01 Nov 2020

Set Operation Aided Network For Action Units Detection

Huiyuan Yang
Missouri University of Science and Technology, hyang@mst.edu

Taoyue Wang

Lijun Yin

Recommended Citation

H. Yang et al., "Set Operation Aided Network For Action Units Detection," *Proceedings - 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2020*, pp. 229 - 235, article no. 9320229, Institute of Electrical and Electronics Engineers, Nov 2020.
The definitive version is available at https://doi.org/10.1109/FG47880.2020.00030

2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020) | 978-1-7281-3079-8/20/$31.00 ©2020 IEEE | DOI: 10.1109/FG47880.2020.00030

Set Operation Aided Network for Action Units Detection

Huiyuan Yang, Taoyue Wang and Lijun Yin
Department of Computer Science
Binghamton University-SUNY, Binghamton, USA
{*hyang51, twang61*}*@binghamton.edu, lijun@cs.binghamton.edu*

***Abstract*— As a large number of parameters exist in deep-model based methods, training such models usually requires many fully AU-annotated facial images. This is true with regard to the number of frames in two widely used datasets: BP4D[31] and DISFA[18], while those frames were captured from a small number of subjects (41, 27 respectively). This is problematic, as subjects produce highly consistent facial muscle movements, adding more frames per subject would only adds more close points in the feature space, and thus the classifier does not benefit from those extra frames. Data augmentation methods can be applied to alleviate the problem to a certain degree, but they fail to augment new subjects. We propose a novel Set Operation Aided Network (SO-Net) for action units detection. Specifically, new features and the corresponding labels are generated by adding set operations to both the feature and label spaces. The generated new features can be treated as a representation of a hypothetical image. As a result, we can implicitly obtain training examples beyond what was originally observed in the dataset. Therefore, the deep model is forced to learn subject-independent features, and is generalizable to unseen subjects. SO-Net is end-to-end trainable, and can be flexibly plugged in any CNN model during training. We evaluate the proposed method on two public datasets, BP4D and DISFA. The experiment shows a state-of-the-art performance, demonstrating the effectiveness of the proposed method.**

I. INTRODUCTION

In recent decades, automatic action units (AU) detection (defined in facial Action Coding System [9]) has been an essential task for facial expression analysis. The conventional automatic AU detection approaches rely on a set of well-defined features (i.e., SIFT[16], HoG [8], LBP [1]), classifying those features by a classifier. Recently, deep-model based methods have shown great progress and been widely used in various computer vision tasks, including image classification[13], image segmentation[5], and object detection[25]. As a result, more researchers are embracing deep-model based methods for AU detection [17] [26] [35] [15] [22] [14] and achieving state-of-the-art performances.

Currently, a number of facial expression databases are commonly used for facial action units detection [17]. However, most of these annotated datasets have a limited number of subjects. For example, the BP4D dataset has around 140,000 frames with AU labels from 41 subjects (3414 frames per subject on average), and the DISFA dataset has 130,000 frames from 27 subjects (4815 frames per subject on average). As mentioned in [11], when using appearance features, increasing the number of subjects in the training set significantly improved performance, while increasing the number of training frames per subject did not. We argue that more subjects is one of the reason why a higher F1-score is usually observed in BP4D (41 subjects) than DISFA (27 subjects). In paper [11], the authors also verified through experiments that 450 frames per subject were enough to achieve competitive classification performance. As subjects are highly consistent in producing facial actions, adding more frames per subject would only be adding more close data points in the feature space, and classification performance would not change.

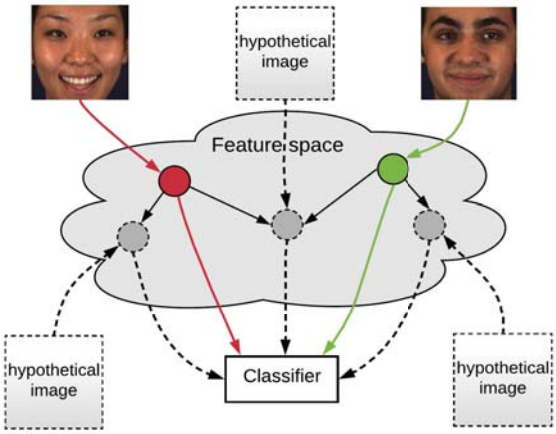

Fig. 1. Set operation is applied to augment features in the feature space, and also the corresponding label sets. The set operation aided network aims to synthesize new features and label sets, which can be considered as hypothetical images in the sample space, for AU detection, so that the model generalizes well to unseen subjects.

To address the problem, one may apply the widely used data augmentation methods (i.e. cropping, rotation, random noise) to add variations to the dataset, but it fails to add subject-related variations. Another recent data augmentation method called *Mixup* [30], which has been proved effective in improving performance by averaging two images with the same label, does not work well for the multi-label problem. There are very few works trying to solve this issue. Girard et al. [11] gave experimental analysis about the subject-related issue. Zhang et al. [33] proposed to learn subject-independent features for AU detection through adversarial training. Niu et al. [21] added a person-specific shape module as regularization to the deep model for AU detection, hoping

978-1-7281-3079-8/20/$31.00 ©2020 IEEE
DOI 10.1109/FG47880.2020.00030

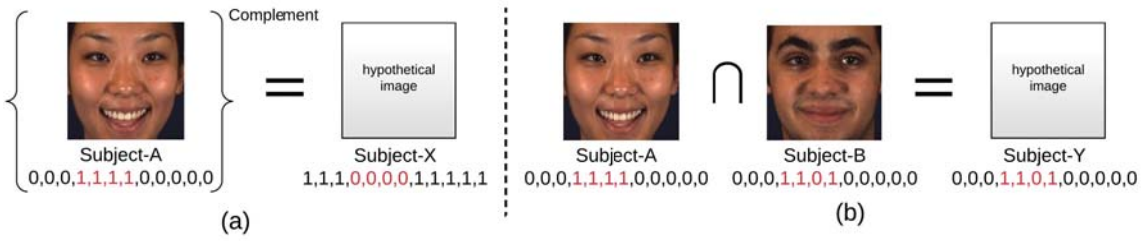

Fig. 2. **Set operation in feature space**: (a) *complement* operation is applied to **feature-A** (extracted from subject-A) and its label sets, then a new **feature-X** (hypothetical image-X) and its corresponding label sets are obtained. (b) **feature-Y** and its label sets are generated by performing intersection operation to **feature-A** and **feature-B** and their label sets. By manipulating the features in the feature space, we can obtain features beyond what was originally observed in the dataset.

to learn a feature that is orthogonal to the person-specific feature.

Inspired by some recent works [19] [24] [20] [2] that manipulate the semantic contents of the samples by arithmetic operations in representation space, we propose to add set operations in the feature and label spaces to augment new training samples. Those set operations can be considered as a kind of data augmentation method that is directly applied to generate new features and labels. An intuitive example can be found in Fig.2. Feature vectors are extracted from a pair of images to represent the corresponding semantic content (i.e., {AU1, AU2, AU7} and {AU1, AU2, AU5} respectively) using a backbone network. The shared (implicit) concept here is the {AU1, AU2}. If an intersection operation is applied to both the feature A and feature B, we should obtain a feature vector which represents AU1, AU2, but no longer represents either of the AU7 and AU5. Based on these set operation concepts, we propose a Set Operation Aided Network (SO-Net) for AU detection which directly augments the data in the feature space. As seen in Fig.1, a shared feature extraction module is used for both input image pairs, and then the set operations (i.e., complement, union, intersection) are applied to generate features and their corresponding AU labels. Each generated feature can be considered as a representation of a hypothetical image from an unknown subject. As a result, we can obtain features beyond what was originally observed in the dataset, and are able to train a model capable of being generalized to unseen subjects.

The contributions of this paper can be summarized as follows:

- An end-to-end trainable set operations aided network is proposed to augment data in the feature space. By augmenting these new features (representation of hypothetical images) for training, the model is forced to learn subject robust features, thus has the ability to generalize well to unseen subjects.
- Our proposed method is very flexible, light-weight, and can be easily plugged into any deep model for training.
- Our proposed method has been evaluated on two public datasets, and achieves better performance than the state-of-the-art methods.

II. RELATED WORK

Facial action unit detection: Many conventional automatic facial action unit detection approaches often perform feature learning to extract robust features from appearance or geometric information from the whole face or local patches (i.e., SIFT, LBP, HoG). The extracted features are selected and classified by a classifier (for example: Adaboost, SVM). Deep models have made great progress in many tasks, and show promising results on AU detection. The deep model based methods can be interpreted as a joint learning of image features and classifier, so a well-designed hand-crafted feature is not needed any more. To improve the performance of deep model based AU detection, many factors have been considered by different researchers. Some works try to combine multiple tasks for AU detection. Shao et al. [26] proposed a deep model for jointly learning facial action unit detection and face alignment tasks, so the two tasks can benefit each other and achieve good performance on two public datasets. Niu et al. [21] used a person-specific shape regularization module to enforce the deep model to learn an orthogonal feature, which will be more discriminative and generalizable for AU detection. AUs are related to different regions of the face, and different facial regions can provide unique information for recognizing AUs. Some works try to detect AUs by focusing on regions of interest. Zhao et al.[35] proposed a unified network (DRML) that simultaneously addresses deep learning and multi-label learning problems for AU detection. The proposed method aims to identify the active sparse facial regions, and as a result, the structural information of the face is also captured. Li et al.[15] add an enhancing and cropping (EAC) net into a pre-trained CNN model; the EAC-net contains both attention layers and cropping layers, which significantly improves the performance for AU detection. Correlations among AUs and AU-Emotion are another consideration for robust AU detection. Zhang et al. [32] proposed a domain-knowledge driven method for jointly learning multiple AU classifiers. The dependency among AUs and expressions make it possible to train a model without using any AU labels. Peng et al. [23] used two kinds of auxiliary information, which exist among AUs

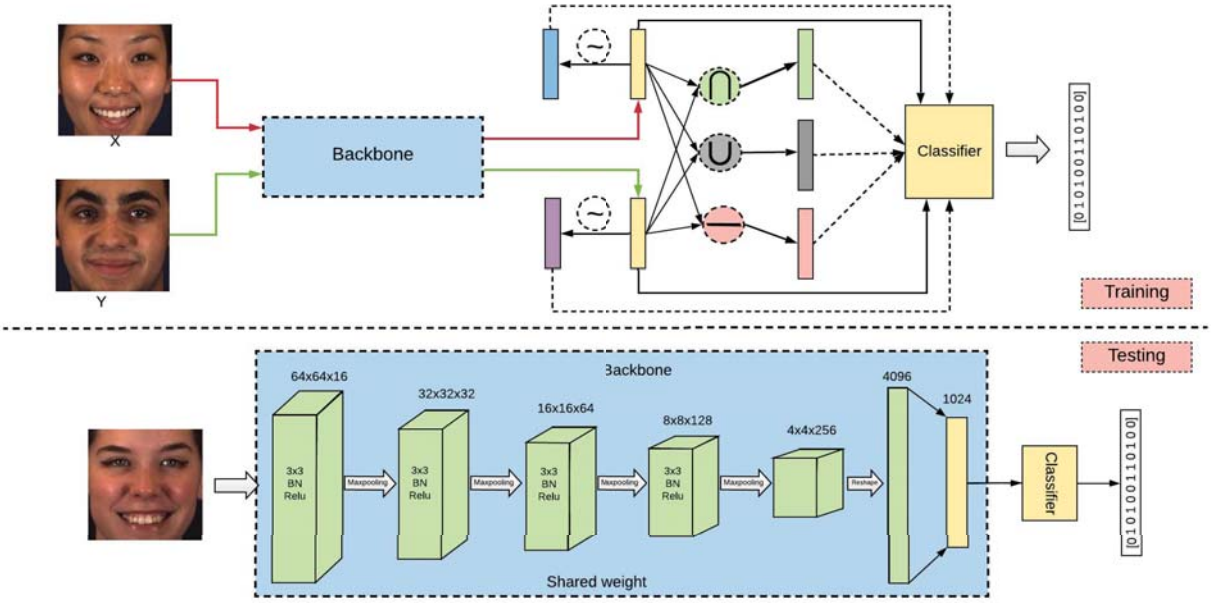

Fig. 3. Framework of proposed SO-Net. A CNN model consists of a backbone for feature extraction and classifier that maps the feature into different AU occurrence. Set operation is added to the last fully connected layer during training, Feature-X (F_X) and Feature-Y (F_Y)are extracted from input image pairs through backbone (shared by two input images); then set operations are applied to the features, including *Complement* (F_X), *Complement* (F_Y), *Union*(F_X,F_Y), *Intersection*(F_X, F_Y), *Subtraction*(F_X, F_Y) , and also applied to the label sets respectively. The extracted features *(solid lines)* and synthesized features *(dotted lines)* are sent to the classifier for AU detection. The combined loss is used to update both the classifier and backbone, so our model is end-to-end trainable. The set operation can be removed during testing, and the model [*backbone* + *classifier*] will run just like a regular CNN model.

and expressions, for AU detection in partially AU-labeled and fully expression-labeled facial images; the proposed network is trained using a dual semi-supervised learning approach. Peng and Wang [22] generated pseudo-AU labels according to the probalilistic dependencies among AUs and expressions, and then designed a weakly supervised AU detection method via adversarial training. When sequence data is available, temporal information is also useful for AU detection. Graphical models, like *e.g.*, HMMs[27], Hidden CRF[3] and Gaussian process models[4], can be used to model this temporal information. However, these models may not work very well for modeling long sequences. Chu et al.[6] used LSTM to model the temporal information, which was further combined with the spatial information for Au detection, and achieved good performance. Li et al.[14] combined the facial region of interest and LSTM-based temporal information, significantly improving the performance.

These methods have achieved promising results on public datasets, e.g., DISFA [18], BP4D [31]. However, deep models usually have millions of parameters to be optimized through training on large accurately labelled images, so overfitting is often observed on a specific dataset due to limited training data. Another issue, as mentioned in [11], is that the number of subjects has an important effect on automatic facial unit detection, but there are only 41 subjects in BP4D, and 27 subjects in DISFA, meaning deep models may not generalize well to unseen subjects.

Arithmetic operation in representation space: To explain arithmetic operation in the learned representation space, we can use one canonical example provided in [24] that *vector("King") - vector("Man") + vector("Woman")* resulted in a vector whose nearest neighbor was *vector("Queen")* in the representation space. Readers may find more details in [24]. Mikolov et al.[19] has demonstrated rich linear structures in representation space by applying some simple arithmetic operations. Radford et al. [24] tried to model face attributes like emotion, hairs, glasses and gender by performing some simple arithmetic operations on the representation space; their work also showed the benefits of developing arithmetic operations in representation space, which could dramatically reduce the amount of data needed for modeling complex images. In order to generalize recognition to unseen attribute-object compositions, Nagarajan and Grauman [20] proposed modeling attributes as operators, learning a semantic embedding space that explicitly factors out attributes from their accompanying objects. Recently, Alfassy et al. [2] proposed a novel method for the multi-label few-shot classification problem. They combined pairs of given examples in both feature space and sets of labels using set operations, so that resulting synthesized feature vectors will correspond to examples whose label sets are obtained through some set operations on the label sets. As a result, the proposed method is able to perform augmentation on examples of novel categories and show promising performance. Our proposed method is inspired by [2], but their differences are in two-

fold. First, the focus and problem domain are different, as paper [2] focuses on the multi-label few-shot classification problem, while our method focuses on the "*large number of frames from a small number of subjects* " problem, and tries to learn a subject-invariant features for AU detection. Second, a new set operation (complement operation) is used in the feature space.

III. METHOD

Our approach is illustrated in Fig.3. To improve the performance of a CNN model, set operations are added to the last fully connected convolutional layer (FC) for training; during testing, the model runs exactly like a regular CNN model (i.e. VGG). Details of the proposed method is in the following sections.

A. Set operation aided neural network

Given a dataset $X = \{(x_i, l_i)\}_{i=1}^{N}$, where x_i represents the training image, and $l_i \in [0,1]^K$ is the AU labels, N and K are the total number of training images and AU labels. As shown in Fig.3, (x, l_x) and (y, l_y) are input images and corresponding set of multiple labels, F_x and F_y are features extracted from a backbone network F. Instead of using a big model, e.g. InceptionV3 [28] or ResNet-34[12], we use a very light-weight VGG-like structure, as we want to train our model from scratch.

Set operations, noted as $SO(.)$, are added between the FC layer and the classifier during training, and removed for testing. The function $SO(.)$'s goal is to synthesize a feature vector in the feature space $\mathcal{F}$:

$$SO(F_x, F_y) = F_z \in \mathcal{F} \tag{1}$$

which corresponds to a hypothetical image Z in the image space $\mathcal{X}$ and its feature extracted by the backbone F_z. Since this is a multiple-class classification problem, $SO(.)$ can also be applied to the label space $\mathcal{L}$ as well:

$$SO(l_x, l_y) = l_z \in \mathcal{L} \tag{2}$$

which means that if an image Z is observed, it would then receive l_z as its label set.

Different set operations:{ *Complement, Union, Intersection, Subtraction* } are used as $SO(.)$. The original feature vectors, F_x and F_y, and the outputs of the $SO(.)$ function, namely F_x^{com}, F_y^{com}, F_z^{uni}, F_z^{int} and F_z^{sub}, are fed into a classifier C. Here $F_x^{com} = SO^{complement}(l_x, \textit{null})$, similar for F_y^{com}. Binary Cross-Entropy multi-label classification loss is used here, which is defined as below:

$$L(\widehat{l}, l) = -\sum_{i=1}^{N} l_i * log(\widehat{l_i}) + (1 - l_i) * log(1 - \widehat{l_i}) \tag{3}$$

where l being the desired binary ground-truth labels vectors, and $\widehat{l}$ denotes the predicted AU occurrence. The whole model is end-to-end trained from scratch, and the total loss is a weighted sum of classification loss from both real training data and hypothetical images:

$$\begin{aligned} L_{total} = {} & \alpha_1 * L\big(C(F_x), l_x\big) + \alpha_2 * L\big(C(F_y), l_y\big) \\ & + \alpha_3 * L\big(C(F_x^{com}), l_x^{com}\big) + \alpha_4 * L\big(C(F_y^{com}), l_y^{com}\big) \\ & + \alpha_5 * L\big(C(F_z^{uni}), l_z^{uni}\big) + \alpha_6 * L\big(C(F_z^{int}), l_z^{int}\big) \\ & + \alpha_7 * L\big(C(F_z^{sub}), l_z^{sub}\big) \end{aligned}$$

TABLE I

SET OPERATION FUNCTION USED IN OUR PROPOSED METHOD

Operator $SO(.)$	Feature Vectors	Multi-Labels
Complement	$1 - F_x$	$\lvert 1 - l_x \rvert$
Union	$max(F_x, F_y)$	$max(l_x, l_y)$
Intersection	$min(F_x, F_y)$	$min(l_x, l_y)$
Subtraction	$ReLU(F_x - F_y)$	$\lvert l_x - l_y \rvert$

B. Set operation functions

Selecting a set of suitable set operation functions $SO(.)$ is not a trivial problem, as it plays a key role in generating synthetic features in the feature space $\mathcal{F}$ and the according labels in the label space $\mathcal{L}$. The generated features and labels will contribute to the training of the multi-label classifier.

In the learned representation spaces, Mikolov et al.[19] has demonstrated that simple arithmetic operations reveal rich linear structure in representation space. Radford et al. [24] performed some simple arithmetic operations on the representation space to model face attributes like emotion, hairs, glasses and gender. Inspired by these works, we employ several simple set operation functions as defined in $SO(.)$, which directly manipulate feature vectors in $\mathcal{F}$ and label in label space $\mathcal{L}$, as shown in Table I.

IV. EXPERIMENT

BP4D[31] and DISFA[18] are two widely used datasets for evaluating AU detection methods. Our proposed method is evaluated on these two datasets, and F1-scores are reported as well. The F1-score is the harmonic mean of the precision and recall. As the number of AUs differ in different datasets, we report both metrics on each AU as well as average metrics over all AUs (denoted as Avg.).

A. Datasets

BP4D: BP4D is a dynamic spontaneous facial expression database, which contains 328 2D and 3D videos collected from 41 subjects (23 females, 18 males) under eight different tasks. Following previous research, only 2D videos are used here. The most expressive frames are manually labeled for AU occurrence, which resulted in a dataset of 140,000 manually FACS-annotated frames. To compare with state-of-the-art methods, 12 AUs are selected to evaluate the performance.

The images are split into 3 folds, where the subjects in any two subsets are mutually exclusive. Then, a 3-fold subject-independent cross validation is performed.

DISFA: 27 subjects (12 females, 15 males) involved in the DISFA dataset. For each subject, two videos were recorded

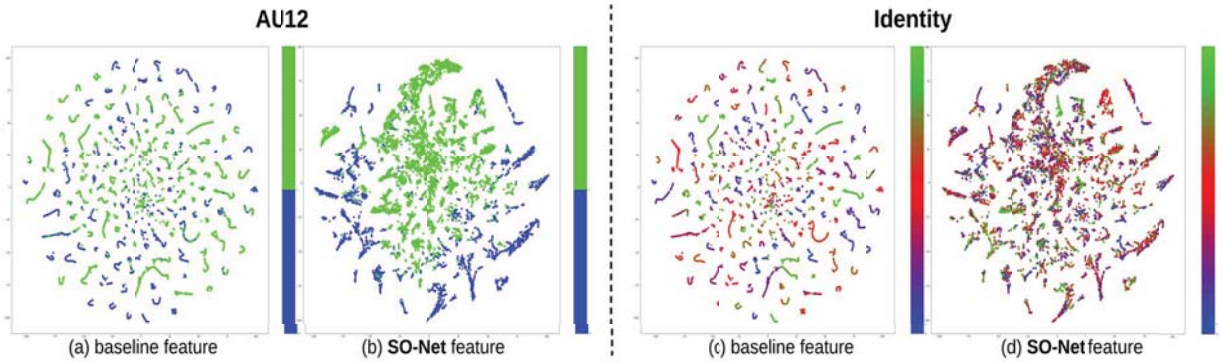

Fig. 4. A visualization of t-SNE embedding using deep features on the BP4D database by coloring each frame in terms of AU12 (a, b) and subject identities (c, d). The features (noted as baseline) used in (a) and (c) is the model training without the set operation module. The clustering effect in (a) and (c) reveal that the baseline features encode information about not only facial AUs but also subject identities. Compared to the baseline feature, our features show clearer separation in terms of AU12 (b), and reduce the influence caused by individual differences (d). Best viewed in color.

using two cameras (left camera and right camera) while watching videos. 12 AUs are labeled with AU intensity from 0 to 5 and 66 facial landmarks were provided. This results in about 130,000 valid AU labeled frames. Following the experimental setting of previous work, 8 of the 12 AUs with AU intensity greater than 0 are used from the left camera.

Subject-exclusive 3-fold cross-validation is performed on the BP4D dataset, and the best model trained on BP4D is further fine-tuned to the DISFA dataset.

B. Implementation details

All the face images are aligned and cropped to the size of 140x140 using affine transformation based on the provided facial landmarks, randomly cropped to 128x128 for training, and center-cropping for testing. Random horizontal flip is also applied for data augmentation.

AUs samples have imbalanced distributions. For example, the frames with occurrence of AU25 are almost 7 times greater than AU2 in DISFA dataset. In order to balance the training data, following the strategy in [15], we manually repeated 4 to 7 times for the less occurring AUs. During the training, an image pair is constructed by randomly selecting two images from the dataset.

An Adam optimizer with an initial learning rate of 0.001 is applied for optimizing the stem network. Our model is trained for 30 epochs, α_1, α_2 are set to 2, and the other α are set to 1. All experiments are implemented using TensorFlow and performed on the NVIDIA Geoforce 1080 Ti GPU.

C. Results

1) Comparison with the state-of-the-art: We compare our proposed method with the state-of-the-art methods based on same setting: subject-exclusive three-fold cross validation. Those methods include traditional methods: Linear SVM (LSVM) [10], Joint Patch and Multi-label Learning (JPML) [34] and deep-model based methods: Deep Region Multi-Label Learning (DRML)[35], Enhancing and Cropping Network (EAC-Net) [15], Finetuned VGG Network (FVGG), Deep Structured Inference Network (DSIN) [7], Joint AU detection and face Alignment (JAA) [26], learning Optical Flow Network (OF-Net) [29] and Local relationship learning with Person-specific shape regularization (LP-Ne) [21]. For fair comparison, we excluded the studies which use sequence for AU detection, such as ROI-LSTM [14]. OF-Net [29] learned temporal information from a single frame for AU detection, so it is still a frame-based AU detection method and added to the table for comparison. For these state-of-the-art methods, we use their reported results from the paper.

Table II shows the results of different methods on the BP4D database. First, we compare with the baseline, which is exactly the same as the SO-Net, except the set operations are disabled. As we can see, our proposed method shows 8.1% improvement over the baseline.

Our proposed method outperforms all the state-of-the-art methods except LP-Net. The LP-Net contains three sub-networks for feature learning, local relationship modeling and person-specific shape regularization, during which the 68 facial landmarks are used for shape regularization. As compared to LP-Net, our proposed method only adds a set-operation module to the feature space during training, which makes it very light-weight, and easy to insert into any baseline model. At the same time, our method achieves a close performance to LP-Net (60.8% vs 61.0%). It also worth mentioning that our method achieves the best F1-score for 8 of 12 AUs (AU2, AU4, *AU6*, AU10, AU12, AU14, AU15 and AU17)

The comparison with the state-of-the-art methods on the DISFA database is reported in Table III. The results reflect that our method gives the best F1-score, around 1.6% improvement over the state-of-the-art method. By comparing with the baseline, our method shows 7.2% improvement.

D. Influence of individual set operation

Experiments have been conducted to evaluate the influence of individual set operation, and the F1 scores are reported on

TABLE II

F1 SCORES IN TERMS OF 12 AUS ARE REPORTED FOR THE PROPOSED METHOD AND THE STATE-OF-THE-ART METHODS ON BP4D DATASET. BRACKETED AND BOLD NUMBERS INDICATE THE BEST PERFORMANCE; BOLD NUMBERS INDICATE THE SECOND BEST.

Method	AU1	AU2	AU4	AU6	AU7	AU10	AU12	AU14	AU15	AU17	AU23	AU24	***Avg***
LSVM [10]	23.2	22.8	23.1	27.2	47.1	77.2	63.7	64.3	18.4	33.0	19.4	20.7	35.3
JPML[34]	32.6	25.6	37.4	42.3.	50.5	72.2	74.1	65.7	38.1	40.0	30.4	42.3	45.9
DRML[35]	36.4	41.8	43.0	55.0	67.0	66.3	65.8	54.1	33.2	48.0	31.7	30.0	48.3
FVGG	27.8	27.6	18.3	69.7	69.1	78.1	63.2	36.4	26.1	50.7	22.8	35.9	43.8
EAC-net[15]	39.0	35.2	48.6	76.1	72.9	81.9	86.2	58.8	37.5	59.1	35.9	35.8	55.9
DSIN [7]	**[51.7]**	40.4	**56.0**	76.1	73.5	79.9	85.4	62.7	37.3	62.9	38.8	**41.6**	58.9
JAA [26]	47.2	44.0	54.9	77.5	74.6.	84.0	86.9	61.9	43.6	60.3	**42.7**	**[41.9]**	60.0
OF-Net [29]	**50.8**	**45.3**	56.6	75.9	**75.9**	80.9	88.4	63.4	41.6	**60.6**	39.1	37.8	59.7
LP-Net [21]	43.4	38.0	54.2	77.1	**[76.7]**	83.8	87.2	**63.3**	**45.3**	60.5	**[48.1]**	54.2	**[61.0]**
Baseline	35.0	35.7	47.2	**[80.0]**	77.2	**[84.9]**	**89.0**	62.0	24.2	57.2	17.8	22.3	52.7
SO-Net	40.2	**[46.2]**	**[56.0]**	**79.3**	73.5	**84.2**	**[90.8]**	**[64.7]**	**[55.9]**	**[61.0]**	37.4	40.2	**60.8**

TABLE III

F1 SCORES IN TERMS OF 8 AUS ARE REPORTED FOR THE PROPOSED METHOD AND THE STATE-OF-THE-ART METHODS ON BP4D DATASET. BRACKETED AND BOLD NUMBERS INDICATE THE BEST PERFORMANCE; BOLD NUMBERS INDICATE THE SECOND BEST.

Method	AU1	AU2	AU4	AU6	AU9	AU12	AU25	AU26	***Avg***
LSVM [10]	10.8	10.0	21.8	15.7	11.5	70.4	12.0	22.1	21.8
DRML [35]	17.3	17.7	37.4	29.0	10.7	37.7	38.5	20.1	26.7
EAC-net [15]	**41.5**	26.4	66.4	**50.7**	**[80.5]**	**[89.3]**	88.9	15.6	48.5
DSIN [7]	42.4	39.0	68.4	28.6	46.8	70.8	**90.4**	42.2	53.6
JAA [26]	**[43.7]**	**[46.2]**	56.0	41.4	44.7	69.6	88.3	58.4	56.0
OF-Net [29]	30.9	34.7	63.9	44.5	31.9	**78.3**	84.7	**60.5**	53.7
LP-Net [21]	29.9	24.7	**[72.7]**	46.8	**49.6**	72.9	**[93.8]**	**[65.0]**	**56.9**
Baseline	25.0	28.3	63.1	47.6	31.8	76.5	80.7	57.9	51.3
SO-Net	33.8	**44.5**	**70.3**	**[57.6]**	39.7	78.2	86.7	57.3	**[58.5]**

TABLE IV

INFLUENCE OF INDIVIDUAL SET OPERATION ON PERFORMANCE. F1 SCORES ARE REPORTED ON BOTH BP4D AND DISFA DATASETS

Set operation	BP4D	DISFA
Baseline	52.7	51.3
Complement	57.2	54.5
Union	59.1	54.3
Intersection	58.3	56.3
Subtraction	59.7	52.0

both BP4D and DISFA datasets in Table.IV. As we can see, *Union* and *Subtraction* operations work better in the BP4D dataset; while *Intersection* operation works the best in the DISFA dataset.

1) Visualization of the learnt features: To answer the question as to why our proposed method has the ability to generalize well to unseen subjects, and to give insight into the feature space, we first extract the features from the BP4D dataset using our method (noted as **SO-Net features**), and use a method without set-operation module (noted as **baseline features**). Fig.4 shows the t-SNE embedding of frames, which are colored in terms of AU12 ((a)(b), different colors means presence or absence of AU12) and subject identities ((c)(d), different colors represents different subjects). As we can see in (a) and (c), the baseline feature shows a strong distributional biases toward subject identity. In other words, the baseline feature encodes information about not only facial action units, but also subject identities. On the other hand, our feature shows a much clearer separation as seen in (b), implying our feature is able to capture necessary information for facial action units detection, and is also subject-independent (d), thereby works well for unseen subjects.

V. CONCLUSION

In this paper, we have presented a novel set operation aided neural network for AU detection. The set operation is added to the extracted features for input image pairs, including *complement, union, intersection and subtraction*, to generate synthesized features. The synthesized features can be treated as features extracted from hypothetical images. As a result, we can obtain samples beyond what was originally observed. The training data and the synthesized features/labels are combined to train the classifier, and the classification error is used to update both backbone and classifier, so our model is end-to-end trainable. By learning on real data and synthesized data through set operation, the model is forced to learn a more general representation, and therefore works well for unseen subjects.

Future work will involve exploring additional operations in the feature and label space, and also the visualization of hypothetical images for synthesized features.

VI. Acknowledgement

The material is based on the work supported in part by the NSF under grant CNS-1629898 and the Center of Imaging, Acoustics, and Perception Science (CIAPS) of the Research Foundation of Binghamton University.

References

[1] T. Ahonen, A. Hadid, and M. Pietikäinen. Face recognition with local binary patterns. In *European conference on computer vision*, pages 469–481. Springer, 2004.

[2] A. Alfassy, L. Karlinsky, A. Aides, J. Shtok, S. Harary, R. Feris, R. Giryes, and A. M. Bronstein. Laso: Label-set operations networks for multi-label few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6548–6557, 2019.

[3] K.-Y. Chang, T.-L. Liu, and S.-H. Lai. Learning partially-observed hidden conditional random fields for facial expression recognition. 2009.

[4] J. Chen, M. Kim, Y. Wang, and Q. Ji. Switching gaussian process dynamic models for simultaneous composite motion tracking and recognition. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2655–2662. IEEE, 2009.

[5] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2018.

[6] W.-S. Chu, F. De la Torre, and J. F. Cohn. Modeling spatial and temporal cues for multi-label facial action unit detection. *arXiv preprint arXiv:1608.00911*, 2016.

[7] C. Corneanu, M. Madadi, and S. Escalera. Deep structure inference network for facial action unit recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 298–313, 2018.

[8] N. Danal. Histgram of oriented gradients for human detection. In *Proc. of IEEE Conference on Conputer Vision and Pattern Recognition (CVPR), 2005*, pages 886–893, 2005.

[9] R. Ekman. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997.

[10] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *Journal of machine learning research*, 9(Aug):1871–1874, 2008.

[11] J. M. Girard, J. F. Cohn, L. A. Jeni, S. Lucey, and F. De la Torre. How much training data for facial action unit detection? In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 1, pages 1–8. IEEE, 2015.

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

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[14] W. Li, F. Abtahi, and Z. Zhu. Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1841–1850, 2017.

[15] W. Li, F. Abtahi, Z. Zhu, and L. Yin. Eac-net: A region-based deep enhancing and cropping approach for facial action unit detection. *arXiv preprint arXiv:1702.02925*, 2017.

[16] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[17] B. Martinez, M. F. Valstar, B. Jiang, and M. Pantic. Automatic analysis of facial actions: A survey. *IEEE transactions on affective computing*, 2017.

[18] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013.

[19] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[20] T. Nagarajan and K. Grauman. Attributes as operators: factorizing unseen attribute-object compositions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 169–185, 2018.

[21] X. Niu, H. Han, S. Yang, Y. Huang, and S. Shan. Local relationship learning with person-specific shape regularization for facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11917–11926, 2019.

[22] G. Peng and S. Wang. Weakly supervised facial action unit recognition through adversarial training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2188–2196, 2018.

[23] G. Peng and S. Wang. Dual semi-supervised learning for facial action unit recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8827–8834, 2019.

[24] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[25] 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*, pages 91–99, 2015.

[26] Z. Shao, Z. Liu, J. Cai, and L. Ma. Deep adaptive attention for joint facial action unit detection and face alignment. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 705–720, 2018.

[27] Y. Sun, M. Reale, and L. Yin. Recognizing partial facial action units based on 3d dynamic range data for facial expression recognition. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*, pages 1–8. IEEE, 2008.

[28] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[29] H. Yang and L. Yin. Learning temporal information from a single image for au detection. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–8. IEEE, 2019.

[30] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[31] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014.

[32] Y. Zhang, W. Dong, B.-G. Hu, and Q. Ji. Classifier learning with prior probabilities for facial action unit recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5108–5116, 2018.

[33] Z. Zhang, S. Zhai, and L. Yin. Identity-based adversarial training of deep cnns for facial action unit recognition. In *BMVC*, page 226, 2018.

[34] K. Zhao, W.-S. Chu, F. De la Torre, J. F. Cohn, and H. Zhang. Joint patch and multi-label learning for facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2207–2216, 2015.

[35] K. Zhao, W.-S. Chu, and H. Zhang. Deep region and multi-label learning for facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3391–3399, 2016.