IEEE Signal Processing Society | IEEE Transactions on Image Processing

# Learning Bayesian Image Set Representations with Deep Neural Networks for Automatic Object and Face Classification

| | |
|---|---|
| Journal: | *Transactions on Image Processing* |
| Manuscript ID | TIP-28234-2022 |
| Manuscript Type: | Regular Paper |
| Date Submitted by the Author: | 03-Aug-2022 |
| Complete List of Authors: | Mirnateghi, Nima; Edith Cowan University, School of Science<br>Shah, Syed Afaq Ali ; Edith Cowan University |
| Subject Category<br>Please select at least one subject category that best reflects the scope of your manuscript: | Image & Video Processing Techniques, Image & Video Sensing, Modeling, and Representation |
| EDICS: | 13. TEC-MLI Machine Learning for Image Processing < Image & Video Processing Techniques |
| | |

# Learning Bayesian Image Set Representations with Deep Neural Networks for Automatic Object and Face Classification

Nima Mirnateghi, Syed Afaq Ali Shah

***Abstract*—The pervasiveness of human computer interaction in the modern world has escalated the collection of images and videos, which, at the same time, demands robust image/video recognition methods. While there is a huge number of studies on single-shot image based recognition, the image set classification has shown to be a rather versatile approach for object recognition and similar tasks from a number of studies. However, to date, existing image set classification methods have mainly focused on deterministic models, which often fail to capture the uncertain regions of neural networks. To address these limitations, we propose a deep Bayesian learning framework for image set-based object recognition and video-based facial classification. We empirically demonstrate a superior performance of the proposed technique on multiple benchmark object recognition datasets, including CIFAR-100, CIFAR-10, MNIST, ETH-80, a subset of ImageNet dataset, and on two video-based facial recognition datasets including CMU MoBo, YouTube Celebrity and USCD/Honda. Our detailed experimental analysis reveals that our proposed method achieves state-of-the-art performance on most datasets.**

***Index Terms*—Image Set Classification, Deep Bayesian Learning, Object Recognition, Face Classification**

## I. Introduction

With the availability of low-cost portable cameras e.g., on mobile devices, the acquisition of videos and images has become a trivial task. Millions of images are captured every day and new photo albums of individuals are emerging on daily basis [1]–[3]. In addition, close loop surveillance and security cameras capture hundred thousands of image frames thus further adding to this data acquisition process. These images and videos are extensively used for computer vision applications such as object/face detection, tracking and recognition [4]–[8]. Despite the availability of such huge data, recognition is traditionally performed by using a single shot of an object or face to make the decision about its class [3]. While a convenient way, the chances of misclassification in the traditional single-shot based approaches are very high [9]–[11].

Image set classification has received significant attention from the research community [3], [11]–[14]. It is defined as the classification from multiple images where each image set contains images of the same object class captured from different viewpoints, illuminations, occlusions, and backgrounds. Unlike traditional methods that recognise individual images as a vector of pixel values assigned to a class, image set classification views the task of classification as a set of vectors as an input [15].

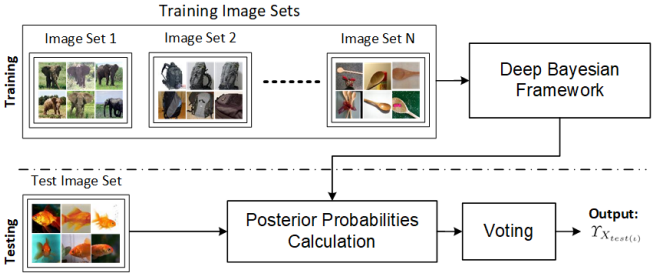



Fig. 1. Block Diagram of our Proposed Deep Bayesian Image Set Classification Method.

When compared to single-shot approaches, image sets inherit rich information about appearance variations, such as non-rigid deformations, different viewing angles, and camera pose changes. This information is particularly useful in challenging real-world scenarios such as complex dynamic scenes. These characteristics of image set classification offer promising performances in real-life applications e.g., biometrics, surveillance, object recognition, video-based face recognition, and person re-identification in a network of security cameras [11]. Image set classification can mitigate the limitations of recognition tasks from single images arising from scene-complexity. Nevertheless, large intra-class variability [16], and inter-class ambiguity [12] resulting from appearance variations may pose a significant challenge to accurately classify image sets.

With respect to the intrinsic properties of image set classification, data uncertainty among image sequences arguably arises from two factors, i.e., class overlaps and the presence of noise [17]. It is, therefore, indispensable to capture model uncertainty in such applications. Recent studies have shown that Bayesian neural networks can offer a better way to capture the uncertain regions of the model [18], [19], while arriving at more reliable decisions about classification predictions [20]. The ability of Deep Bayesian neural networks to minimise the Kullback-Leibler (KL) divergence from data distributions motivates the need to apply image set classification to address these issues.

This paper proposes a deep Bayesian-learning framework for image set classification (Figure I). The proposed framework is based on Monte Carlo Dropout as Bayesian approx-

N. Mirnateghi and S. Shah are with School of Science and the Centre for AI and Machine Learning, Edith Cowan University, Perth, Australia.
Manuscript received April 19, 2021; revised August 16, 2021.

imation for automatic object and face recognition [19]. The contributions of the paper are summarised as follows:

- We propose a Deep Bayesian Image Set classification approach. To the best of our knowledge, this is the first image set based deep Bayesian learning framework.
- We demonstrate the effectiveness of three different voting strategies in our proposed image set framework.
- We extensively evaluate our proposed technique on five different publicly available object datasets and two face datasets.

The rest of the paper is organised as follows. In Section II, we provide an overview of the existing image set classification techniques. The proposed method is described in Section III. Experimental results are reported in Section IV. The paper is concluded in Section VI.

## II. Related Work

Image set classification techniques can be classified into three broad categories: parametric, non-parametric, and deep learning-based methods.

### A. Parametric Methods

Parametric methods [21] view the task of image set classification from a statistical perspective. These models measure the similarities and differences between the gallery and test image set samples. For instance, Manifold Density Divergence (MDD) [21] uses probability density distributions. The density functions between the image sets are often estimated using Kullback-Leibler divergence (KL-Divergence) and in some cases Monte Carlo (MC) algorithm. A general assumption in parametric methods is to have a certain distribution (i.e. Gaussian distribution) for the population [22]. However, due to the distributional hypothesis, the major shortcoming of such techniques is their generalisation ability in the real-world applications in case there is a weak statistical relationship between the training and test image sets.

### B. Non-parametric Methods

To address the drawbacks of parametric methods, non-parametric models are proposed. These models use linear and non-linear representations [23]–[28]. Non-parametric models characterise image sets either on a geometric surface or by its exemplar images. A broad range of distance metrics have been developed in the literature to calculate the set-to-set distance depending on the type of representation. For example, in terms of exemplar representations for image sets, Wang et al. [29] compute the mean of image sets using Euclidean distance as a metric for between-set similarity. On the contrary, projection kernel metric [30], geodesic distance [31], [32], and log-map distance [33] are used to project image sets on geometric manifolds, such as Grassmannian manifold, and Lie group of Riemannian manifold.

To tackle the challenge of large data variability in video sequences in realistic surveillance scenarios, Cevikalp and Triggs [34] proposed to adaptively learn the image set samples from affine hull or convex hull models. The set-to-set distance is then termed as Affine Hull Image Set Distnace (AHISD), and Convex Hull Image Set Distance (CHISD), respectively. Hu et al. [35] introduced Sparse Approximated Nearest Points (SANP) by combining the affine hull with sparse representation to improve the AHISD algorithm. Despite the high accuracy of SANP based models on facial recognition tasks, they are computationally expensive as they perform one-to-one matching of sets together. Their performance is also affected by outliers. Yang et al. [27] addressed this problem by proposing an efficient iterative solver, called Regularized Nearest Points (RNP) method for image set based face recognition.

Furthermore, the majority of research focusing on non-parametric methods project image sets as one point on a geometric manifold [26], [29], [36], [37]. Discriminant analysis has been applied as a common strategy to compare image sets on different manifold surfaces in the following studies; Discriminative Canonical Correlations (DCC) [15], Covariance Discriminative Learning (CDL) [26], Graph Embedding Discriminant Analysis (GEDA) [37], Manifold Discriminant Analysis (MDA) [36], Discriminant Analysis on Riemannian manifold of Gaussian distributions (DARG) [38], Duplex Metric Learning (DML) [39], and Group Collaborative Representations (GCRs) [16] to name a few. Nevertheless, non-parametric techniques often require a larger dimension of the feature vectors, and relatively high-resolution images, compared to the total number of image sets in the training and test sets. As a result, their performance is only limited to small image sets. On the other hand, some studies [23], [24] extend the Linear Regression Classification (LRC) technique to perform image set based face and object recognition using small training data and low-resolution images.

### C. Deep Learning Methods

To date, image set classification using deep learning approaches have only been employed in a few studies. Lu et al. [40] introduce Multi-Manifold Deep Metric Learning (MMDML) motivated by the non-linearity learning capability of deep learning models and the MDD [21] method. Hayat et al. [11] posit a Deep Reconstruction Model (DRM) to represent image sets with a deep learning model. The structure of their framework is based on an Auto-Encoder (AE), which is composed of an encoder and a decoder. The input images are reconstructed with three hidden layers in each of the encoder and decoder, connected by a common hidden layer. In spite of the promising performance of DRM, and MMDML, the computational complexity of these techniques are high. In particular, the performance of DRM is undermined in the absence of LBP features, as opposed to using raw images for superior performance.

The success of covariance matrix and Symmetric Positive Definite (SPD) manifold methods for image set characterisation inspired Wang et al. [41] to propose a simple SPD manifold deep learning network, named as SymNet. The main components of SymNet include SPD Matrix Mapping Layer, Rectifying Layer, and SPD Matrix Pooling Layer. However, it is argued that the algorithm is not able to handle data variability on large-scale datasets resulting in lower performance,

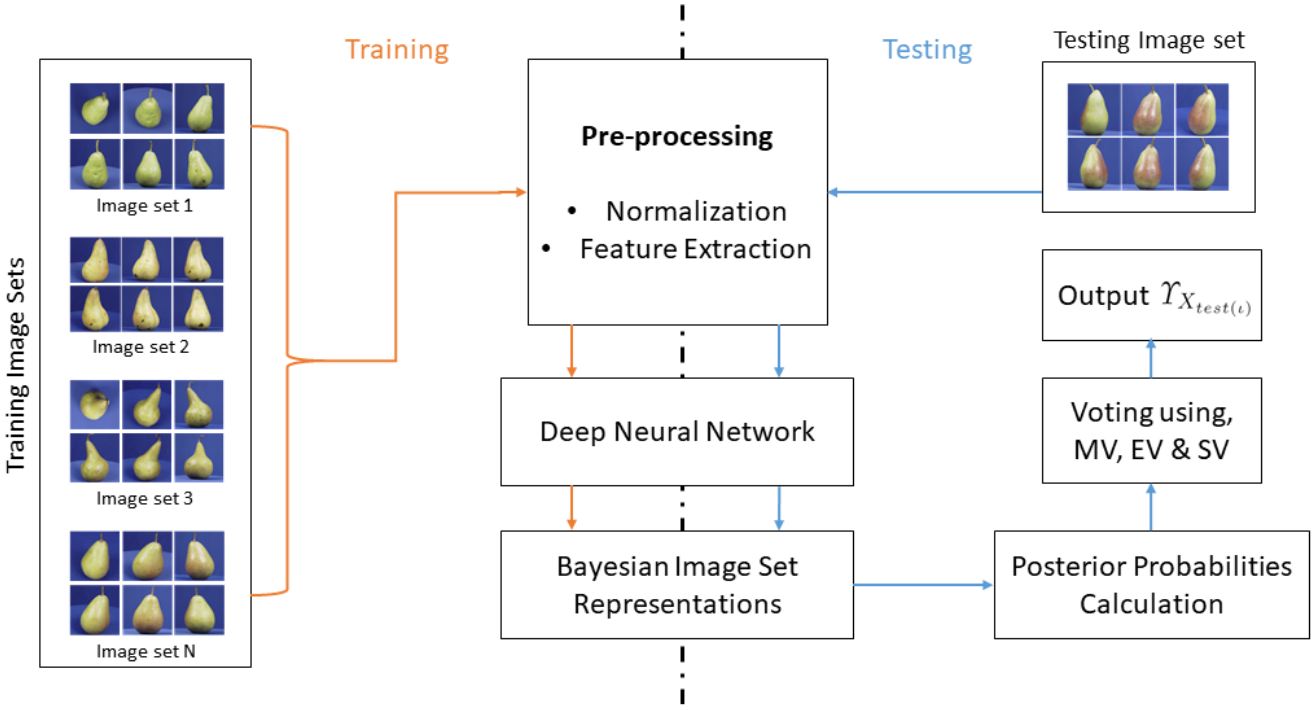



Fig. 2. Framework of the proposed deep Bayesian image set model. The framework is composed of two stages; training and testing. During training, we first implement and train the deep Bayesian neural network with the training image set to learn class-specific representations. During testing, Monte Carlo posterior probabilities are calculated for a given test image set. The final decision is then determined based on the three proposed voting strategies.

compared to the state-of-the-art techniques. The development of Convolutional Neural Networks (CNNs) has also attracted some researchers in the computer vision community to adopt the class-specific representation of image sets with CNN-based deep learning models in recent years [22]. Shah et al. [42] proposed an Iterative Deep Learning Model (IDLM), using one CNN layer and recurrent neural networks to recognise objects with image set classification. The characteristic of IDLM to iteratively learn feature representations provides an advantage to overcome the limitations of previous deep learning models [11] in terms of speed, performance, and hidden layer size. In contrast, the IDLM method is constrained to the fine-tuning of several hyper-parameters, and hand-crafted extraction of LBP features to achieve high performance.

In this paper, we overcome the limitations of the existing technique and propose Bayesian Image Set Representation for face and object recognition using deep learning. In the following section, we present our proposed method, which uses raw input images for training and testing.

## III. Proposed Methodology

The block diagram of our proposed framework is shown in Figure 2. It can be divided into four stages i.e., input, image pre-processing, training, and testing the network using three different voting strategies. All the images in each image set are first pre-processed before passing each set to the network for training purposes. Empirically, we found that image normalisation was an appropriate pre-processing step for object recognition task, and histogram equalisation for facial recognition. A detailed description of training is provided in Section III-A1. Once the network is trained, the network collects a Monte Carlo sample on the given pre-processed test image set and then calculates the posterior probabilities. The decision about the class of the test image set is then made based on three different voting strategies, including majority voting, soft voting and weighted voting (see section III-A2).

We adapt the architecture of Res-Net50 [43] as the backbone of our proposed model. We build the model by adding one dropout layer [44] with a ratio of 0.5 between the last activation layer and the last fully connected dense layer, where the network weights are initialised. This will collect a sample of Monte Carlo estimates [19] at both training and testing stages. The pre-trained network with ImageNet weights is then trained with the training data using fine-tuning.

### A. Image Set Formation

This section describes the formation of image sets followed by training and testing procedures of the proposed method in Section III-A1, and Section III-A2, respectively.

Given an image set $X_\iota$, which consists of multiple images, $I_1$, $I_2$, $I_3$, ...., $I_\nu$ with similar appearance variability of the same class $\varphi$, where $\nu$ is the number of images in each image set, an image set $X_\iota$ is formulated as follows:

$$X_\iota = \{I_1, I_2, I_3, ...., I_\nu\} \tag{1}$$

Let class labels be $\Phi = \{1, 2, 3, ..., \varphi\}$, a gallery of images $\Gamma_\varphi$ for each class of $\Phi$ can be represented as a group of image sets corresponding to a class $\varphi$ as in Equation 2:

$$\Gamma_\varphi = \{X_1, X_2, ..., X_\mu\} \ \forall \varphi \in \Phi \tag{2}$$

where $X_1, X_2, X_3, ...., X_\mu$ are $\mu$ image sets of a gallery $\Gamma_\varphi$, and $\Phi$ is the set of labels. Equation 3 defines an image set-based dataset $D$.

$$D = \{\Gamma_1, \Gamma_2, \Gamma_3, ...., \Gamma_\varphi\} \tag{3}$$

Since each image set $X_\iota \subset \Gamma_\varphi$ and $X_\iota \subset D$ , each gallery $\Gamma_\varphi$ contains $\nu \times \mu$ images, as formulated in Equation 4:

$$\forall \iota \in \mu \ \exists \ X_\iota \in \Gamma_\varphi \rightarrow \ \Gamma_\varphi = \{I_1, I_2, I_3, ...., I_{\nu \times \mu}\} \forall \varphi \in \Phi \tag{4}$$

*1) Training:* Let $M$ denote the total number of sets created for a dataset $D$, all the image sets of training galleries are combined to form one single image set, named as $X_{train}$, assuming there are a multiple numbers of training image sets in each gallery $\Gamma_\varphi$:

$$X_{train} = \{X_1, X_2, X_3, ...., X_M\} \tag{5}$$

If a gallery $\Gamma_\varphi$ does not contain any image sets, $\mu$ number of image sets $X_\iota$ are first created for the gallery before forming $X_{train}$. The images are all normalised by dividing with the maximum pixel value in each image set and then fed to model $f$ for training. Our training algorithm is provided in Algorithm 1.

**Algorithm 1** Training Procedure Using Image Sets

```
for all Γ_φ in D_train do
    if Γ_φ ∌ X_ι then
        Create μ number of image sets X_ι^μ for Γ_φ
    end if
end for
X_train = ⋃_{μ=1}^{M} X_μ
Input: Training image sets X_train
Normalise ∀I ∈ X_train
Train f(X_train)
Output: f
```

*2) Testing:* The ensemble of posterior probabilities $P_{I_\nu}$ for each image $I_\nu \in X_{test(\iota)}$ is approximated from Monte Carlo distribution of the image $Z^*_{I_\nu}$, by finding the mean of $T$ stochastic forward passes:

$$P_{I_\nu} = \frac{1}{T} \sum_{t=1}^{T} f_(Z^*_{I_\nu}) \tag{6}$$

The proposed deep Bayesian image set based technique (Algorithm 2) first calculates the posterior Monte Carlo probabilities $P_{X_{test(\iota)}}$ of the image set classifier $f$ based on each test image set $X_{test(\iota)} \in \Gamma_{test(\varphi)}$, which is made up of $\nu$ number of normalised test images $I_1, .., I_\nu$. Hence, the posterior probability of an image set $X_{test(\iota)}$ is a set of $\nu$ number of predicted posterior probabilities generated by the classifier, as represented in Equation 7:

$$P_{X_{test(\iota)}} = \{P_{I_1}, P_{I_2}, ..., P_{I_\nu}\} \tag{7}$$

Secondly, each testing image set $X_{test(\iota)}$ cast a vote for $\Gamma_{test(\varphi)} \forall \varphi \in \Phi$. The final decision about the class of the image set $\Upsilon_{X_{test(\iota)}}$ is then declared in accordance with the proposed voting strategies.

To classify image sets, we experimented our proposed technique with different voting strategies, including majority voting, soft voting and exponential weighted voting. Empirically, exponential weighted voting strategy demonstrated superior performance for most of the datasets. This is due to the fact that other voting strategies may choose the predicted class that the model is not most confident about but also the most uncertain. The weighted voted strategy can minimise this effect by assigning a weight to the predicted labels for image sets.

**Algorithm 2** Image Set Classification Algorithm

```
Input: D_test, T
for each gallery Γ_test(φ) in D_test do
    for each image set X_test(ι) in Γ_test(φ) do
        Normalise X_test(ι)
        for t in 1 to T do
            P_{X_test(ι)} =Predict(X_test(ι), f)   ▷ See Eq. (6-7)
        end for
        MV: Υ_{X_test(ι)} = mode(Υ̂_{X_test(ι)})   ▷ See Eq. (9-10)
        WV: Υ_{X_test(ι)} = arg max(wv_{X_test(ι)})   ▷ See Eq. (13)
        SV: Υ_{X_test(ι)} = arg max(δ_sv)   ▷ See Eq. (14-17)
    end for
end for
return Label y_test of X_test
```

*3) Majority Voting (MV):* Let $X_{test(\iota)}$ be a test image set as in Equation 1, the predicted label for an image $I \in X_{test(\iota)}$ is determined by the maximum predicted posterior probability $P_{I_\nu}$ of that image in the test set for a class $\varphi$ (Equation 8).

$$\hat{\Upsilon_{I_\nu}} = \arg\max(P_{I_\nu}) \tag{8}$$

The set of votes for a test image set is represented as follows:

$$\hat{\Upsilon}_{X_{test(\iota)}} = \{\hat{\Upsilon}_{I_1}, \hat{\Upsilon}_{I_2}, ..., \hat{\Upsilon}_{I_\nu}\} \tag{9}$$

Then, each test image $I$ casts an equal vote $\hat{\Upsilon_{I_\nu}}$ for the class $\varphi$. The most frequent label $\forall I_\nu \in X_{test(\iota)}$, which has the highest number of votes for class $\varphi$ is declared as the final class of the test image set $X_{test(\iota)}$:

$$\Upsilon_{X_{test(\iota)}} = mode(\hat{\Upsilon}_{X_{test(\iota)}}) \tag{10}$$

*4) Exponential Weighted Voting (EWV):* In Exponential Weighted Voting, each image $I_\nu \in X_{test(\iota)}$, casts a vote $\delta^I_\varphi$ for each class $\varphi \in \Phi$. Given $\beta$ is a constant, the weight of the vote $\delta^I_\varphi$ of a test image $I$ is defined as follows:

$$\delta^I_\varphi = e^{-\beta P_{I_\nu}} \tag{11}$$

Subsequently, the weights are accumulated for all images of the test set $X_{test(\iota)}$:

$$WV_{X_{test(\iota)}} = \sum_{I=1}^{\nu} \delta_{\varphi}^{I} \quad \forall \varphi \in \Phi \tag{12}$$

The final decision for $X_{test(\iota)}$ is determined by the class with the maximum accumulated weighted vote as:

$$\Upsilon_{X_{test(\iota)}} = \arg\max(WV_{X_{test(\iota)}}) \tag{13}$$

*5) Soft Voting (SV):* In soft voting as an ensemble voting strategy, the preference is given to a vote based on the degree of model $f$ certainty for an image set $X_{test(\iota)}$ per each iteration of $T$ stochastic forward pass to predict the gallery label $\varphi$ of the set. Equation 14 calculates the mean of predicted posterior probability of an image set per each class $\varphi \in \Phi$, which is distributed over the collected Monte Carlo Sample $Z^*$ in an iteration of $T$.

$$P_{(\varphi|X_{test(\iota)})} = \frac{1}{\nu} \sum_{l=1}^{\nu} P_{(I_\nu|\varphi)} \quad \forall \varphi \in \Phi \tag{14}$$

Thus, at each iteration of the stochastic forward pass $T$, there is a set of $\varphi$ predicted probabilities $\Psi_t, \forall t = 1 \in T$ for the given image set $X_{test(\iota)}$ :

$$\Psi_t = \{P_{(\Phi_1|X_t est\iota)}, P_{(\Phi_2|X_{test(\iota)})}, ..., P_{(\Phi_\varphi|X_{test(\iota)})}\} \tag{15}$$

The vote is cast to a class with the highest probability as the candidate label of an image set $\Upsilon_{X_{test(\iota)}}$.

$$\delta_{sv} = \max\left(P_{(\Phi_\varphi|X_{test(\iota)})\in\forall T_t}\right) \tag{16}$$

$$\Upsilon_{X_{test(\iota)}} = \arg\max(\delta_{sv}) \tag{17}$$

## IV. Experimental Results

We extensively evaluate our proposed framework on the publicly available object recognition datasets, including CIFAR-10 [45], CIFAR-100 [45], Tiny-ImageNet [46], MNIST [47] and ETH-80 [48]. We also evaluate our proposed model on CMU MoBo [49] and USCD/Honda [50] for facial recognition. All the experiments are conducted without any data augmentation. In the following, we provide details of these datasets and then discuss our experimental results.

### A. Object Recognition Datasets

*1) CIFAR-10 Dataset:* The CIFAR-10 dataset [45] is a collection of 10 object classes, including airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck. Each colour image is of size $32 \times 32$. There are 50,000 images for training and 10,000 images for testing, resulting in the total number of 60,000 images in the dataset. In our experiments, each class is considered as a gallery. We also form training and testing image sets based on different types of objects in each gallery. For instance, as shown in Figure. 3, all images of police cruiser, taxi, and convertibles are assigned to three separate image sets, corresponding to the automobile gallery. Test image sets that contain more than 4 images are used in our experiments. We perform our experiments for the total number of 9,869 images belonging to 287 test image sets and 10 object classes.


Fig. 3. An Example of CIFAR-10 Image Sets for Automobile Gallery. A random selection of images from five different sets, including Police Cruiser, Taxi, Convertible, Compact Cars, and Coupe are shown.

*2) CIFAR-100 Dataset:* CIFAR-100 [45] contains colour images of size $32 \times 32$ pixels, belonging to 100 different object classes. The training and testing sets are composed of 50,000 images and 10,000 respectively. In other words, there are 500 images per training class and 100 images per testing class. We follow the same procedure as CIFAR-10 to create image sets based on object types per gallery. After creating sets, the training set consists of 553 image sets, whereas 699 image sets for the test set. We observe that the difference in the number of image sets is due to the fact that there are some image sets in the training set that do not exist in the test set. Therefore, to balance the dataset, we only keep 538 common image sets in each train and test sets. After this modification, there are 538 training image sets containing 49,371 images in total. The testing dataset resulted in 9,890 images belonging to 538 testing images sets. Our experiments are conducted on test image sets containing more than 5 images with a total of 9,649 images for all the 100 galleries.

*3) MNIST Dataset:* MNIST dataset [47] consists of 70,000 black and white handwritten digit images, among which 60,000 are for training and 10,000 for testing, with the original image size of $28 \times 28$ pixels. There are 10 classes from zero to nine digits. We categorise each class as a gallery. Because of the similar appearance variations of images in each gallery, we create image sets that contain 100 images. The resulting test set is composed of 105 image sets for the ten galleries in our experiments. The images are upscaled to the resolution of $32 \times 32$ pixels.

*4) Tiny-ImageNet Dataset:* One of the challenging datasets for the task of object recognition is Tiny-ImageNet [46], which is a subset of ImageNet dataset. All images of the dataset are derived from ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [51] and downscaled to $64 \times 64$ pixels. The dataset contains three sets, including training, validation and test set. In total, there are 100,000 images belonging to 200 object classes. The validation set contains 50 images per class, while the test set has 10,000 unlabelled images. The significant appearance variability of images in this dataset can evaluate the generalisation ability of our proposed technique.

To perform our experiments, we create 10 image sets per gallery and assign 50 images per set in the training set. Moreover, for testing purposes, all the 50 images of the validation set are considered as one set corresponding to their relative galleries. Besides, we randomly select two image sets from each training gallery for our test set. Hence, there are 8 image sets per gallery for training, and each gallery of the new test set is composed of 3 sets. Overall, there are 30,000 images for testing and 80,000 images that constitute 1600 image sets of 200 galleries for training. We conduct our experiments on this testing set for all the 200 classes, which are composed of 3 sets and altogether 150 images per gallery.

*5) ETH-80 Dataset:* ETH-80 dataset [48] is one of the most popular datasets for the task of image set based object classification. The dataset consists of eight object categories including apples, cars, cows, cups, dogs, horses, and tomatoes. Each object is a gallery of 10 image sets. Each image set is composed of 41 images of an object taken from various viewpoints. Each gallery has image sets representing distinct types of an object. For example, different models of car, different breeds of dogs or different apple varieties. There are collectively 3280 images, 80 image sets, and 8 galleries in the dataset. We conduct our experiments on $128 \times 128$ cropped images. We also keep a split ratio of 80:20 by randomly selecting 32 images from each image set for training and 9 images for testing.

### B. Video-based Face Recognition Datasets

*1) CMU MoBo Dataset:* The CMU Motion of Body Dataset (CMU MoBo) [49] is a collection of videos capturing 25 individuals walking on a treadmill. The dataset objective was to advance biometric research on human gait analysis to identify people from the way they walk [49]. The videos are captured from 6 different viewpoints. All the subjects except the last one have four different walking patterns, such as slow walk, fast walk, inclined walk and holding a ball while walking. For the purpose of image set classification, we only use videos from the front cameras in our experiments. We followed the experimental protocol of previous works [11], [13], [52], and used the video sequences of the first 24 individuals containing all the four walking patterns. Each frame in the videos are considered as an image set. The face from each frame was detected using the Viola and Jones face detection algorithm [53]. Following the standard protocol of [11], [29], [34], [35], we randomly select one video sequence of a walking pattern for each subject as training image sets and the remaining three sequences as the test image sets. The images were rescaled to the resolution of $20 \times 20$ pixels, and converted to grayscale. Histogram equalisation was also applied to increase the contrast while minimising illumination variations in all the images. As opposed to the majority of existing techniques extracting features such as LBP, our experiments are performed on only raw images. For a fair comparison, we repeated the experiments 10 times with different random selection of training and test image sets.

*2) USCD/Honda Dataset:* The main objective of USCD/Honda dataset [50] is to evaluate face tracking and recognition algorithms. The dataset consists of 59 videos of 20 individuals. There are one to five number of videos for each subject with significant head rotations, pose variations and partial occlusions. Similar to the experimental protocols in [11], [29], [35], [36], [50], the face from each video frame was detected using Viola and Jones face detection algorithm [53]. We downsampled the detected faces to the resolution of $20 \times 20$. All the images were converted to grayscale before applying histogram equalisation. For the training set, we randomly selected one video sequence for each individual and used the remaining 39 as test image sets.

*3) YouTube Celebrities Dataset:* YouTube Celebrities (YTC) dataset [6] contains 1910 video clips of 47 celebrities and politicians downloaded from YouTube. The task of recognition for this dataset is comparatively challenging owing to high compression rates and low quality of images. There is also large variations in illumination, pose, and expressions in almost all of the face images. Since Viola and Jones face detection algorithm [53] failed to capture faces for a large number of frames, we used Incremental Learning Tracker [6] to crop the faces, similar to [11]. It was observed that there were many tracking errors because of the low resolution of images. Despite the cropped face region not being uniform across video frames, no refinement was applied to faces automatically tracked by the detection algorithm. To evaluate our experiments, the standard protocols proposed in [11], [26], [35] were followed. We divide the dataset into five folds with minimum overlap between the different folds. Each fold consists of 423 video clips in total, among which there are 9 videos per subject. All the tracked images are resampled to the resolution of $20 \times 20$, and converted to grayscale. Histogram equalisation is also applied to enhance image contrast. We randomly select 6 video clips per subject for testing and the remaining 3 image sets for training from each fold.

### C. Implementation Details and Network Hyperparameters

We implemented the proposed framework in Python using Tensorflow and Keras libraries, and ran the experiments on a Windows machine with 16GB RAM, 2.80 GHz Intel Core i7 CPU and 8GB NVDIA GeForce GTX GPU. The proposed ResNet model was designed with Adam optimiser with a learning rate of $e^{-4}$, and the batch size of 64 is used for CIFAR-10, CIFAR-100, MNIST, and Tiny-ImageNet. The batch size of 32 is selected for ETH-80, CMU MoBo and USCD/Honda. The number of epochs for each dataset is set between 10 to 15.

### D. Qualitative Results

Table I summarises the evaluation of our proposed technique on object recognition datasets based on three different voting strategies. Our empirical results suggest that the weighted voting strategy performs comparatively better than the other two strategies. The proposed image set-based deep Bayesian framework achieves the best performance on CIFAR-10 [45], and MNIST [47] with 100% accuracy using EWV strategy. Despite the large class variability manifested in Tiny-ImageNet dataset [46], our framework achieves a superior performance

for large-scale image recognition with 99.83% accuracy using soft voting. Similarly, the proposed network achieves 94.49% accuracy on CIFAR-100 [45] using soft voting. These results suggest that the proposed image set deep network can, to some extent, minimise the inter-class ambiguity even when there are a large number of classes such as 200 and 100 classes in the case of TinyImageNet and CIFAR-100 respectively. It is also evident that soft voting performs better when there are a large number of classes.

TABLE I
PERFORMANCE EVALUATION OF THE PROPOSED METHOD FOR OBJECT RECOGNITION (CLASSIFICATION ACCURACY) ON CIFAR-10, CIFAR-100, MNIST, ETH-80, AND TINY-IMAGENET DATASETS.

| Dataset | SV | MV | EWV |
|---|---|---|---|
| CIFAR-10 [45] | **100.00%** | 99.30% | **100.00%** |
| CIFAR-100 [45] | **94.97%** | 92.58% | 94.49% |
| MNIST [47] | **100%** | **100%** | **100%** |
| ETH-80 [48] | 97.50% | **98.75%** | **98.75%** |
| Tiny-ImageNet [46] | **99.83%** | 99.66% | 99.66% |

We evaluate the performance of our proposed technique for facial recognition on video-based datasets and report the average accuracy with standard deviation in Table II and Table III, along with image set based object recognition. The proposed framework achieves the best performance on USCD/Honda dataset [50] with 100% accuracy. On CMU MoBo [49], we achieved the average identification rate of 94.3% with standard deviation of 2.8%, followed by 63.74% average accuracy with standard deviation of 4.6% on YTC [6]. Moreover, as reported in Table I, the proposed framework achieves 98.75% accuracy on ETH-80 [48]. Overall, our technique performs better to recognise objects than facial identification.

### *E. Comparison with Image Set Classification Techniques*

We also compare the performance of our proposed framework for both object recognition and facial identification tasks with the recent state-of-the-art techniques. Results in [3], [11], [16], [26], [29], [34], [40], [42], [54], [55] are reported in this paper for comparison purposes. Table II reports that our proposed technique achieves superior performance on UCSCD/Honda dataset with 100% accuracy for all the three voting strategies. Our technique is on par with the existing state-of-the-art, OFDFL image set technique [3], as well as deep learning based image set techniques, such as IDLM [42], MMDML [40] and DRM [11].

Despite the state-of-the art performance of non-parametric CSSSIRT method [57], the performance of our technique on CMU/MoBo is nearly on par with the current CNN-based state-of-the-art IDLM [42], which relies on hand-crafted LBP features for good performance while our technique uses raw images/videos. Furthermore, as reported in Table III, the proposed technique achieves 63.74 % accuracy for YouTube Celebrity dataset, which is on par with existing methods. The complexity of the data set as discussed in Section IV is likely to contribute to the relatively lower performance of our technique, when compared with existing methods. For ETH-80, despite the 100% classification score of Duplex Metric Learning [39], which is a non-parametric method, our proposed deep learning technique still outperforms the current image set-based deep learning methods, including SymNet [41], DRM [11], MMDML [40], and IDLM [42] as reported in Table IV.

TABLE II
THE AVERAGE FACIAL IDENTIFICATION RATES AND STANDARD DEVIATION OF THE PROPOSED METHOD ON VIDEO-BASED DATASETS (I.E CMU MOBO, USCD/HONDA), COMPARED WITH THE STATE-OF-THE-ART TECHNIQUES.

| Datasets → Methods ↓ | CMU MoBo [49] | USCD/Honda [50] |
|---|---|---|
| CDL [26] | 92.5% ± 2.9% | 97.4% ± 1.3% |
| GCR [16] | 93.3% ± 2.5% | 99.7% ± 0.8% |
| MMD [29] | 93.6% ± 2.9% | 96.9% ± 2.3% |
| DFRV [54] | 94.4% ± 2.3% | 97.4% ± 1.9% |
| CHISD [34] | 94.2% ± 1.3% | 94.9% ± 1.9% |
| LMKML [55] | 94.5% ± 2.5% | 98.5% ± 2.5% |
| NDENP [56] | 94.6% ± 1.1% | **100.0% ± 0.0%** |
| ODFDL [3] | 97.5% ± 1.1% | **100.0% ± 0.0%** |
| MMDML [40] | 97.8% ± 1.0% | **100.0% ± 0.0%** |
| DRM [11] | 97.9% ± 0.7% | **100.0% ± 0.0%** |
| IDLM [42] | **98.4% ± 0.3%** | **100.0% ± 0.0%** |
| CSSIRT [57] | **98.5% ± 0.2%** | **100.0% ± 0.0%** |
| Ours (SV) | 94.3% ± 2.8% | **100.0% ± 0.0%** |
| Ours (MV) | 93.2% ± 2.5% | **100.0% ± 0.0%** |
| Ours (EWV) | 94.3% ± 2.8% | **100.0% ± 0.0%** |

TABLE III
THE AVERAGE FACIAL IDENTIFICATION RATES OF THE PROPOSED METHOD ON YOUTUBE CELEBRITY (YTC)

| Methods | Accuracy |
|---|---|
| DARG [38] | 51.74% |
| AHISD [34] | 54.18% |
| JDRML [58] | 57.87% |
| DRM [11] | 61.06% |
| PLRC [13] | 62.59% |
| DLRC [38] | 63.9% |
| SSDML [28] | 64.4% |
| Ours (SV) | 63.56% ± 4.7% |
| Ours (MV) | 62.94% ± 4.6% |
| Ours (EWV) | 63.74% ± 4.6% |

TABLE IV
THE CLASSIFICATION SCORE OF THE PROPOSED METHOD ON IMAGE SET BASED OBJECT RECOGNITION DATASET ETH-80 AND COMPARISON WITH THE STATE-OF-THE-ART TECHNIQUES.

| Methods | Accuracy |
|---|---|
| SFDL [14] | 90.5% |
| MMDML [40] | 94.5% |
| SymNet-v2 [41] | 97.0% |
| DRM [11] | 98.3% |
| IDLM [42] | 98.6% |
| ODFDL [3] | 98.7% |
| DML [39] | **100.0%** |
| Ours (SV) | 97.50% |
| Ours (MV) | 98.75% |
| Ours (EWV) | 98.75% |

TABLE V
CLASSIFICATION SCORE (%) OF THE PROPOSED METHOD FOR OBJECT RECOGNITION ON CIFAR-10, CIFAR-100, MNIST, AND COMPARISON WITH THE STATE-OF-THE-ART TECHNIQUES.

| Datasets → / Methods ↓ | CIFAR 10 [45] | CIFAR 100 [45] | MNIST [47] |
|---|---|---|---|
| ViT-H/14 [59] | 99.50 | **94.55** | - |
| BiT-L [60] | 99.37 | 93.51 | - |
| ResNet-152-SAM [61] | 98.2 | 87.80 | - |
| EfficientNet+DCL [62] | 98.21 | 88.35 | 84.08 |
| VBGS-KD [63] | 88.36 | - | 98.37 |
| IRCNN [64] | 91.83 | 92.89 | 99.71 |
| RandAlexNet [65] | 66.88 | - | 87.50 |
| Bayesian AlexNetHalf [66] | 79.00 | 38.00 | 99.00 |
| SOPCNN [67] | 94.29 | 72.96 | 99.83 |
| RadDropout [68] | - | 56.78 | 98.53 |
| DAN/K-DAN [69] | 87.79 | | 99.46 |
| ScatNets [70] | 93.00 | 73.50 | - |
| Ours (SV) | **100.0** | 94.49 | **100.0** |
| Ours (MV) | **99.30** | 92.56 | **100.0** |
| Ours (EWV) | **100.0** | 94.49 | **100.0** |

TABLE VI
THE CLASSIFICATION SCORE OF THE PROPOSED METHOD ON TINY-IMAGENET AND COMPARISON WITH THE STATE-OF-THE-ART TECHNIQUES.

| Methods | Accuracy |
|---|---|
| EfficientNet-B1 (DCL) [62] | 84.39% |
| IRCNN [64] | 51.92% |
| RadDropout [68] | 51.86% |
| DAN/K-DAN [69] | 78.02% |
| ScatNets [70] | 62.1% |
| Ours (SV) | **99.83**% |
| Ours (MV) | **99.66**% |
| Ours (EWV) | **99.66**% |

### *F. Comparison with State-of-the-art Single Image based Recognition Approaches*

To demonstrate the effectiveness of our proposed image set representation, we also compare the performance of our proposed technique with traditional one image based recognition methods. We report the results in [59]–[62], [62]–[64], [64]–[68], [68], [69], [69], [70], [70] in this paper for comparison purposes. Our results are reported in Table V and Table VI. For object recognition task, our proposed image set framework outperforms the current state-of-the-art deep learning techniques on CIFAR-10 and MNIST (see Table V). In comparison with the vision transformer model of ViT-H/14 [59], our technique achieves a relatively similar performance on CIFAR-100. Furthermore, as reported in Table VI, the proposed deep Bayesian image set network significantly outperforms the current state-of-the-art techniques on Tiny-ImageNet.

## V. ABLATION STUDY

We investigate the generalisation ability of our proposed Bayesian image set-based deep neural network under different situations. In this ablation study, the experiments are conducted on CIFAR-100 dataset using image sets to assess the performance of the proposed model against image noise, various training set sizes as well as variations in image sizes. There are two main reasons for selecting the CIFAR-100 dataset. Compared to the other datasets analysed in this paper, the performance of our technique was relatively low on CIFAR-100. In addition, the dataset contains low-resolution images with high number of galleries (100 classes), high image variations in each image set and comparatively the highest number of testing image sets (699 sets), which can sometimes be challenging to achieve high performance. We further compare image set classification performance with single-shot classification for the proposed deep Bayesian network.

TABLE VII
THE EFFECT OF IMAGE RESOLUTION ON THE PROPOSED TECHNIQUE FOR CIFAR-100 DATASET.

| S. No. | Image Size | SV | MV | EWV |
|---|---|---|---|---|
| 1 | $10 \times 10$ | 66.28% | 59.57% | 65.31% |
| 2 | $20 \times 20$ | 91.63% | 86.12% | 91.87% |
| 3 | $28 \times 28$ | 92.10% | 90.90% | 92.34% |
| 4 | $32 \times 32$ | 94.97% | 92.58% | 94.49% |
| 5 | $40 \times 40$ | 94.74% | 92.11% | 94.50% |

TABLE VIII
THE EFFECT OF IMAGE NOISE ON THE PROPOSED TECHNIQUE USING GAUSSIAN NOISE FOR CIFAR-100 DATASET

| S. No. | Mean | SV | MV | EWV |
|---|---|---|---|---|
| 1 | 0.5 | 21.77% | 20.81% | 20.57% |
| 2 | 0.1 | 49.76% | 45.69% | 50.96% |
| 3 | 0 | 49.76% | 46.17% | 50.71% |
| 4 | -0.1 | 48.56% | 44.98% | 48.56% |
| 5 | -0.5 | 19.14% | 17.46% | 19.13% |

### *A. Image Resolution Analysis*

The performance of our proposed technique against various image resolutions (i.e. $10 \times 10$, $20 \times 20$, $28 \times 28$, and $40 \times 40$ pixels) is reported in Table VII. With respect to the CIFAR-100 dataset original resolution of $32 \times 32$, variations in image sizes does not exhibit a significant impact on the proposed network performance. Although the down-sampling of images to a comparatively low-resolution of $20 \times 20$ reduces the performance by approximately $3\%$, the accuracy of $91.87\%$ is still among the top five state-of-the-art techniques according to Table V.

### *B. Image Noise Analysis*

To evaluate the robustness of our proposed network against image noise, we introduce Gaussian Noise to only test image sets. The training image sets remained without Gaussian Noise. Our experiments are carried out by varying the Gaussian mean of random distribution with a variance of 0.01 (See Table VIII). Notably, Gaussian noise has significantly impacted the performance of our proposed model as the noise was unseen to the network increasing its uncertain region across multiple classes.

TABLE IX
THE EFFECT OF TRAINING IMAGE SET SIZE ON THE PROPOSED TECHNIQUE USING CIFAR-100.

| S. No. | Training Size | SV | MV | EWV |
|---|---|---|---|---|
| 1 | 50% | 81.21% | 78.50% | 79.75% |
| 2 | 60% | 84.95% | 81.08% | 85.59% |
| 3 | 70% | 89.04% | 87.06% | 88.82% |
| 4 | 80% | 91.61% | 88.66% | 91.16% |
| 5 | 90% | 93.02% | 89.53% | 92.79% |

### C. Training Size Analysis

In this experiment, we train the proposed model with different proportions ranging from $50\%$ to $90\%$ of training image sets to assess its performance on smaller number of training image sets. We randomly select the training image sets based on the given rates and the remaining training image sets are added to the test set in order to keep the total number of image sets in the dataset constant. Table IX reports that as the number of training image sets reduces, the accuracy of the proposed deep Bayesian Network reduces to $81.21\%$ when trained on half of training image sets.

### D. Comparison of Image Set classification Vs. Single-Shot Classification

To demonstrate the superiority of our proposed Bayesian image-set based deep learning technique compared to the single-shot classification. In this experiment, we train the proposed Bayesian neural network on CIFAR-100 original training set with the same configurations mentioned in Sec IV-C without any hyper-tuning or data augmentation. The network achieved $53.20\%$ accuracy for single-shot classification while achieving $94.49\%$ for image set classification. These results demonstrate the effectiveness of our Bayesian image Set representations for classification task.

## VI. CONCLUSION

We propose a Bayesian image set representation using deep learning framework for facial and object classification. We evaluate the proposed technique on several benchmark datasets. Specifically, the experiments were performed without any data augmentation on a wide range of image galleries and image resolutions as low as $20 \times 20$ and as high as $128 \times 128$ pixels, such as low-resolution facial images (CMU MoBo, YouTube Celebrity and USCD/Honda), high resolution object recognition (ETH-80, Tiny-ImageNet), and low resolution object classification (CIFAR-10 and CIFAR-100). We also demonstrate that image set classification can be applied to any application even on single-shot datasets and yet achieve high accuracy. Our proposed Bayesian image set representations based model outperforms the majority of state-of-the-art deep learning techniques for both object and video-based facial recognition tasks.

## ACKNOWLEDGMENTS

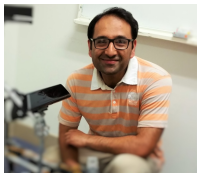
This work is supported by Edith Cowan University, Australia.

## REFERENCES

[1] B. Liu, L. Jing, J. Li, J. Yu, A. Gittens, and M. W. Mahoney, "Group collaborative representation for image set classification," *International Journal of Computer Vision*, vol. 127, pp. 181–206, 2018.
[2] X. Wang, Y. Han, V. C. M. Leung, D. Niyato, X. Yan, and X. Chen, "Convergence of edge computing and deep learning: A comprehensive survey," *IEEE Communications Surveys Tutorials*, vol. 22, no. 2, pp. 869–904, 2020.
[3] G. Zhang, J. Yang, Y. Zheng, Z. Luo, and J. Zhang, "Optimal discriminative feature and dictionary learning for image set classification," *Information Sciences*, vol. 547, pp. 498–513, 2021.
[4] C. Ding and D. Tao, "Trunk-branch ensemble convolutional neural networks for video-based face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, pp. 1002–1014, 2018.
[5] J. Yang, P. Ren, D. Zhang, D. Chen, F. Wen, H. Li, and G. Hua, "Neural aggregation network for video face recognition," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5216–5225, 2017.
[6] M. Kim, S. Kumar, V. Pavlovic, and H. Rowley, "Face tracking and recognition with visual constraints in real-world videos," *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2008.
[7] H. Joo, T. Simon, and Y. Sheikh, "Total capture: A 3d deformation model for tracking faces, hands, and bodies," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8320–8329, 2018.
[8] J. Lin, C. Gan, and S. Han, "Tsm: Temporal shift module for efficient video understanding," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7082–7092, 2019.
[9] C. Ding, C. Xu, and D. Tao, "Multi-task pose-invariant face recognition," *IEEE Transactions on Image Processing*, vol. 24, pp. 980–993, 2015.
[10] R. Ranjan, V. Patel, and R. Chellappa, "Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, pp. 121–135, 2019.
[11] M. Hayat, M. Bennamoun, and S. An, "Deep reconstruction models for image set classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 4, pp. 713–727, 2015.
[12] D. Liu, L. Liu, Y. Tie, and L. Qi, "Multi-task image set classification via joint representation with class-level sparsity and intra-task low-rankness," *Pattern Recognition Letters*, vol. 132, pp. 99–105, 2020.
[13] Q. Feng, Y. Zhou, and R. Lan, "Pairwise linear regression classification for image set retrieval," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4865–4872.
[14] J. Lu, G. Wang, and J. Zhou, "Simultaneous feature and dictionary learning for image set based face recognition," *IEEE Transactions on Image Processing*, vol. 26, no. 8, pp. 4042–4054, 2017.
[15] T.-K. Kim, J. Kittler, and R. Cipolla, "Discriminative learning and recognition of image set classes using canonical correlations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 1005–1018, 2007.
[16] B. Liu, L. Jing, J. Li, J. Yu, A. Gittens, and M. W. Mahoney, "Group collaborative representation for image set classification," *International Journal of Computer Vision*, vol. 127, no. 2, pp. 181–206, 2019.
[17] B. T. Phan, "Bayesian deep learning and uncertainty in computer vision," Master's thesis, University of Waterloo, 2019.
[18] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" *arXiv preprint arXiv:1703.04977*, 2017.
[19] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*. PMLR, 2016, pp. 1050–1059.
[20] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya *et al.*, "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," *Information Fusion*, 2021.
[21] O. Arandjelovic, G. Shakhnarovich, J. Fisher, R. Cipolla, and T. Darrell, "Face recognition with image sets using manifold density divergence," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, 2005, pp. 581–588 vol. 1.
[22] Z.-Q. Zhao, S.-T. Xu, D. Liu, W.-D. Tian, and Z.-D. Jiang, "A review of image set classification," *Neurocomputing*, vol. 335, pp. 251–260, 2019.
[23] S. A. A. Shah, U. Nadeem, M. Bennamoun, F. Sohel, and R. Togneri, "Efficient image set classification using linear regression based image reconstruction," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 601–610.
[24] U. Nadeem, S. A. A. Shah, M. Bennamoun, R. Togneri, and F. Sohel, "Real time surveillance for low resolution and limited data scenarios:

An image set classification approach," *Information Sciences*, vol. 580, pp. 578–597, 2021.
[25] E. G. Ortiz, A. Wright, and M. Shah, "Face recognition in movie trailers via mean sequence sparse representation-based classification," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3531–3538.
[26] R. Wang, H. Guo, L. S. Davis, and Q. Dai, "Covariance discriminative learning: A natural and efficient approach to image set classification," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2496–2503.
[27] M. Yang, P. Zhu, L. Van Gool, and L. Zhang, "Face recognition based on regularized nearest points between image sets," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 2013, pp. 1–7.
[28] P. Zhu, L. Zhang, W. Zuo, and D. Zhang, "From point to set: Extend the learning of distance metrics," in *2013 IEEE International Conference on Computer Vision*, 2013, pp. 2664–2671.
[29] R. Wang, S. Shan, X. Chen, and W. Gao, "Manifold-manifold distance with application to face recognition based on image set," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.
[30] J. Hamm and D. D. Lee, "Grassmann discriminant analysis: a unifying view on subspace-based learning," in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 376–383.
[31] P. Turaga, A. Veeraraghavan, A. Srivastava, and R. Chellappa, "Statistical computations on grassmann and stiefel manifolds for image and video-based recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 11, pp. 2273–2286, 2011.
[32] M. Hayat and M. Bennamoun, "An automatic framework for textured 3d video-based facial expression recognition," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 301–313, 2014.
[33] M. T. Harandi, C. Sanderson, A. Wiliem, and B. C. Lovell, "Kernel analysis over riemannian manifolds for visual recognition of actions, pedestrians and textures," in *2012 IEEE Workshop on the Applications of Computer Vision (WACV)*, 2012, pp. 433–439.
[34] H. Cevikalp and B. Triggs, "Face recognition based on image sets," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2567–2573.
[35] Y. Hu, A. S. Mian, and R. Owens, "Face recognition using sparse approximated nearest points between image sets," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 10, pp. 1992–2004, 2012.
[36] R. Wang and X. Chen, "Manifold discriminant analysis," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 429–436.
[37] M. T. Harandi, C. Sanderson, S. Shirazi, and B. C. Lovell, "Graph embedding discriminant analysis on grassmannian manifolds for improved image set matching," in *CVPR 2011*, 2011, pp. 2705–2712.
[38] W. Wang, R. Wang, Z. Huang, S. Shan, and X. Chen, "Discriminant analysis on riemannian manifold of gaussian distributions for face recognition with image sets," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 151–163, 2018.
[39] G. Cheng, P. Zhou, and J. Han, "Duplex metric learning for image set classification," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 281–292, 2018.
[40] J. Lu, G. Wang, W. Deng, P. Moulin, and J. Zhou, "Multi-manifold deep metric learning for image set classification," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1137–1145.
[41] R. Wang, X.-J. Wu, and J. Kittler, "Symnet: A simple symmetric positive definite manifold deep learning method for image set classification," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.
[42] S. A. A. Shah, M. Bennamoun, and F. Boussaid, "Iterative deep learning for image set based face and object recognition," *Neurocomputing*, vol. 174, pp. 866–874, 2016.
[43] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
[44] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.
[45] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.
[46] Y. Le and X. Yang, "Tiny imagenet visual recognition challenge," *CS 231N*, vol. 7, p. 7, 2015.
[47] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
[48] B. Leibe and B. Schiele, "Analyzing appearance and contour based methods for object categorization," in *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, vol. 2, 2003, pp. II–409.
[49] R. Gross and J. Shi, "The cmu motion of body (mobo) database," 2001.
[50] K.-C. Lee, J. Ho, M.-H. Yang, and D. Kriegman, "Video-based face recognition using probabilistic appearance manifolds," in *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, vol. 1. IEEE, 2003, pp. I–I.
[51] 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.
[52] L. Chen, "Dual linear regression based classification for face cluster recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 2673–2680.
[53] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.
[54] Y.-C. Chen, V. M. Patel, P. J. Phillips, and R. Chellappa, "Dictionary-based face and person recognition from unconstrained video." *IEEE Access*, vol. 3, no. 3, pp. 1783–1798, 2015.
[55] J. Lu, G. Wang, and P. Moulin, "Image set classification using holistic multiple order statistics features and localized multi-kernel metric learning," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 329–336.
[56] Z. Ren, Q. Sun, and C. Yang, "Nonnegative discriminative encoded nearest points for image set classification," *Neural Computing and Applications*, vol. 32, no. 13, pp. 9081–9092, 2020.
[57] Z. Ren, B. Wu, X. Zhang, and Q. Sun, "Image set classification using candidate sets selection and improved reverse training," *Neurocomputing*, vol. 341, pp. 60–69, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231219303121
[58] W. Yan, Q. Sun, H. Sun, and Y. Li, "Joint dimensionality reduction and metric learning for image set classification," *Information Sciences*, vol. 516, pp. 109–124, 2020.
[59] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
[60] A. Kolesnikov, L. Beyer, X. Zhai, J. Puigcerver, J. Yung, S. Gelly, and N. Houlsby, "Big transfer (bit): General visual representation learning," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*. Springer, 2020, pp. 491–507.
[61] X. Chen, C.-J. Hsieh, and B. Gong, "When vision transformers outperform resnets without pretraining or strong data augmentations," *arXiv preprint arXiv:2106.01548*, 2021.
[62] Y. Luo, Y. Wong, M. Kankanhalli, and Q. Zhao, "Direction concentration learning: Enhancing congruency in machine learning," *IEEE transactions on pattern analysis and machine intelligence*, 2019.
[63] Y. Ming, H. Fu, Y. Jiang, and H. Yu, "Variational bayesian group-level sparsification for knowledge distillation," *IEEE Access*, vol. 8, pp. 126 628–126 636, 2020.
[64] M. Z. Alom, M. Hasan, C. Yakopcic, T. M. Taha, and V. K. Asari, "Inception recurrent convolutional neural network for object recognition," *Machine Vision and Applications*, vol. 32, no. 1, pp. 1–14, 2021.
[65] P. Tabarisaadi, A. Khosravi, and S. Nahavandi, "Bayesian randomly wired neural network with variational inference for image recognition," in *International Conference on Neural Information Processing*. Springer, 2020, pp. 197–207.
[66] K. Shridhar, F. Laumann, and M. Liwicki, "A comprehensive guide to bayesian convolutional neural network with variational inference," *arXiv preprint arXiv:1901.02731*, 2019.
[67] Y. Assiri, "Stochastic optimization of plain convolutional neural networks with simple methods," *arXiv preprint arXiv:2001.08856*, 2020.
[68] H. Wang, W. Yang, Z. Zhao, T. Luo, J. Wang, and Y. Tang, "Rademacher dropout: An adaptive dropout for deep neural network via optimizing generalization gap," *Neurocomputing*, vol. 357, pp. 177–187, 2019.
[69] C.-Y. Low, J. Park, and A. B.-J. Teoh, "Stacking-based deep neural network: deep analytic network for pattern classification," *IEEE transactions on cybernetics*, vol. 50, no. 12, pp. 5021–5034, 2019.
[70] F. Cotter and N. Kingsbury, "A learnable scatternet: Locally invariant convolutional layers," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 350–354.

**Nima Mirnateghi** received Master of Information Technology in Data Science at Murdoch University, Perth, WA, Australia. He completed his Bachelor of Computer Science from the University of Wollongong (UOW), NSW, Australia. He has worked as a research assistant at the School of Science and is a member of the centre for AI and Machine Learning, Edith Cowan University, WA, Australia. He has actively participated in a number of IEEE WA competitions, and software/game development hackathons. His current research interests include computer vision, pattern recognition, statistical analysis, object recognition, facial detection, machine learning, and image processing.



**Syed Afaq Ali Shah** received the PhD degree in computer vision/machine learning from The University of Western Australia (UWA), Australia. He is currently a Senior Lecturer at Edith Cowan University, Australia and an Adjunct Senior Lecturer with the Department of Computer Science and Software Engineering, UWA, Perth, WA, Australia. His current research interests include deep learning, object/face recognition, Scene understanding, and image processing. Dr. Shah was a recipient of the Start Something Prize for Research Impact through Enterprise for 3-D Facial Analysis Project funded by the Australian Research Council. He has authored over 50 research articles and co-authored a book, A guide to convolutional neural networks for computer vision.