Model III of [11] in salt & pepper noise and occlusion cases, finding the performance to be hurt sometimes.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2908802, IEEE Transactions on Image Processing

III. Image Based Visual Recognition under Single or Mixed Adverse Conditions

A. Problem Statement

We start by introducing single image based visual recognition models in this section, and extend to the video recognition models later. We define the visual recognition model $\mathcal{M}$ that predicts the category labels $\{l_i\}_{i=1}^N$ from the images $\{\mathbf{y}_i\}_{i=1}^N$. Due to the adverse conditions, $\{\mathbf{y}_i\}_{i=1}^N$ can be viewed as low-quality (LQ) images, degraded from high-quality (HQ) ground truth images $\{\mathbf{x}_i\}_{i=1}^N$. For now, we treat the original training datasets as HQ images $\{\mathbf{x}_i\}$, and generate LQ images $\{\mathbf{y}_i\}$ using synthetic degradation. In testing, our model operates with only LQ inputs.

We define a CNN based image recognition model $\mathcal{M}$ with d layers. The first d_1 layers are convolutional, while the remaining $d - d_1$ layers are fully connected. The i-th convolutional layer, denoted as $conv_i$ ($i = 1, \cdots, d_1$), contains n_i filters of size $c_i \times c_i$, with default stride size 1 and zero-padding. The j-th fully connected (fc) layer, denoted as fc_j ($j = 1, \cdots, d - d_1$), has m_j nodes. We use ReLU activation and apply dropout with a rate of 0.5 to fully connected layers. Cross-entropy loss is adopted for classification, while mean square error (MSE) is used for reconstruction.

B. Robust Adverse Pre-training of Sub-models

Building a classifier $\mathcal{M}$ directly on $\{\mathbf{y}_i\}$ is usually not robust due to the severe information loss caused by adverse conditions. Training $\mathcal{M}$ over $\{\{\mathbf{x}_i\}, \{l_i\}\}$ also does not perform well when tested on $\{\mathbf{y}_i\}$ due to the domain mismatch [11], [4]. Our main intuition is to regularize and enhance the feature extraction from $\{\mathbf{y}_i\}$, via injecting auxiliary information from $\{\mathbf{x}_i\}$. With the help of $\{\mathbf{x}_i\}$, the model better discriminates the true signal from the severe corruption, and learns more robust filters from low-quality inputs. The entire $\mathcal{M}$ can be well adapted for the mapping from $\{\mathbf{y}_i\}$ to $\{l_i\}$ by a joint optimization step followed.

To pre-train $\mathcal{M}$, we first define the sub-model $\mathcal{M}_s$ with k layers. Its first k_p layers are configured the same as the first k_p layers from $\mathcal{M}$. The last $k - k_p$ layers reconstruct the input image from the output feature maps of the k_p-th layer. We generate $\{\mathbf{y}_i\}$ from $\{\mathbf{x}_i\}$, based on a degradation process parameterized by the *adverse factor* α[2], in order to meet the adverse conditions in testing. We then train $\mathcal{M}_s$ to reconstruct $\{\mathbf{x}_i\}$ from $\{\mathbf{y}_i\}$. We empirically find that pre-training only a part of convolutional layers (i.e., $k_p \leq d_1$) maintains a good balance between the feature extraction and the discrimination ability, with the best performance. After $\mathcal{M}_s$ is trained, its first k_p layers are exported to initialize the first k_p layers of $\mathcal{M}$. $\mathcal{M}$ is then jointly tuned for the recognition task over $\{\mathbf{y}_i, \{l_i\}$. The algorithm, termed as *Robust Adverse Pre-training* (**RAP**), is outlined in Algorithm 1.

[2]Here the *adverse factor* is defined in a broad sense. It can be the downsampling factor for low-resolution, the proportion of image for noise corruption, the degree of blur and so on.

Algorithm 1 Robust adverse pre-training

Input: Configuration of $\mathcal{M}$; $\{\mathbf{x}_i\}$ and $\{l_i\}$, $i = 1, 2, ..., N$; the choice of k; the adverse factor α.

1: Generate $\{\mathbf{y}_i\}$ from $\{\mathbf{x}_i\}$, based on a degradation process parameterized by α
2: Construct the k-layer sub-model $\mathcal{M}_s$. Its first k_p layers are configured identically to those of $\mathcal{M}$.
3: Train $\mathcal{M}_s$ to reconstruct $\{\mathbf{x}_i\}$ from $\{\mathbf{y}_i\}$, under MSE.
4: Export the first k_p layers from $\mathcal{M}_s$ to initialize the first k_p layers of $\mathcal{M}$, where $k_p < k$.
5: Tune $\mathcal{M}$ over $\{\{\mathbf{y}_i\}, \{l_i\}\}$, under the cross-entropy loss.

Output: $\mathcal{M}$.

C. Aggressively Robust Adverse Pre-training

Different from testing when only LQ data is available, we have the flexibility to synthesize LQ images for training at our will. While the RAP algorithm trains $\mathcal{M}$ and $\mathcal{M}_s$ under the same adverse condition, we continue to explore when the $\mathcal{M}_s$ pre-training and $\mathcal{M}$ joint-tuning are performed under different levels of adverse conditions. This is motivated by the denoising autoencoders [33], where the pre-training was conducted by noisy data and the subsequent classification model was learned with clean data. Our conjecture is that pre-training $\mathcal{M}_s$ in severer degradation can actually help $\mathcal{M}_s$ capture more robust feature mappings. This leads to the *Aggressively Robust Adverse Pre-training* (**ARAP**), a variant of RAP, outlined in Algorithm 2. We assume the degradation process of $\{\mathbf{y}_i\}$ to be identical to the target testing data, while $\{\mathbf{z}_i\}$ is a more heavily degraded set independently generated from $\{\mathbf{x}_i\}$. The larger adverse factor indicates the severer degradation, and thus in this case the adverse factor β for generating $\{\mathbf{z}_i\}$ is larger than α for $\{\mathbf{y}_i\}$. RAP can be a special case of ARAP where α and β coincide.

Algorithm 2 Aggressively robust adverse pre-training

Input: Configuration of $\mathcal{M}$; $\{\mathbf{x}_i\}$ and $\{l_i\}$, $i = 1, ..., N$; the choice of k; two adverse factors α and β ($\beta > \alpha$).

1: Generate $\{\mathbf{y}_i\}$, $\{\mathbf{z}_i\}$ from $\{\mathbf{x}_i\}$, based on two degradation processes parameterized by α and β, respectively.
2: Construct the sub-model $\mathcal{M}_s$ same as in Algorithm 1.
3: Train $\mathcal{M}_s$ to reconstruct $\{\mathbf{x}_i\}$ from $\{\mathbf{z}_i\}$, under MSE.
4: Export the first k_p layers from $\mathcal{M}_s$ to initialize the first k_p layers of $\mathcal{M}$, where $k_p < k$.
5: Tune $\mathcal{M}$ over $\{\{\mathbf{y}_i\}, \{l_i\}\}$, under the cross-entropy loss.

Output: $\mathcal{M}$.

D. Experiments on Benchmarks

1) Object Recognition on the CIFAR-10 Dataset: In order to validate our algorithm, we first conduct object recognition on the CIFAR-10 dataset [34], which consists of 60,000 color images of 32×32 pixels from 10 classes (we convert all to grayscale ones). Each class has 5,000 training images and 1,000 test images. We generate LQ images as per each specific type of adverse conditions, where the adverse factors α or β become concrete degradation hyper-parameters such as

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2908802, IEEE Transactions on Image Processing

Table I
THE TOP-1 AND TOP-5 CLASSIFICATION ACCURACY (%) ON THE CIFAR-10 DATASET, WHERE LQ IMAGES ARE GENERATED BY DOWNSAMPLING THE ORIGINAL IMAGES WITH A FACTOR OF α = 2.

	HQ	LQ-2	RAP-2-non-joint	RAP-2	ARAP-2-4	ARAP-2-8	ARAP-2-12	ARAP-2-16
Top-1	67.43	60.79	46.89	62.12	62.80	**63.31**	62.91	62.56
Top-5	96.61	95.32	90.77	95.10	95.52	**95.80**	95.34	95.10

Table II
THE TOP-1 AND TOP-5 CLASSIFICATION ACCURACY (%) ON THE CIFAR-10 DATASET, WHERE LQ IMAGES ARE GENERATED BY ADDING α = 50% SALT & PEPPER NOISE.

	HQ	LQ-50%	RAP-50%-non-joint	RAP-50%
Top-1	67.43	33.46	38.64	**50.32**
Top-5	96.61	83.22	86.86	**92.03**

Table III
THE TOP-1 AND TOP-5 FACE IDENTIFICATION ACCURACY (%) ON THE MSRA-CFW DATASET, WHERE LQ IMAGES ARE GENERATED BY ADDING α = 50% SALT & PEPPER NOISE.

# of pre-trained layers	0	1	2	3	4
Top-1	14.75	27.04	**49.86**	42.17	38.64
Top-5	36.28	45.36	**72.14**	63.09	59.85

downsampling factor, noise level, or blur kernel. We perform no other data augmentation beyond generating LQ images.

We choose $\mathcal{M}$ with $d = 4$, with $d_1 = 3$ convolutional layers, followed by $d - d_1 = 1$ fully connected layer with m_1 always equaling the number of classes. Unless otherwise stated, we set $\mathcal{M}_s$ as a fully convolutional network with the empirical values $k = 3, k_p = 2$, which work well in all experiments. The default configuration of convolutional layers are: n_1 = 64, c_1 = 9; n_2 = 32, c_2 = 5; n_3 = 20, c_3 = 5. We first train $\mathcal{M}_s$ with learning rate 0.0001, and then jointly tune $\mathcal{M}$ with a learning rate 0.001 for the first k_p layers and 0.01 for the rest $d - k_p$ layers. Both learning rates are reduced by a factor of 10 every 5,000 iterations.

a) Low-Resolution: We generate LQ (low-resolution) images $\{\mathbf{y}_i\}$ by following the process in [35], [36]: first downsampling the HQ (high-resolution) images $\{\mathbf{x}_i\}$ by a factor of α, then upsampling back to the original size with bicubic interpolation. We use the same process for all the following experiments of low-resolution degradation, unless otherwise stated. We compare the following approaches:

- **HQ:** $\mathcal{M}$ is trained and tested on $\{\{\mathbf{x}_i\}, \{l_i\}\}$.
- **LQ-α:** $\mathcal{M}$ is trained and tested on $\{\{\mathbf{y}_i\}, \{l_i\}\}$.
- **RAP-α-non-joint:** $\mathcal{M}_s$ is pre-trained using the Step 3 of Algorithm 1 on $\{\{\mathbf{y}_i\}, \{\mathbf{x}_i\}\}$. The remaining $d - k_p$ layers of $\mathcal{M}$ are then trained on $\{\{\mathbf{y}_i\}, \{l_i\}\}$, with the first k_p pre-trained layers fixed. It is identical to RAP except for no jointly tuning $\mathcal{M}$.
- **RAP-α:** $\mathcal{M}$ is trained using RAP (Algorithm 1).
- **ARAP-α-β:** $\mathcal{M}$ is trained using ARAP (Algorithm 2), where β is a larger downsamping factor than α.

The evaluation of $\mathcal{M}$s is all performed on the testing set of LQ images (except for the HQ baseline), downsampled by the factor α. The first two baselines aim to examine how much the adverse condition affects the performance.

Table I displays the results at $\alpha = 2$, which is a challenging problem of recognizing objects from images of 16×16 pixels. Such an adverse condition dramatically affects the performance, by dropping the top-1 accuracy for nearly 7%, after comparing LQ-2 with HQ. It might be unexpected that the performance of RAP-2-non-joint is much inferior to that of LQ-2. As observed in this and many following experiments, the reconstruction based pre-training step, if not jointly optimized for the recognition step, often hurts the performance rather than does any help. By adding the joint tuning step, RAP-2 gains a 1.33% advantage over LQ-2 in the top-1 accuracy, which is owning to the $\mathcal{M}_s$ pre-training that involves auxiliary yet beneficial information from HQ data.

It is noteworthy that all four ARAP methods ($\beta = 4, 8, 12, 16$) show superior results over RAP-2. ARAP-2-8 achieves the best accuracy of 63.31% (top-1) and 95.80% (top-5). The observation confirms our conjecture that more robust feature extractions could be achieved by purposely pre-training $\mathcal{M}_s$ in severer degradation ($\beta > \alpha$). As β grows with α fixed at 2, the performance of ARAP first improves and then drops, with the peak at $\beta = 8$. That is also explainable, since if $\{\mathbf{z}_i\}$ are too much degraded, little information is left for training $\mathcal{M}_s$.

b) Noise: Since adding moderate Gaussian noise has been standard for data augmentation, we focus on the more destructive salt & pepper noise. The LQ images $\{\mathbf{y}_i\}$ are generated by randomly choosing $\alpha = 50\%$ pixels in each HQ image $\mathbf{x}_i$ to be replaced with either 0 or 255. We compare HQ, LQ-α, RAP-α-non-joint, and RAP-α, all of which are similarly defined as in the low-resolution case. We tried RAP-α-β, but did not get much performance improvement over RAP-α as we did for low-resolution. In Table II, the severe information loss by 50% salt & pepper noise is reflected on the 34% top-1 accuracy drop from HQ to LQ-50%. After only pre-training the first few layers, there is a 5.18% increase in the top-1 accuracy, obtained by RAP-50%-non-joint. RAP-50% achieves the closest accuracy to the HQ baseline, and outperforms RAP-50%-non-joint by 11.68% and 5.17%, in terms of top-1 and top-5 accuracy, respectively. Those results re-confirm the necessity of both per-training and end-to-end tuning for RAP.

To determine the number of pre-trained layers in $\mathcal{M}$, we conduct an ablation study by changing the number of pre-trained layers in the model under the adverse condition of salt & pepper noise. The top-1 and top-5 face identification

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2908802, IEEE Transactions on Image Processing

Table IV
THE TOP-1 AND TOP-5 CLASSIFICATION ACCURACY (%) ON THE CIFAR-10 DATASET, WHERE LQ IMAGES ARE GENERATED BY BLURRING ORIGINAL IMAGES (HQ), WITH GAUSSIAN KERNEL OF STD α = 2.

	HQ	LQ-2	RAP-2-non-joint	RAP-2	RAP-2-5	RAP-2-8	RAP-2-9
Top-1	67.43	52.62	39.80	54.73	54.77	**55.67**	54.35
Top-5	96.61	92.70	87.34	93.24	93.50	**93.52**	93.15

Table V
THE TOP-1 AND TOP-5 FACE IDENTIFICATION ACCURACY (%) ON THE MSRA-CFW DATASET, WHERE LQ IMAGES ARE GENERATED BY DOWNSAMPLING ORIGINAL IMAGES BY A FACTOR OF α = 4.

	HQ	LQ-4	CDP	RAP-4-non-joint	RAP-4	ARAP-4-6
Top-1	57.25	50.79	53.61	50.50	**54.23**	54.10
Top-5	76.89	72.81	73.82	72.88	74.06	**74.97**

Table VI
THE TOP-1 AND TOP-5 FACE IDENTIFICATION ACCURACY (%) ON THE MSRA-CFW DATASET, WHERE LQ IMAGES ARE GENERATED BY ADDING α = 50% SALT & PEPPER NOISE.

	HQ	LQ-50%	CDP	RAP-50%-non-joint	RAP-50%
Top-1	57.25	14.75	17.54	26.20	**49.86**
Top-5	76.89	36.28	40.38	51.59	**72.14**

Table VII
THE TOP-1 AND TOP-5 FACE IDENTIFICATION ACCURACY (%) ON THE MSRA-CFW DATASET, WHERE LQ IMAGES ARE GENERATED BY BLURRING THE ORIGINAL IMAGES (HQ), WITH GAUSSIAN KERNEL OF STD α = 5.

	HQ	LQ-5	CDP	RAP-5-non-joint	RAP-5	ARAP-5-8
Top-1	57.25	49.96	51.04	45.66	**52.19**	51.94
Top-5	76.89	72.51	72.46	69.08	73.73	**73.88**

Table VIII
THE TOP-1 AND TOP-5 ACCURACY (%) ON MSRA-CFW, WHERE LQ IMAGES ARE GENERATED WITH RANDOM SYNTHETIC OCCLUSIONS.

	HQ	LQ-α	RAP-α-non-joint	RAP-α
Top-1	59.41	32.62	34.91	**43.96**
Top-5	78.11	56.32	60.16	**67.20**

accuracy (%) on the MSRA-CFW dataset is shown in Table III. It is observed that when there are very few layers pre-trained, the model extracts very limited information from the LQ-HQ reconstruction. In contrast, when too many layers are pre-trained, the model is very biased to the reconstruction task and less suitable to the recognition task. We choose pre-training the first two layers in our experiments, which achieves the highest top-1 and top-5 accuracy.

c) Blur: Images commonly suffer from various types of blurs, such as simple Gaussian blur, motion blur, out-of-focus blur, or their complex combinations [19]. We focus on the Gaussian blur, while similar strategies can be naturally extended to other types. The LQ images $\{\mathbf{y}_i\}$ are generated by convolving the HR images $\{\mathbf{x}_i\}$ with a Gaussian kernel with std $\alpha = 2$, and the fixed kernel size of 9×9 pixels. We compare HQ, LQ-α, RAP-α-non-joint, RAP-α, and ARAP-α-β (β denotes a larger std than α), all similarly defined.

Table IV demonstrates similar findings as the low-resolution case. The non-adapted restoration in RAP-α-non-joint only leaves it worse than LQ-α. RAP-α gains 1.21% over LQ-α in top-1 accuracy. Two out of three ARAP methods ($\beta = 5, 8$) yield greatly improved results than RAP-α, while $\beta = 9$ is only marginally inferior. Using Algorithm 2, $\mathcal{M}_s$ trained with heavier blurs tends to produce more discriminative features, when applied to LQ data with lighter blurs, which benefits recognition tasks.

2) Face Identification on the MSRA-CFW Dataset: We conduct face identification on the MSRA Dataset of Celebrity Faces on the Web (MSRA-CFW) [37], which includes $202,792$ cropped and centered face images of 64×64 pixels in around 1600 classes. We select a subset including all the 123 classes of more than 300 images, to ensure the sufficient amount of training data for our deep network model. We split 90% images of each class for training and 10% for testing. We perform the face identification task, under highly challenging adverse conditions, such as very low resolution, noise, blur, occlusion and mixed cases. The visual examples are displayed in Figure 1.

For the low resolution, noise or blur case, we set $\mathcal{M}$ with $d = 8$, $d_1 = 6$. The convolutional layers are configured as: $n_1 = 32$, $c_1 = 9$; $n_2 = 16$, $c_2 = 5$; $n_3 = 20$, $c_3 = 4$; $n_4 = 40$, $c_4 = 3$; $n_5 = 60$, $c_5 = 3$; $n_6 = 80$, $c_6 = 2$. fc_1 has $m_1 = 160$ and fc_2 has $m_2 = 123$. For occlusion, we modify $n_1 = 16$, $c_1 = 21$; $n_2 = 8$, $c_2 = 1$, and leave other six layers unchanged. Here the low-level filters perform in-painting, and thus needs larger receptive fields to predict missing pixels from neighborhoods.

a) Low-Resolution, Noise, Blur: The three adverse conditions follow similar settings and comparison methods to CIFAR-10. To validate the effectiveness of RAP and ARAP, we utilize for pre-training the domain adaptation scheme in [17], [18], termed Cross-Domain Pre-training (CDP), as an additional baseline method for comparison. We adopt a larger downsampling factor 4 in the low-resolution case, and a larger blur std 5 for the blur case. The conclusions drawn from Tables V, VI and VII are also consistent to those of CIFAR-10: RAP boosts performance in all cases compared to LQ, CDP, and RAP-non-joint, and ARAP achieves considerably higher results for the two cases of low-resolution and blur.

b) Occlusion: Prior studies in [7] discovered that the periocular occlusion degraded the face recognition performance most. We follow [7] to synthesize the occlusions for the periocular regions, in the shape of either rectangle or ellipse (chosen with equal probability). The size of either shape, as well as the pixel values within the synthetic occlusion,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2908802, IEEE Transactions on Image Processing

Table IX
THE TOP-1 AND TOP-5 ACCURACY (%) ON MSRA-CFW, WHERE LQ IMAGES ARE GENERATED BY FIRST DOWNSAMPLING ORIGINAL IMAGES BY α = 2 AND THEN ADDING GAUSSIAN NOISE WITH STD 25.

	HQ	LQ-2	RAP-2-non-joint	RAP-2	ARAP-2-4
Top-1	57.25	45.57	44.30	48.63	**50.34**
Top-5	76.89	69.82	68.00	71.89	**73.76**

Table X
THE TOP-1 AND TOP-5 ACCURACY (%) ON MSRA-CFW, WHERE LQ IMAGES ARE GENERATED BY FIRST DOWNSAMPLING ORIGINAL IMAGES BY α = 4 AND THEN BLURRED WITH THE GAUSSIAN KERNEL OF STD 2.

	HQ	LQ-4	RAP-4-non-joint	RAP-4	ARAP-4-8
Top-1	57.25	49.39	48.76	52.30	**52.68**
Top-5	76.89	71.29	70.99	73.80	**74.51**

is drawn from uniform distributions. The center locations of synthetic occlusions are picked randomly in a bounding box, whose boundaries are determined by eye landmark points. We emphasize that the occlusion masks are unknown and changing for both training and testing, corresponding to the toughest blind inpainting problem [38].

We evaluate HQ, LQ-α, RAP-α-non-joint and RAP-α in Table VIII. The parameter α generally denotes the controlled shape/size/location variations. We also tried large β via enlarging the maximal size of occlusions, but observed no visible improvement from ARAP-α-β. The occlusion causes much worse corruptions than previous adverse conditions: it completely masks a facial region that is known to be critical for recognition. The lost pixel information is harder to be restored than the salt & pepper noise case, due to the missing neighborhood. As expected, the challenging random occlusions result in very significant drops from HQ to LQ. RAP-non-joint only marginally raises the accuracy (e.g., 2% in top-1). RAP achieves the most encouraging improvements of 11.34% and 10.88%, in terms of top-1 and top-5 accuracy, respectively.

c) Mixed Adverse Conditions: In real-world applications, multiple types of degradation may appear simultaneously. To this end, we examine if our algorithms remain effective under a mixture of multiple adverse conditions. We evaluate two settings: 1) first downsampling HQ images by α = 2 and then adding Gaussian noise with std 25; 2) first downsampling HQ images by α = 4 and then blurring with the Gaussian kernel of std 2. We compare HQ, LQ-α, RAP-α-non-joint, RAP-α and ARAP-α-β, where α and β both only consider the downsampling factor for simplicity. ARAP and RAP seamlessly generalize to the mixed adverse conditions, and obtain the most promising performance in Tables IX and X

3) Digit Recognition on the SVHN Dataset: The Street View House Number (SVHN) dataset [39] contains $73,257$ digit images of 32×32 pixels for training, and $26,032$ for testing. We focus on investigating the impact of low-resolution and blur on the SVNH digit recognition. Our model has a default configuration of $d = 4$, $d_1 = 2$; $n_1 = 20$, $c_1 = 5$; $n_2 = 50$, $c_2 = 5$; $m_1 = 500$; $m_2 = 10$ (class number used). $conv_1$ is followed by 2×2 max pooling.

Figure 2. Digit image samples from the SVHN dataset.

a) Low-Resolution: Table XII compares HQ, LQ-α, RAP-α-non-joint, RAP-α and ARAP-α-β, in the low-resolution case with α = 8. While the LQ-α accuracy drops disastrously, satisfactory top-1 and top-5 accuracy is achieved by ARAP-α-β (β = 16) and RAP-α. We observe that more than half of digit images could still be correctly predicted at the extremely low-resolution of 4×4 pixels by the proposed methods.

b) Blur: Table XI compares those methods in the Gaussian blur case with standard deviation α = 2. To our astonishments, ARAP-α-β not only improves over LQ-α, but also surpasses the performance of HQ in terms of top-1 accuracy. That is because the original SVNH images (treated as HQ) are real-world photos that unavoidably suffer from certain blur, which can be found in Figure 2. Convolved with the synthetic Gaussian blur kernel (α = 2), the actual blur kernel's standard deviation becomes larger than 2. Hence ARAP-α-β is potentially able to remove the inherent blurs in HQ images, besides the synthetically added blurs.

4) Image Classification on the ImageNet Dataset: We validate our algorithm on a large-scale dataset, ImageNet dataset [40], for image classification of 1,000 classes. We utilize 1.2 million images of ILSVRC2012 training set for training, and 50,000 images of its validation set for testing. We study the degradation of low-resolution on the ImageNet image classification. In our experiment, we customize a popular classification model: *VGG-16* [41] to work on color images directly. Specifically, we add three convolutional layers to the beginning of VGG-16, in order to increase the model capacity for handling the low-resolution degradation. We choose $k = 3, k_p = 3$ for $\mathcal{M}_s$ and the configuration of the first three convolutional layers is n_1 = 64, c_1 = 9; n_2 = 32, c_2 = 1;

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2908802, IEEE Transactions on Image Processing

Table XI
THE TOP-1 AND TOP-5 DIGIT RECOGNITION ACCURACY (%) ON THE SVHN DATASET, WHERE LQ IMAGES ARE GENERATED BY BLURRING THE ORIGINAL IMAGES (HQ), WITH THE GAUSSIAN KERNEL OF STANDARD DEVIATION $\alpha = 2$.

	HQ	LQ-2	RAP-2-non-joint	RAP-2	ARAP-2-5	ARAP-2-8
Top-1	89.23	85.40	83.84	82.47	**89.40**	88.29
Top-5	98.57	97.55	96.92	96.82	**98.32**	98.09

Table XII
THE TOP-1 AND TOP-5 DIGIT RECOGNITION ACCURACY (%) ON THE SVHN DATASET, WHERE LQ IMAGES ARE DOWNSAMPLING THE ORIGINAL IMAGES (HQ) BY A FACTOR OF $\alpha = 8$.

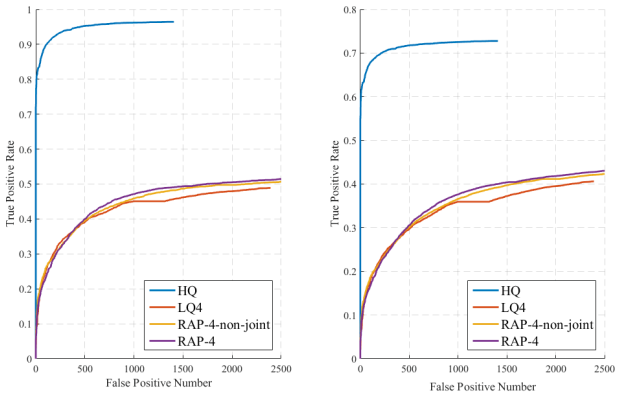
	HQ	LQ-8	RAP-8-non-joint	RAP-8	ARAP-8-16
Top-1	89.23	19.60	45.98	51.00	**51.17**
Top-5	98.57	65.44	87.08	**89.15**	89.06

$n_3 = 3$, $c_3 = 5$. The rest architecture is the same as VGG-16. We use the VGG-16 model released by its authors as the initialization of it, in order to boost the convergence rate. We follow the conventional protocols in [41] for data pre-processing, including image resizing, random cropping and mean removal of each color channel.

Table XIII compares HQ, LQ-α, RAP-α-non-joint and RAP-α, in the low-resolution case with $\alpha = 4$ and 8. RAP-4 outperforms LQ-4 and RAP-4-non-joint in terms of both top-1 and top-5 accuracy. When the low-resolution degradation becomes severe, RAP-8 is superior to LQ-8 and RAP-8-non-joint by a larger margin. Specifically, RAP-8 beats LQ-8 by 0.55% in top-1 accuracy and 0.77% in top-5 accuracy, and beats RAP-8-non-joint by 1.85% in top-1 accuracy and 1.72% in top-5 accuracy, respectively.

5) Face Detection on the FDDB Dataset: We further generalize our proposed algorithm to the face detection task. We use the training images of the WIDER Face dataset [42] as our training set, which consists of 12,880 images and the annotations of 159,424 faces. and adopt the Face Detection Data Set and Benchmark (FDDB) [43] as our test set, which contains the annotations for 5,171 faces in a set of 2,845 images. We study the degradation of low-resolution for the face detection task. In our experiment, we customize a popular detection model: *Faster R-CNN* [44] to work on color images directly. Similar to Section III-D4, we add three convolutional layers to the beginning of Faster R-CNN, in order to increase the model capacity for handling the low-resolution degradation. We choose $k = 3, k_p = 3$ for $\mathcal{M}_s$ and the configuration of the first three convolutional layers is $n_1 = 64$, $c_1 = 9$; $n_2 = 32$, $c_2 = 1$; $n_3 = 3$, $c_3 = 5$. The rest architecture is the same as Faster R-CNN. We use the VGG-16 model in [41] released by its authors as initialization, in order to accelerate the convergence speed.

Figure 3 shows the discrete and continuous ROC curves of HQ, LQ-α, RAP-α-non-joint and RAP-α, in the low-resolution case with $\alpha = 4$. We can observe that there is an obvious performance drop due to the low-resolution degradation. RAP-4 outperforms LQ-4 and RAP-4-non-joint in terms of recall rate with the same number of false positives. For example, RAP-4 recalls 50.49% faces with 2,000 false positives, which is 0.73% higher than RAP-4-non-joint and 2.55% higher than LQ-4, respectively. We obtain the same comparison result in the case of 1,500 false positives, where RAP-4 recalls 48.68% faces, being 0.67% higher than RAP-4-non-joint and 3.15% higher than LQ-4, respectively.



(a) (b)

Figure 3. (a) Discrete ROC curve and (b) Continuous ROC curve on FDDB dataset, where LQ images are downsampled by a factor of $\alpha = 4$.

E. Analysis and Visualization

1) Cause of Failure of Deep Models When Using LQ Data: The failures caused by LQ inputs to deep models can vary a lot, both by degradation type and by the subsequent high-level vision tasks, and sometimes being model-specific too. We pick several examples below to explain this complicacy (all below refer to ordinary deep models without robust pre-training, and our proposed approach alleviates those problems in general):

- Low-resolution recognition: the learned filter responses tend to be weak (lack of sharp, recognizable patterns), and inter-channel feature maps tend to be less variable, resulting in poorer discriminativeness of features.
- Recognition with salt & pepper noise: the learned feature maps contain moderately perturbed patterns, and those perturbed patterns can dominate over true features when the noise level gets high (e.g., when 50% pixels are corrupted) and thus confuse classification.
- Low-resolution object detection: when the two-stage detectors are used, low-resolution causes fewer object proposals to be generated in the first region proposal network stage. As a proof-of-concept, we tried to generate proposals on high-resolution images, then applied those

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2908802, IEEE Transactions on Image Processing

Table XIII
THE TOP-1 AND TOP-5 CLASSIFICATION ACCURACY (%) ON THE IMAGENET VALIDATION SET, WHERE LQ IMAGES ARE DOWNSAMPLED BY A FACTOR OF α = 4 OR 8.

	HQ	LQ-4	RAP-4-non-joint	RAP-4	LQ-8	RAP-8-non-joint	RAP-8
Top-1	71.46	61.92	61.16	62.03	46.67	45.37	47.22
Top-5	90.62	84.13	83.65	84.35	71.55	70.60	72.32

bounding box locations to low-resolution images to train classifiers, finding the performance loss to be restored to a large extent. The same observations were found in hazy images [20].

2) Convolutional and Additive Adverse Conditions: We have tested four adverse conditions so far. RAP and ARAP improves the recognition in all cases, which shows that the pre-training of image restoration achieves feature enhancement in the recognition model and benefits the visual recognition task. We note that low-resolution and blur clearly receive extra bonus from ARAP than RAP. In the other two cases, i.e., noise and occlusion, RAP and ARAP perform approximately the same. Such contrastive behaviors hint that some adverse conditions might be more suitable for ARAP to perform than the others.

In the general image degradation model, the observed image Y is usually represented as

$$Y = \mathcal{F} * X + e, \tag{1}$$

where $\mathcal{F}$ denotes the point spread function, X is the clean image, and e is the noise. Low-resolution and blur are usually modeled in $\mathcal{F}$ as low-pass filters, while noise and occlusion can be incorporated in e as additive perturbations. We term the former category as **convolutional adverse conditions**, and the latter as **additive adverse conditions**. We conjecture that the additive adverse condition causes pixel-wise corruptions but still retains some structural information, while the convolutional adverse condition results in global detail loss and smoothening, which may be more challenging for recognition and thus needs more robust feature extractions by purposely pre-training $\mathcal{M}_s$ in heavier adverse conditions. This hypothesis will be further justified experimentally when we extend our framework to video cases.

3) Effects of End-to-End Tuning in RAP: To further analyze our proposed RAP, we focus on the following two questions: How the joint tuning of $\mathcal{M}$ modifies the features learned in the pre-trained $\mathcal{M}_s$, and why it improves the recognition in almost all adverse conditions?

To answer these questions, we visualize and compare the features in the first k_p-th layers of $\mathcal{M}$ before and after the end-to-end tuning, denoted as $\mathcal{F}_k$ and $\mathcal{F}'_k$, respectively. Recall that in the pre-training step of RAP, $\mathcal{M}_s$ reconstructs the images by feeding $\mathcal{F}_k$ to $k-k_p$ additional layers, that are removed in the joint tuning step. We pass both $\mathcal{F}_k$ and $\mathcal{F}'_k$ through the fixed mapping of these $k-k_p$ layers (obtained when training $\mathcal{M}_s$). The output, which is of the same dimension as HQ images, is used to visualize of $\mathcal{F}_k$ or $\mathcal{F}'_k$. Note that the visualizations of $\mathcal{F}_k$ are just the reconstruction results of $\mathcal{M}_s$.

Figure 4. Visualized features for successful examples of joint tuning, i.e. those correctly classified by RAP but misclassified by RAP-non-joint. Column (a): original HQ images from MSRA-CFW. (b): LQ images from the first mixed adverse condition setting. (c): visualized $\mathcal{F}_k$ (intermediate features by RAP-non-joint). (d): visualized $\mathcal{F}'_k$ (intermediate features by RAP).

Figure 4 presents feature visualizations for five MSRA-CFW images that are correctly classified by RAP but misclassified by RAP-non-joint. As shown in column (c), the $\mathcal{F}_k$ features from the un-tuned $\mathcal{M}_s$ are heavily over-smoothed, with much discriminative information lost. In contrast, the visualizations of $\mathcal{F}'_k$ yield a few impressive restoration results in column (d). The joint tuning step enables the closed-loop consideration of two information sources (HQ data and labels) for two related tasks (restoration and recognition). It thus boosts not only the recognition accuracy, but also the restoration: column (d) results contain much richer and finer details, and are apparently more recognizable than column (c).

IV. VIDEO RECOGNITION IN ADVERSE CONDITIONS

A. Temporal Fusion for Video Based Models

Temporal fusion of feature representations is usually adopted in deep learning based methods for video-related

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2908802, IEEE Transactions on Image Processing

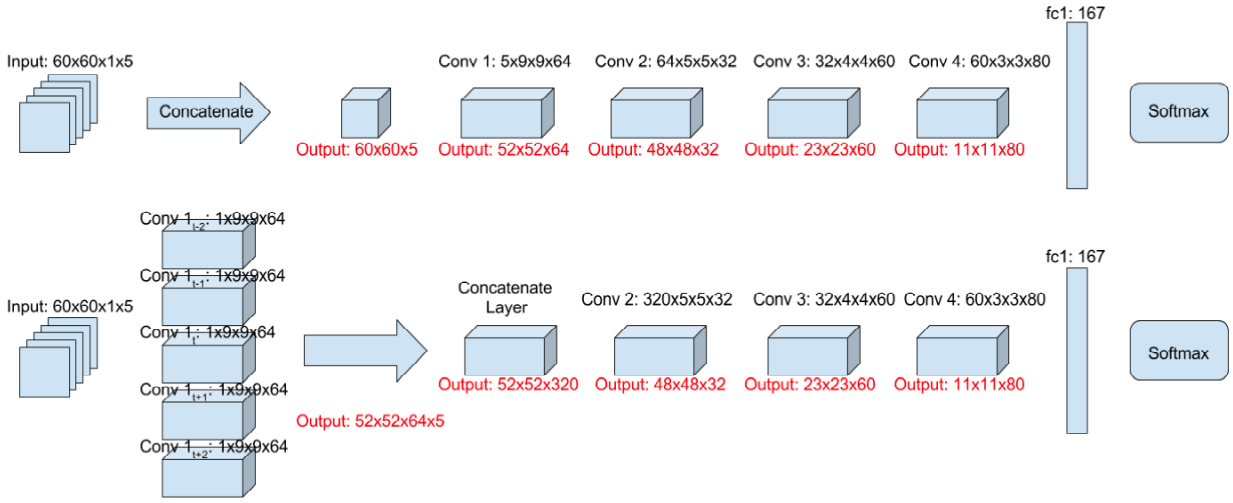


Figure 5. Model architectures for YTF video recognition experiments. Top: early fusion. Bottom: slow fusion.

tasks. Karpathy et al. [27] first provided an extensive empirical evaluation of CNNs on large-scale video classification. In addition to the single frame baseline, [27] discussed three connectivity patterns. The **early fusion** combines frames within a time window immediately in the pixel level. The **late fusion** separately extracts features from each frame and does not merge them until the first fully connected layer. The **slow fusion** is a balanced mix between the two, which slowly unifies temporal information throughout the network by progressively merging features from individual frames.

B. Robust Adverse Pre-training for Video Recognition

Following [27], [45], [46], we treat each video as a number of short, fixed-sized clips. Each clip is set to contain $2T+1$ contiguous frames in time. The video based CNN model $\mathcal{M}_v$ takes a clip as its input. To extend $\mathcal{M}_v$ to adverse conditions, we first pre-train a single image model $\mathcal{M}$ using RAP or ARAP, by treating all frames as individual images and formulating an image based recognition problem. We then convert $\mathcal{M}$ to $\mathcal{M}_v$ based on different fusion strategies, and initialize the weights of $\mathcal{M}_v$ from $\mathcal{M}$ using the weight transfer proposed in [45]. $\mathcal{M}_v$ is then tuned in the video setting. Since we find the late fusion results to be always inferior to the other two, we omit discussing the case of late fusion hereinafter.

For early fusion, we copy the $conv_1$ layer of $\mathcal{M}$ (n_1 filters of $c_1 \times c_1$) for $2T+1$ times, and divide the weights of all filters by $2T+1$[3]. We then use them in the new $conv_1$ layer of $\mathcal{M}_v$ with the size $n_1 \times c_1 \times c_1 \times (2T+1)$, to fuse information in the first layer. All other layers of $\mathcal{M}_v$ are identical with $\mathcal{M}$ in both configuration and weight transfer.

For slow fusion, we copy the $conv_1$ layer of $\mathcal{M}$ for $2T+1$ times into the new $conv_1$ layer of $n_1 \times c_1 \times c_1 \times (2T+1)$, without changing the weights. We then stack the filters of the $conv_2$ layer of $\mathcal{M}$ (n_2 filters of $c_c \times c_c$) for $2T+1$ times and divide all weights by $2T+1$, constituting the new $conv_2$ layer of $n_2 \times c_2 \times c_2 \times (2T+1)$ to fuse information in the second layer. All other layers of $\mathcal{M}_v$ remain identical to $\mathcal{M}$.

[3]Detailed reasoning follows Section III.C of [45]. Our early and slow fusion models resemble their architectures (a) and (b).

C. Experiments on Benchmarks

We use a video face dataset: the *YouTube Face* (YTF) benchmark [47] to validate our algorithm. We choose the 167 subject classes that contain 4 video sequences. For each class, we randomly pick one video for testing and the rest for training. The face regions are cropped using the given bounding boxes. As the majority of cropped faces have side lengths between 56 and 68, we slightly resize them all to 60×60 for simplicity, and refer to those as the *original YTF set* hereinafter. We densely sample clips of 5 ($T = 2$) frames from each video with a stride of one frame, and present each clip individually to the model. The class predictions are averaged to produce an estimate of the video-level class probabilities. For the single image model, we chose $d = 5$, $d_1 = 4$, with each layer: n_1 = 64, c_1 = 9; n_2 = 32, c_2 = 5; n_3 = 60, c_3 = 4; n_4 = 80, c_4 = 3, m_1 = 167. All video based models start from the same pre-trained single frame model, and then split filters differently. We enforce filter symmetry as in [45]. The detailed architectures are drawn in Figure 5.

Similarly to image based experiments, Tables XIV and XV compare HQ, LQ-α, RAP-α, and ARAP-α-β, in the settings of low resolution (α = 2) and salt & pepper noise (α = 50%). ARAP/RAP bring substantially improved performance within each fusion. Recall that the best fusion models in [27] displayed only modest improvement over single frame models (from 59.3% to 60.9%), we consider that our 1.11% top-5 gain by early fusion in the low resolution setting, and 13.37% top-5 gain by slow fusion in the noise setting, are both reasonably good.

While [27] advocated slow fusion for normal visual recognition problems, the situations seem more complicated when adverse conditions step in. Our results imply that additive adverse conditions favor slow fusion, while convolutional adverse conditions prefer early fusion. We tried experiments in the blur case, whose observations are close to the low

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2908802, IEEE Transactions on Image Processing

Table XIV
THE TOP-1 AND TOP-5 ACCURACY (%) ON YTF, IN THE LOW RESOLUTION SETTING, WITH DIFFERENT FUSION STRATEGIES.

		HQ	LQ-2	RAP-2	ARAP-2-4	ARAP-2-8
Single Frame	Top-1	37.32	38.30	39.16	41.05	38.58
	Top-5	60.01	59.56	59.94	61.97	60.33
Early Fusion	Top-1	38.11	37.73	39.83	**41.11**	38.05
	Top-5	58.48	62.42	62.74	**63.85**	60.79
Slow Fusion	Top-1	35.99	37.76	39.60	40.98	39.67
	Top-5	53.20	58.79	60.86	63.03	61.50

Table XV
THE TOP-1 AND TOP-5 ACCURACY (%) ON YTF, IN THE SALT & PEPPER NOISE SETTING, WITH DIFFERENT FUSION STRATEGIES.

		HQ	LQ-50%	RAP-50%
Single Frame	Top-1	37.32	15.81	31.64
	Top-5	60.01	30.93	48.48
Early Fusion	Top-1	38.11	18.86	21.20
	Top-5	58.48	36.59	38.01
Slow Fusion	Top-1	35.99	21.97	**34.55**
	Top-5	53.20	39.00	**52.37**

resolution case. We conjecture that early fusion becomes the preferred option when the data is already heavily "filtered" by degradation operators or blur kernels, such that it cannot afford extra information loss after more filtering. The diverse fusion preferences manifest the unique complication brought by adverse conditions.

As the last finding, in the low resolution case, the RAP and ARAP results using LQ data can even surpass HQ results notably. We input the original YTF set to the trained ARAP-2-4 models, and also witnesses much improved accuracy in Table XVI, than feeding the same set through the HQ models. The best top-1 and top-5 results in Table XVI also surpass all results in Table XIV. We suspect that although the original YTF set is treated as clean and high-quality, it was actually contaminated by degradations during image collection, and is thus low-quality from that viewpoint. Applying RAP and ARAP compensates part of the unknown information loss. From another perspective, training a model on LQ data and then applying on HQ data is related to a special data augmentation introduced in [11], that blends HQ and LQ data for training. While [11] confirmed its effectiveness in recognizing LQ subjects, we discover its usefulness for normal (HQ) visual recognition too.

Table XVI
THE TOP-1 AND TOP-5 ACCURACY (%) BY FEEDING THE ORIGINAL YTF SET TO THE TRAINED ARAP-2-4 MODELS.

	Single Frame	Early Fusion	Slow Fusion
Top-1	41.31	41.60	**42.20**
Top-5	62.30	**64.04**	63.10

V. COPING WITH UNKNOWN ADVERSE CONDITIONS: A TRANSFER LEARNING APPROACH

In all previous experiments, we train with $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^N$ pairs. That is equivalent to assuming a pre-known degradation process from $\{\mathbf{x}_i\}_{i=1}^N$ to $\{\mathbf{y}_i\}_{i=1}^N$. Such an assumption, as made in [11], is impractical for real-world LQ data and restricts our experiments to synthesized test data so far. In this section, we develop a transfer learning approach to significantly relax this strong assumption. It ensures the wide applicability of our algorithms, even when the degradation parameters cannot be accurately inferred.

For convolutional adverse conditions, the recognition accuracy is usually peaked at some optimal $\beta^* > \alpha$. The additive adverse conditions seems insensitive to β. However, the performance ARAP-α-β results ($\beta > \alpha$) are observed to be always better, or at least comparable to ARAP-α, even when β deviates far away from β^*.

Algorithm 3 Transfer ARAP Learning

Input: Configurations of $\mathcal{M}$ and $\mathcal{M}'$; the choice of k; the clean source dataset $\{\mathbf{x}_i\}$ and $\{l_i\}$; the target dataset $\{\mathbf{x}'_i\}$ and $\{l'_i\}$, with unknown α'.

1: Decide the major degradation type in $\{\mathbf{x}'_j\}$, and choose β' such that it overestimates α'.
2: Generate $\{\mathbf{y}_i\}$ from $\{\mathbf{x}_i\}$, based on the degradation processes of the major type, parameterized by β'.
3: Perform Steps 3 - 6 in Algorithm 1, to train $\mathcal{M}$ on the source dataset.
4: Export the first k layers from $\mathcal{M}$ to initialize the first k layers of $\mathcal{M}'$.
5: Tune $\mathcal{M}'$ over $\{\{\mathbf{x}'_i\}, \{l'_i\}\}$.

Output: $\mathcal{M}'$.

On a target dataset with real-world corruptions, it is reasonable to assume that the major type of adverse condition(s) can still be identified, but the parameter α' of the underlying degradation process cannot be accurately estimated. Observing the robustness of ARAP w.r.t. β, we propose the *Transfer ARAP Learning* (**T-ARAP**) approach, as detailed in Algorithm 3. The core idea is to first choose β' that we empirically believe $\beta' > \alpha'$, then performing RAP (with β') to train $\mathcal{M}$ on a source dataset. Next, we transfer the learned sub-model of $\mathcal{M}$ to initialize $\mathcal{M}'$, which is later tuned for the target dataset. Note that β' is not necessarily very close to α'. In practice,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2908802, IEEE Transactions on Image Processing

Table XVII
THE TOP-1 AND TOP-5 ACCURACY (%) ON THE ORIGINAL YTF SET, BY TRANSFERRING FROM THE MSRA-CFW RAP-4 MODEL.

	LQ_d	LQ_p	T-ARAP-non-joint	T-ARAP
Top-1	32.65	32.35	33.67	**34.77**
Top-5	45.37	47.73	48.11	**53.11**

one may safely start with some large β', and scan backwards for an optimal value.

We validate the approach via conducting the following experiment: improving face identification on (original) YTF by referring to a RAP model on MSRA-CFW. For simplicity, here we perform the task of single-image face identification, and treat the original YTF set as an image collection without utilizing temporal coherence. We visually observe that the original YTF images have inherently lower quality, which is also supported by Table XVI. We select low resolution as our target adverse condition, and not too aggressively, choose $\beta' = 4$. We hence take the $\mathcal{M}_s$ part from the RAP-4 model trained on MSRA-CFW, to initialize the first 2 layers of $\mathcal{M}'$. Meanwhile, we design three baselines for comparison: 1) $\mathbf{LQ}_d$ model trained directly end-to-end on YTF; 2) $\mathbf{LQ}_p$ model trained on YTF with classical unsupervised layer-wise pre-training; 3) **T-ARAP-non-joint**, taking the untuned $\mathcal{M}_s$ of RAP-4 for $\mathcal{M}'$ initialization. In Table XVII, T-ARAP improves the top-5 recognition accuracy by nearly 8% over the naive $\mathbf{LQ}_d$, with no strong prior knowledge about the degradation process nor its parameter, which demonstrates the effectiveness of our proposed transfer learning approach.

VI. CONCLUSION AND DISCUSSIONS

This paper systematically improves deep learning models via robust pre-training for image and video recognition under adverse conditions. We thoroughly evaluate our proposed algorithm on various datasets and degradation settings, and analyze our results in depth, which shows the effectiveness of our proposed algorithm. A transfer learning approach is also proposed to enhance the real-world applicability.

REFERENCES

[1] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," in *Workshop on Faces in Real-Life Images: detection, alignment, and recognition*, 2008.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.

[3] M. De Marsico, *Face recognition in adverse conditions*. IGI Global, 2014.

[4] W. W. Zou and P. C. Yuen, "Very low resolution face recognition problem," *IEEE TIP*, 2012.

[5] A. Dutta, R. Veldhuis, and L. Spreeuwers, "The impact of image quality on the performance of face recognition," *Technical Report, Centre for Telematics and Information Technology, University of Twente*, 2012.

[6] A. Abaza, M. A. Harrison, T. Bourlai, and A. Ross, "Design and evaluation of photometric image quality measures for effective face recognition," *IET Biometrics*, vol. 3, no. 4, pp. 314–324, 2014.

[7] S. Karahan, M. K. Yildirum, K. Kirtac, F. S. Rende, G. Butun, and H. K. Ekenel, "How image degradations affect deep cnn-based face recognition?" in *Biometrics Special Interest Group (BIOSIG), 2016 International Conference of the*. IEEE, 2016, pp. 1–5.

[8] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *BMVC*, 2015.

[9] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[10] S. Basu, M. Karki, S. Ganguly, R. DiBiano, S. Mukhopadhyay, and R. Nemani, "Learning sparse feature representations using probabilistic quadtrees and deep belief nets," in *ESANN*, 2015.

[11] Z. Wang, S. Chang, Y. Yang, D. Liu, and T. S. Huang, "Studying very low resolution recognition using deep networks," in *CVPR*. IEEE, 2016.

[12] A. Torralba, R. Fergus, and W. T. Freeman, "80 million tiny images: A large data set for nonparametric object and scene recognition," *TPAMI*, 2008.

[13] R. Fergus, B. Singh, A. Hertzmann, S. T. Roweis, and W. T. Freeman, "Removing camera shake from a single photograph," in *ACM Transactions on Graphics (TOG)*, vol. 25, no. 3. ACM, 2006, pp. 787–794.

[14] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE TIP*, 2010.

[15] D. Liu, Z. Wang, Y. Fan, X. Liu, Z. Wang, S. Chang, and T. Huang, "Robust video super-resolution with learned temporal dynamics," in *ICCV*, 2017, pp. 2507–2515.

[16] P. H. Hennings-Yeomans, S. Baker, and B. V. Kumar, "Simultaneous super-resolution and feature extraction for recognition of low-resolution faces," in *CVPR*. IEEE, 2008.

[17] X. Glorot, A. Bordes, and Y. Bengio, "Domain adaptation for large-scale sentiment classification: A deep learning approach," in *ICML*, 2011, pp. 513–520.

[18] Z. Wang, J. Yang, H. Jin, E. Shechtman, A. Agarwala, J. Brandt, and T. S. Huang, "Deepfont: Identify your font from an image," in *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 2015, pp. 451–459.

[19] H. Zhang, J. Yang, Y. Zhang, N. M. Nasrabadi, and T. S. Huang, "Close the loop: Joint blind image restoration and recognition with sparse representation prior," in *ICCV*. IEEE, 2011, pp. 770–777.

[20] B. Li, X. Peng, Z. Wang, J. Xu, and D. Feng, "Aod-net: All-in-one dehazing network," in *ICCV*, 2017.

[21] L. Stasiak, A. Pacut, and R. Vincente-Garcia, "Face tracking and recognition in low quality video sequences with the use of particle filtering," in *International Carnahan Conference on Security Technology*. IEEE, 2009, pp. 126–133.

[22] C.-C. Chen and J.-W. Hsieh, "License plate recognition from low-quality videos." in *MVA*, 2007, pp. 122–125.

[23] Y.-l. Tian, "Evaluation of face resolution for expression analysis," in *CVPR Workshop*. IEEE, 2004, pp. 82–82.

[24] C. Shan, S. Gong, and P. W. McOwan, "Recognizing facial expressions at low resolution," in *IEEE Conference on Advanced Video and Signal Based Surveillance*, 2005, pp. 330–335.

[25] O. Arandjelovic and R. Cipolla, "A manifold approach to face recognition from low quality video across illumination and pose using implicit super-resolution," in *ICCV*. IEEE, 2007, pp. 1–8.

[26] C. Herrmann, D. Willersinn, and J. Beyerer, "Low-quality video face recognition with deep networks and polygonal chain distance," in *International Conference on Digital Image Computing: Techniques and Applications (DICTA)*. IEEE, 2016, pp. 1–7.

[27] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *IEEE CVPR*, 2014, pp. 1725–1732.

[28] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[29] S. Dodge and L. Karam, "Understanding how image quality affects deep neural networks," *arXiv preprint arXiv:1604.04004*, 2016.

[30] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *ICML*. ACM, 2008, pp. 1096–1103.

[31] D. Erhan, P.-A. Manzagol, Y. Bengio, S. Bengio, and P. Vincent, "The difficulty of training deep architectures and the effect of unsupervised pre-training," in *AISTATS*, 2009.

[32] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber, "Stacked convolutional auto-encoders for hierarchical feature extraction," *Artificial Neural Networks and Machine Learning–ICANN 2011*, pp. 52–59, 2011.

[33] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *Journal of Machine Learning Research*, vol. 11, no. Dec, pp. 3371–3408, 2010.

[34] A. Krizhevsky, "Learning multiple layers of features from tiny images," 2009.

[35] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *European Conference on Computer Vision*. Springer, 2014, pp. 184–199.
[36] D. Liu, Z. Wang, B. Wen, J. Yang, W. Han, and T. S. Huang, "Robust single image super-resolution via deep networks with sparse prior," *IEEE Transactions on Image Processing*, vol. 25, no. 7, pp. 3194–3207, 2016.
[37] X. Zhang, L. Zhang, X.-J. Wang, and H.-Y. Shum, "Finding celebrities in billions of web images," *IEEE Transactions on Multimedia*, 2012.
[38] J. Xie, L. Xu, and E. Chen, "Image denoising and inpainting with deep neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 341–349.
[39] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading digits in natural images with unsupervised feature learning," in *NIPS workshop on deep learning and unsupervised feature learning*, vol. 2011, no. 2, 2011, p. 5.
[40] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.
[41] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[42] S. Yang, P. Luo, C. C. Loy, and X. Tang, "Wider face: A face detection benchmark," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
[43] V. Jain and E. Learned-Miller, "Fddb: A benchmark for face detection in unconstrained settings," University of Massachusetts, Amherst, Tech. Rep. UM-CS-2010-009, 2010.
[44] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.
[45] A. Kappeler, S. Yoo, Q. Dai, and A. K. Katsaggelos, "Video super-resolution with convolutional neural networks," *IEEE Transactions on Computational Imaging*, vol. 2, no. 2, pp. 109–122, 2016.
[46] D. Liu, Z. Wang, Y. Fan, X. Liu, Z. Wang, S. Chang, X. Wang, and T. S. Huang, "Learning temporal dynamics for video super-resolution: A deep learning approach," *TIP*, vol. 27, no. 7, pp. 3432–3445, 2018.
[47] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 529–534.