01 May 2019

Learning Temporal Information From A Single Image For AU Detection

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

Lijun Yin

Recommended Citation

H. Yang and L. Yin, "Learning Temporal Information From A Single Image For AU Detection," *Proceedings - 14th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2019*, article no. 8756556, Institute of Electrical and Electronics Engineers, May 2019.
The definitive version is available at https://doi.org/10.1109/FG.2019.8756556

Learning Temporal Information From A Single Image For AU Detection

Huiyuan Yang and Lijun Yin
Department of Computer Science
State University of New York at Binghamton, USA
hyang51@binghamton.edu; lijun@cs.binghamton.edu

***Abstract*— Automatic Facial Action Units (AUs) detection is the recognition of the facial appearance changes caused by the contraction or relaxation of one or more related facial muscles. Compared to the sequence-based methods, a decreased performance is observed for the static image-based AU detection, due to the loss of temporal information. To solve this problem, we propose a novel method that implicitly learns temporal information from a single image for AU detection by adding a hidden optical-flow layer to concatenate two Convolutional Neural Networks (CNNs) models: optical-flow net (OF-Net) and AU detection net (AU-Net). The OF-Net is designed to estimate the facial appearance changes (optical flow) from a single input image through unsupervised learning. The AU-Net accepts the estimated optical-flow as input and predicts the AU occurrence. By training both OF-Net and AU-Net jointly, our model achieves better performance than training them separately, as the AU-Net provides semantic constraints for the optical-flow learning and helps generate more meaningful optical-flow. In return, the estimated optical-flow, which reflects facial appearance changes, benefits the AU-Net. Our proposed method has been evaluated on two benchmarks: BP4D and DISFA, and the experiments show significant performance improvement as compared to the state-of-the-art methods.**

I. Introduction

Action Units (AUs), defined by the Facial Action Coding System (FACS)[7], describe the contraction or relaxation of one or more facial muscles. Detecting facial action units has become an essential task for facial analysis, *i.e.*, facial expression recognition, depression analysis [23], and the measurement of pain in patients [16].

As shown in Fig.1, AU4 is described as *brow lowerer* with related facial muscles, *i.e.*, *depressor glabellae, depressor supercilli, currugator*; following that, a frame sequence is given to show the dynamic process from neutral face to peak state of AU4. A motivation of this work is that facial appearance changes (temporal information) play an important role in AU detection. Compared to the sequence-based methods [25][10][13][5], which utilize the temporal information extracted from the sequence for AU detection, a decreased performance is usually observed when using the static image-based AU detection methods because of a loss of temporal information [21][20][34][35][14]. Temporal information is crucial for AU detection, however is unavailable for static image. It is a big challenge to learn the temporal information from a single image. This paper addresses this issue by exploiting a new learning model with a hidden optical-flow learning layer.

978-1-7281-0089-0/19/$31.00 ©2019 IEEE

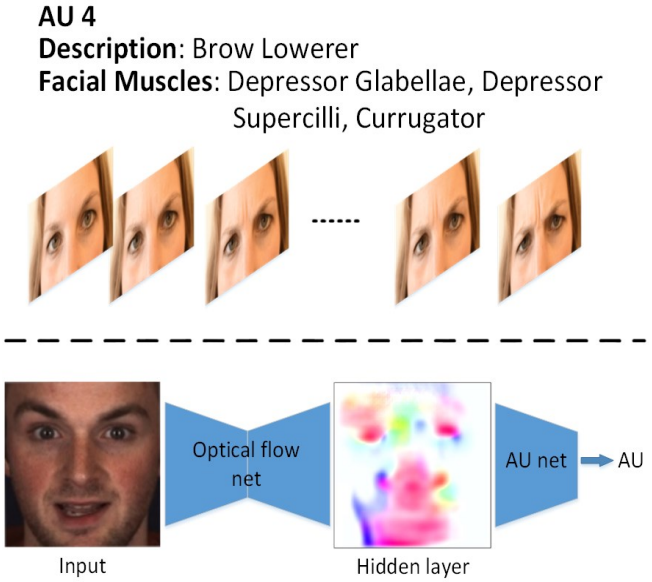

Fig. 1. The upper part is the definition of AU4 by semantic description and related facial muscle movement. The lower part is our proposed method, where the temporal information is implicitly learned from a static image.

Optical flow is used to describe the relative motion between two images which are collected at times t and $t+\Delta t$ at each pixel position. On the other hand, each AU is related to facial muscle movement, and those muscle movements produce motion in facial appearance, which could be detected and represented by optical flow. Therefore, it's quite straightforward to utilize optical flow for AU detection [15][29][8]. But there are two disadvantages of those methods: first, the optical flow is computationally expensive and storage intensive; second, in order to compute optical flow, at least two frames are needed, which make it useless for static AU detection.

Recently, convolutional neural networks (CNNs) have been widely used in various computer vision tasks, for example, image classification[12], image segmentation[4], object detection[24], and achieved better performance than the corresponding traditional methods. However, applying CNNs to predict the optical flow is still a challenging task as the CNNs need not only learn image features but also correspondence among them. FlowNet[6] and several related works [22][9] exploited the ability of CNNs to predict optical flow using synthesized datasets. Although those works

have successfully demonstrated that CNNs could be used to estimate the optical flow, they may not work well in real world scenarios due to the domain gap between the synthesized data and real world data. Since a large size of optical flow ground truth is extremely difficult to obtain, some works[1][36][19] proposed to train the optical flow estimation networks using an unsupervised method, where, the optical flow estimation problem is considered as an image reconstruction problem[11].

Inspired by those works, we proposed a method to learn optical flow from static image for AU detection. Our proposed method contains an optical flow net (OF-Net) and an AU detection net (AU-Net). The OF-Net is designed to predict optical flow from a single frame, and is trained by unsupervised learning. The AU-Net is used to detect AUs from the estimated optical flow. During our experiments, we found that training both models jointly worked much better than training them separately, because the AU loss provides a semantic constraint for the optical flow learning. For example, AU1 refers to *inner brow raiser*, AU4 refers to *brow lowerer* and AU17 refers to *chin raiser*; they all provide some semantic clues about the facial appearance changes, and those changes can be represented by the corresponding optical flow. With the semantic constraints and also other pixel-level constraints, the OF-Net learns to predict a much more meaningful optical flow, which makes it a better representation for AU detection.

Our approach has the following contributions:

1) A unique semantic AU loss is added to the loss functions of the OF-Net, which helps to predict an optical flow that focus on AU-related facial appearance changes. The semantic loss is quite unique for the optical flow learning, and differentiates our method from the others.
2) To the best of our knowledge, this is the first creative application for improving static AU detection using learned temporal information (optical flow), which is estimated from a single image.
3) Our proposed method has been evaluated on two public datasets, and achieves better performance than the state-of-the-art methods.

II. Related work

AU detection has been studied for decades, and many approaches have been proposed. For those approaches, two types of features: hand-crafted or learning based features, were usually used. The hand-crafted and learned features can also be classified by the utilization of temporal information into non-temporal and temporal features. Accordingly, the AU detection approaches can be applied to either static images or image sequences. Zhao et al.[34] proposed a joint patch and multi-label learning (JPML) method for AU detection, which considered the pair-wise relationship (positive correlations and negative competitions) among AUs. Compared to the local pair-wise AU dependency, Wang et al.[27] proposed learning AU dependencies from training data generated by using a restricted Boltzmann machine that learns the high-level AU semantical relationships. From the relationships, they were able to predict AUs. Furthermore, Zhang et al.[31] proposed an approach to learn AU classifiers without AU labels by considering the prior probabilities of both expression-independent and expression dependent AUs. As AUs are active on specific facial regions, some works then try to detect AUs by focusing on the regions of interest. Zhao et al.[35] proposed a deep region and multi-label learning (DRML) method that forced the CNN model to focus on important facial regions for different AUs. Li et al.[14] designed an enhancing and cropping (EAC) net, which contained attention layers and cropping layers. The EAC net was added into a pre-trained CNN model and significantly improved the performance for AU detection.

Temporal information plays a crucial role in AU detection, and graphical models were widely used for modeling the temporal information, *e.g.*, HMMs[25], Hidden CRF[2] and Gaussian process models[3]. Recently, researchers are embracing the Long Short-Term Memory (LSTM) for modeling temporal information for long sequences. Chu et al.[5] used a hybrid network to jointly address spatial and temporal features, and AU correlation. The temporal information, which was modeled by LSTM, played a crucial role for the improved performance. Li et al.[13] further proposed to combine the facial region of interest and LSTM-based temporal fusing, and significantly improved the performance.

Optical Flow has been applied for AU detection (or facial expression recognition) for decades [28] [15] [17] , but none of those works was built on learned optical flow, not to mention learning optical flow from a single image. Considering the advantages of covolutional neural networks in various computer vision tasks, researchers are looking to apply CNNs to optical flow estimation. FlowNet is a supervised end-to-end convolutional networks for optical flow estimation, which was first introduced by Fischer et al.[6]. As getting optical flow ground truth for realistic video is extremely difficult, they generated a large synthetic Flying Chairs dataset using the rendering of 3D chairs with a random background image, to train their model. However, the method may not work well in real world scenarios due to the gap between the synthesized data and real world data. Based on FlowNet, Ilg et al.[9] introduced the FlowNet2, a slower, more complex, but much more accurate supervised end-to-end neural networks. Ranjan and Black [22] proposed a spatial pyramid network (SPyNet) which fused a spatial pyramid into FlowNet. The large motions were then estimated in a coarse-to-fine way.

However, all those above-mentioned approaches relied on supervised learning, and the performance depends on the availability of large amount of optical flow ground truth or the quality of synthesized data. Because a large size of optical flow ground truth is extremely difficult to obtain, and the domain gap between the synthesized data and real world data, researchers have started focusing on end-to-end unsupervised learning methods. The first work was proposed by Ahmadi and Patras [1], they converted the optical flow estimation problem to an image reconstruction problem, by

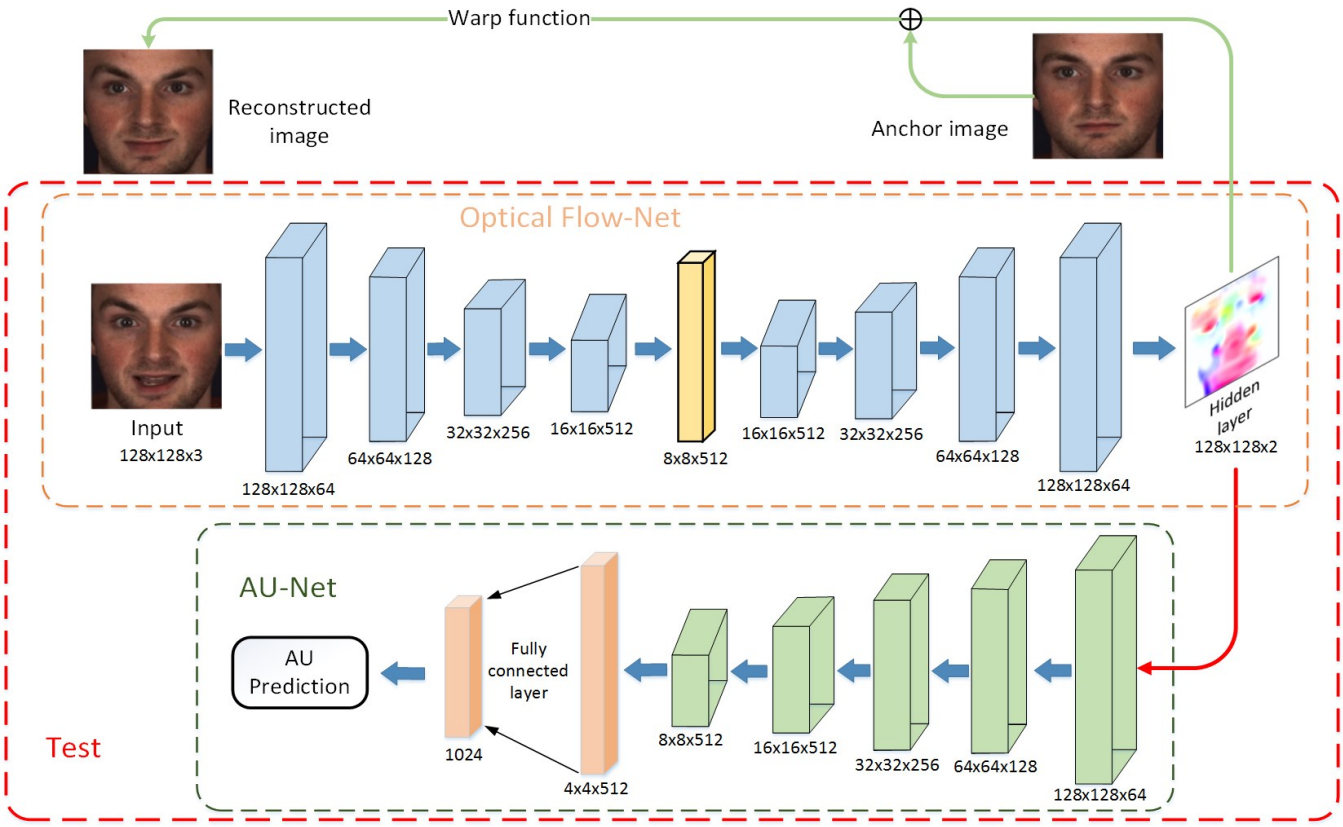

Fig. 2. Framework of our proposed method, which contains an OF-Net and an AU-Net. The OF-Net is designed to estimate the optical flow, which uses unsupervised training that reverses the optical flow and anchor image back to the input image. The AU-Net learns from the optical flow for AU detection. Both the OF-Net and the AU-Net are trained jointly. During training, an image pair is needed to learn the hidden optical flow layer (visualized in color), while only a single input image is needed for testing. This figure is best viewed in color.

defining a new cost function based on the pixel difference between the input image and reconstructed image, and also smoothness constraint, so the model can be trained in an unsupervised way. However, the proposed method failed to outperform the classical FlowNet and FlowNet2. Zhu et al.[36] tried to add more constraints to the unsupervised loss functions; *i.e.*, pixel-wise loss, smoothness loss and SSIM loss. Meister et al.[19] proposed an unsupervised bidirectional end-to-end framework to predict optical flow. By performing a second pass with the two input image pairs, the occlusion problem would also be considered explicitly. To leverage the gap between supervised methods and unsupervised methods for optical flow estimation, Wang et al.[26] proposed an occlusion aware method, which was able to model occlusion explicitly and learn large motions. Zhu et al.[36] combined the unsupervised optical flow learning with action recognition by implicitly estimating optical flow between adjacent frames.

The most related work is [36], but the difference lies in two-fold: first, the application domains are different, as [36] is to recognize actions from video, while our task aims to detect AUs from static image. Second, the learned optical flow (OF) is more preferable to our application domain. As mentioned in [36], the OF learned by MotionNet has lots of background noise, thus performs worse than the traditional OF-based estimation method. However, our method focuses on the AU-related facial action which is by nature correlated with optical flow as seen in Fig.3, thus it performs much better than the traditional OF-based estimation methods (Table IV).

III. Method

Our method contains two parts: optical flow estimation network (OF-Net) and AU detection network (AU-Net). OF-Net is trained to estimate the optical flow between an input image I_1 and an anchor image I_2, and the training process is unsupervised. The AU-Net is designed to detect AU occurrence from the hidden layer (optical flow). We also compared different strategies to train the model: training the OF-Net and AU-Net jointly and separately.

A. Optical Flow Network

Optical flow estimation can be treated as an image reconstruction problem[11]. Basically, given an image pair $< I_1, I_2 >$, we want to estimate optical flow $F = (F_u, F_v)$ from image I_1. Then we hope to reconstruct the image $I_1^{'}$ using the predicted optical flow F and image I_2, *i.e.*, $I_1^{'} = \mathcal{T}(I_2, F)$, where $\mathcal{T}$ is the inverse warping function. By minimizing the pixel-wise difference between I_1 and $I_1^{'}$, the OF-Net can then be trained in an unsupervised way.

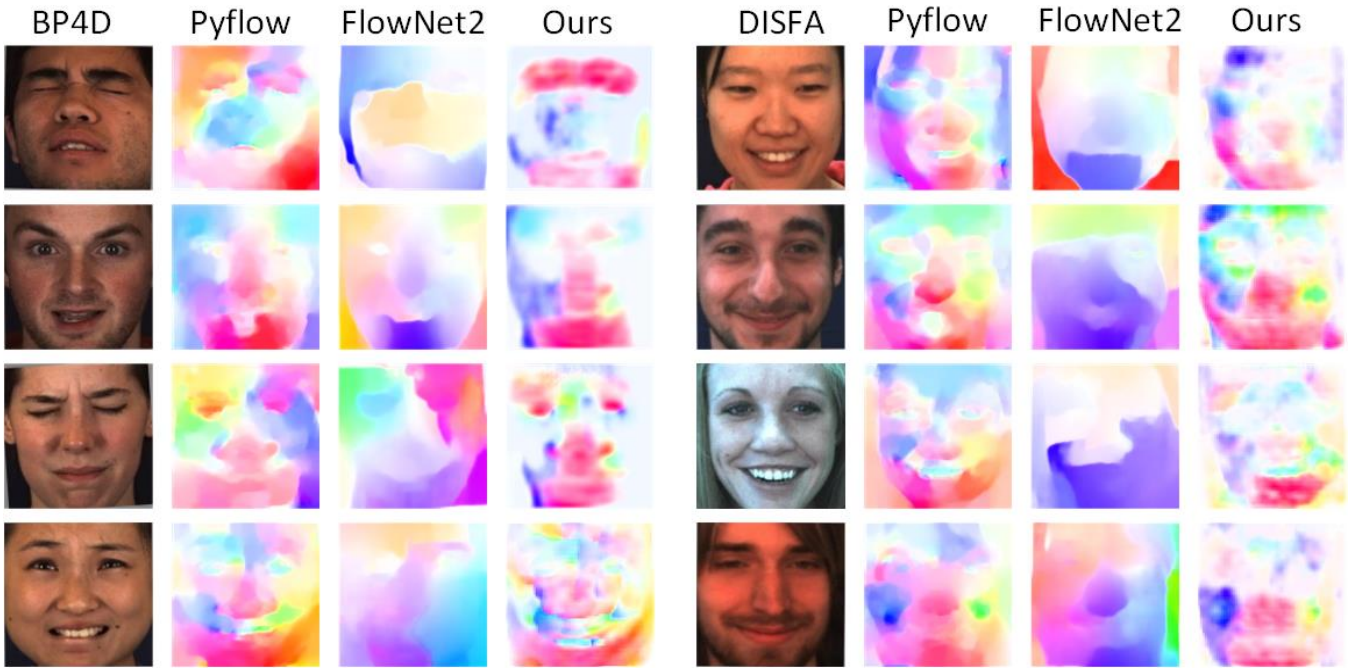

Fig. 3. Visual comparison of estimated optical flows from Pyflow, FlowNet2 and ours. Left: optical flow images on BP4D dataset. Right: optical flow images on DISFA dataset. This figure is best viewed in color.

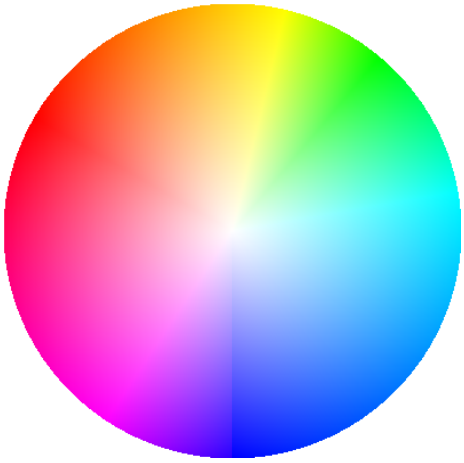

Fig. 4. Optical flow color coding. The central point is the original point, and the other colors represent the displacement (orientation and magnitude) of every other point to the original point. This figure is best viewed in color.

Three loss functions that have been proved effective in optical flow learning [1] [11] [19] [36], are applied, and one extra AU loss is also added, which is quite unique for an optical flow estimation task. Their definitions are as follows.

- The pixel-wise loss function describes the pixel level difference between I_1 and reconstructed $I_1^{'}$:

$$L_{pixel} = \frac{1}{N}\sum_{x,y}^{N}\rho\big(I_1(x,y) - I_1^{'}(x,y)\big) \tag{1}$$

$$I_1^{'} = \mathcal{T}(I_2, F) \tag{2}$$

N is the total number of input images. The inverse warping function $\mathcal{T}$ transforms the anchor image I_2 to $I_1^{'}$ based on $(F_u.F_v)$ of the estimated optical flow F. And $\rho(\cdot)$ is the generalized Charbonnier penalty function: $\rho(x) = (x^2 + \epsilon^2)^{\alpha}$.

- The smoothness loss function is a first order smoothness constraint on the optical flow, which encourages collinearity of adjacent flows and has been proven as an effective method of regularization in [19].

$$L_{smooth} = \rho\big(\nabla F_u\big) + \rho\big(\nabla F_v\big) \tag{3}$$

$\nabla F_u = [\partial F_u/\partial x, \partial F_u/\partial y]^T$ are the gradients on optical flow F_u in the horizontal and vertical directions. And so is $\nabla F_v = [\partial F_v/\partial x, \partial F_v/\partial y]^T$.

- The structure dissimilarity loss function utilizes the structure similarity index (SSIM) to measure the dissimilarity of two images.

$$L_{dssim} = \frac{1}{N}\sum_{i}^{N}\big(1 - SSIM(I_1^i, I_1^{'i})\big) \tag{4}$$

The experiment in [36] proved that the L_{dssim} loss helped to generate much clearer motion boundaries.

B. AU-Net

AUs are the contraction or relaxation of one or more facial muscles, which reflect instant changes in facial appearance. Similarly, optical flow describes the motion between two image frames which are taken at times t and $t + \Delta t$ at every pixel position. If a frame at time t is a neutral face image, and another frame at time $t + \Delta t$ is the face image after one or more facial muscles movement, then the optical flow extracted from the frames also contains AU-related

information. The AU-Net accepts the optical flow as input, and outputs the occurrence of one or more AUs. AUs loss measures the recognition error in AUs detection, which is a multiple labels classification problem, as multiple AUs can co-occur at the same time. The loss is calculated as:

$$L_{au} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} Y_{ic}\log\widehat{Y_{ic}} + (1-Y_{ic})\log(1-\widehat{Y_{ic}}) \quad (5)$$

Where C is the number of AUs, N is the number of samples, Y is ground truth labels $Y \in \{0,1\}^{N\times C}$, and $\widehat{Y}$ is the prediction of AUs $\widehat{Y} \in [0,1]^{N\times C}$.

The total loss is a weighted sum of the above-mentioned losses:

$$L_{total} = \lambda_1 \cdot L_{pixel} + \lambda_2 \cdot L_{smooth} + \lambda_3 \cdot L_{dssim} + \lambda_4 \cdot L_{au} \quad (6)$$

Where $\lambda_1, \lambda_2, \lambda_3$ and λ_4 are the weights, controlling the relative importance of different loss terms.

C. Train them jointly or individually

There are two approaches to train the model. The first is to train the model as a two-stage process, where the optical flow estimation and AUs detection are performed separately. The OF-Net is first trained using all the loss functions except L_{au}, and then all the estimated optical flow values are saved as input for the AU-Net, which is optimized based on a single L_{au} loss function. The other approach is to train both of them jointly, which has two advantages over the two-stage approach.

- First, by training them jointly, our model is end-to-end trainable, which makes the computation more efficient.
- Second, and the most important, OF-Net and AU-Net benefit each other. With the extra constraint from the AU loss, the OF-Net will learn to focus on appearance changes related to AUs, and in return, achieves better performance for AU detection. For example, AU2 describes the raiser of the outer brow, AU9 describes the wrinkler of the nose and AU17 describes the raiser of the chin; all of them provide some semantic descriptions about facial appearance changes, which will be used by the OF-Net.

IV. Experiment

Our proposed approach is evaluated on two widely used datasets for AU detection: BP4D[30] and DISFA[18] . We report the results using two metrics, *i.e.*, Accuracy (denoted as ACC) and F1-score as they are widely used in AU detection. F1-score is the harmonic mean of the precision and recall, while ACC reflects the relationship between true positives and false positives. 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. Implementation details

Some AUs have more samples than others, as shown in the second row of Table II, where AU25 is almost 7 times greater than AU2. In order to balance the training data, we manually repeated 4 to 7 times for the less occurring AUs. The new data distribution is shown in the third row of Table II.

Face images are first registered to the size of 140x140 based on landmarks. For data augmentation, each face is randomly cropped into 128x128, rotated by a random angle between -15^o and 15^o, and randomly flipped in horizontal direction. During training, the first frame of each sequence is usually selected as anchor image. Unless the first frame is obviously not a neutral face, then we manually picked the first available neutral face in the sequence as anchor image. This procedure is not needed for testing.

Our model is initialized with a learning rate of 0.0002 and a momentum of 0.9. AU loss is ignored during the first several epochs, then added to the total loss, and trained together with the OF-Net and AU-Net. Our model is trained for 30 epochs, and $\lambda_1, \lambda_2, \lambda_3$and λ_4 are empirically set to 30, 0.16, 1 and 30 respectively. All experiments are implemented using TensorFlow and performed on four NVIDIA Geoforce 1080Ti GPUs.

B. Visual comparisons of estimated optical flow

We compare the estimated optical flows from the OF-Net with the traditional optical flow: Pyflow [17] and the recent state-of-the-art approach: FlowNet2 [9] on both datasets. The visualizations of the flow $[F_u, F_v]$ are shown in Fig. 3, where different colors, representing the orientation and magnitude of optical flow, are coding in Fig.4 . FlowNet2 captures less information than others, because FlowNet2, as a supervised learning method, is trained on a synthesized dataset, which makes it not work well on the face datasets. Both Pyflow and our method are able to capture optical flow with more details, but, Pyflow relies on the input image pairs by calculating the correspondence between them; while our method is an unsupervised end-to-end learning framework, which is able to predict optical flow from a single input image.

C. Evaluation and results

BP4D: there are 328 2D and 3D videos collected from 41 subjects (23 females, 18 males) under eight different tasks. The most expressive frames are AU labeled for occurrence, which resulted in a dataset of 140,000 images with AU labels. The images are split into 3 folds, and the subjects in any two subsets are mutually exclusive. Then, a 3-fold subject-independent cross validation is performed.

Our method is compared to JPML[34], DRML[35] , EAC-net[14] , ATF[33] and a baseline: fine-tuned VGG (FVGG) using both F1 score and accuracy except ATF, which only reports F1 score. The average F1 score and accuracy of the 3 runs are reported in Table I. As we can see, our proposed method shows better accuracy than others. At the same time, our method significantly improves the performance on F1 score, showing around 4% improvement over that of the state-of-the-art method. Note that, for recognition rates, our method performed the best on *AU1, AU2, AU4, AU7, AU12, AU15, AU17 and AU23* and the second best on *AU6, AU10 and AU24*.

TABLE I

F1 SCORE AND ACCURACY IN TERMS OF 12 AUS ON BP4D DATASET. BRACKETED AND BOLD NUMBERS INDICATE THE BEST PERFORMANCE; BOLD NUMBERS INDICATE THE SECOND BEST.

	F1 Score						Accuracy				
AU	JPML[34]	DRML[35]	FVGG	EAC-net[14]	ATF[33]	**Ours**	JPML[34]	DRML[35]	FVGG	EAC-net[14]	**our**
1	32.6	36.4	27.8	39.0	**39.2**	**[50.8]**	40.7	55.7	27.2	**68.9**	**[75.9]**
2	25.6	**41.8**	27.6	35.2	35.2	**[45.3]**	42.1	54.5	56.0	**73.9**	**[79.7]**
4	37.4	43.0	18.3	**48.6**	45.9	**[56.6]**	46.2	58.8	**80.5**	78.1	**[80.5]**
6	42.3	55.0	69.7	**[76.1]**	71.6	**75.9**	40.0	56.6	72.3	**[78.5]**	**74.1**
7	50.5	67.0	69.1	**72.9**	71.9	**[75.9]**	50.0	61.0	64.1	**69.0**	**[69.3]**
10	72.2	66.3	78.1	**[81.9]**	79.0	**80.9**	**75.2**	53.6	72.4	**[77.6]**	73.0
12	74.1	65.8	63.2	**86.2**	83.7	**[88.4]**	60.5	60.8	69.1	**84.6**	**[85.3]**
14	**[65.7]**	54.1	36.4	58.8	**65.5**	63.4	53.6	57.0	52.8	**[60.6]**	**58.0**
15	38.1	33.2	26.1	**37.5**	33.8	**[41.6]**	50.1	56.2	67.4	**[78.1]**	**77.9**
17	40.0	48.0	50.7	59.1	**60.0**	**[60.6]**	42.5	50.0	61.2	**[70.6]**	**68.3**
23	30.4	31.7	22.8	35.9	**37.3**	**[39.1]**	51.9	53.9	72.2	**81.0**	**[81.4]**
24	42.3	30.0	35.9	35.8	**[41.8]**	**37.8**	53.2	53.9	77.0	**82.4**	**[83.8]**
Avg	45.9	48.3	43.8	**55.9**	55.4	**[59.7]**	50.5	56.0	64.4	**75.2**	**[75.6]**

TABLE II

DATA BALANCE ON DISFA FOR AU OCCURRENCE

AU	1	2	4	6	9	12	25	26
Before	0.05	0.04	0.15	0.12	0.04	0.18	0.27	0.15
After	0.12	0.11	0.16	0.10	0.10	0.11	0.20	0.11

DISFA: there are 27 subjects involved in this dataset. Each of them was recorded in two videos using two cameras (left camera and right camera). Sixty six facial landmarks and AU intensity (0-5 scale) are provided. The frames collected under left camera, with AU intensity greater than 0 of 8 AUs are selected, resulting in a 130,000 images dataset. The dataset was further split into 3 subject-independent subsets. We followed the protocols used in [14], fine-tuned the pre-trained model from BP4D on DISFA, and performed 3 folds cross validation.

The average F1 score and accuracy on recognizing 8 AUs over 3 runs is shown in Table III. Note that, before our method, EAC-net [14] was the state-of-the-art, which significantly improved the performance of F1 score from 40.2% (FVGG) to 48.5%, and better than other methods, *i.e.*, APL[35] and DRML[35]. ATF[33] was published in 2018, and only showed a slight improvement on the F1 score . Our method achieves comparable performance on accuracy and the highest F1 score, showing 4.5% improvement than that of the state-of-the-art methods. It is also worth mentioning that our model shows 21.3% improvement of recognizing AU26 (*Jaw Drop*).

D. Comparison to optical flow + CNN approach

To better understand that our model has the capability of estimating temporal information from a static image, we compared our method to the traditional optical flow based method. For a fair comparison, the same AU-Net is used in OF-Net + AU-Net *vs* optical flow + AU-Net. To calculate the traditional optical flow, an input image pair is needed. Here, we apply the Pyflow [17], which is a dense optical flow calculation method with python wrapper, to calculate the optical flow from the image pairs. The result is reported in Table IV. Note that, our model is able to estimate the optical flow from a single frame, while at least two frames are needed for Pyflow. Our method shows 4.1% and 4.3% improvement of F1 score, 2.6% and 7.5% improvement of accuracy on BP4D and DISFA respectively. The improved performance demonstrates the two merits of our proposed method: first, unlike the traditional optical flow based methods that rely on a reference frame, our method is able to capture the temporal information from a single input image; second, the estimated optical flow captures the AU-related facial appearance changes, which make it a better representation for the AU detection task.

E. Training separately or jointly

Here, we further compare our method under different settings. Training the OF-Net and AU-Net separately means the OF-Net is trained first using three losses: L_{pixel}, L_{smooth} and L_{dssim}; then the estimated optical flows from the OF-Net are stored and used as input for the AU-Net, which is trained by minimizing the AU loss: L_{au}. Under this training setting, the AU-Net relies on the estimated optical flow from the OF-Net, and the AU loss cannot be back-propagated to the OF-Net. On the other hand, by training them jointly, the total loss function contains both the $L_{pixel}, L_{smooth}, L_{dssim}$ and L_{au}, and the AU loss is able to back-propagate to the OF-Net.

As shown in Table V, the performance of training them jointly have 4.6%, 6.5% improvement of F1 score and 4.3%, 4.8% improvement of accuracy for BP4D and DISFA respectively. By comparing Table IV and Table V, we find that even under the separate training setting, our model still has a comparable performance with the traditional optical flow + CNN method, but why does training them jointly has such a performance improvement? Our analysis is, as shown in Fig. 3, both Pyflow and our method are able to capture facial appearance changes. However Pyflow is based on pixel-wise correspondence, and not every pixel change is

TABLE III

F1 SCORE AND ACCURACY IN TERMS OF 8 AUS ON DISFA DATASET. BRACKETED AND BOLD NUMBERS INDICATE THE BEST PERFORMANCE; BOLD NUMBERS INDICATE THE SECOND BEST.

	F1 Score						Accuracy				
AU	APL[35]	DRML[35]	FVGG	EAC-net[14]	ATF[33]	**Ours**	APL[35]	DRML[35]	FVGG	EAC-net[14]	**our**
1	11.4	17.3	32.5	**41.5**	**[45.2]**	30.9	32.7	53.3	82.7	**[85.6]**	**84.7**
2	12.0	17.7	24.3	26.4	**[39.7]**	**34.7**	27.8	53.2	83.6	**84.9**	**[90.6]**
4	30.1	37.4	61.0	**[66.4]**	47.1	**63.9**	37.9	60.0	**74.1**	**[79.1]**	72.1
6	12.4	29.0	34.2	**[50.7]**	**48.6**	44.5	13.6	54.9	64.2	**69.1**	**[72.8]**
9	10.1	10.7	1.67	**[80.5]**	**32.0**	31.9	64.4	51.5	87.1	**[88.1]**	**87.9**
12	65.9	37.7	72.1	**[89.3]**	55.0	**78.3**	**94.2**	54.6	67.8	**[90.0]**	82.6
25	21.4	38.5	**87.3**	**[88.9]**	86.4	84.7	50.4	45.6	78.6	**[80.5]**	**78.8**
26	26.9	20.1	7.1	15.6	**39.2**	**[60.5]**	47.1	45.3	61.7	**64.8**	**[71.4]**
Avg	23.8	26.7	40.2	48.5	**49.2**	**[53.7]**	46.0	52.3	74.9	**[80.6]**	**80.1**

TABLE IV

COMPARISON OF OUR METHOD TO THE TRADITIONAL OPTICAL FLOW + CNN METHOD

	BP4D		DISFA	
Method	F1 Score	Accuracy	F1 Score	Accuracy
Optical flow + CNN	55.6	73.0	49.4	72.6
Ours	**59.7**	**75.6**	**53.7**	**80.1**

TABLE V

PERFORMANCE COMPARISON UNDER DIFFERENT SETTING: TRAIN THEM SEPARATELY OR JOINTLY.

	BP4D		DISFA	
Method	F1 Score	Accuracy	F1 Score	Accuracy
Training Separately	55.1	71.3	47.2	75.3
Training Jointly	**59.7**	**75.6**	**53.7**	**80.1**

equally useful for AU detection; while our method is trained with an extra L_{au} loss, which provides a semantic constraint for the OF-Net. For example, AU1 (Inner Brow Raiser), AU4 (Brow Lowerer), and AU16(Lower Lip Depressor), which help the OF-Net to focus on the AU-related motions, ignore unrelated areas, and in return, have better performance on AU detection.

V. CONCLUSION

In this paper, we have proposed a novel hidden optical flow learning approach for AU detection. A hidden optical flow layer is proposed to concatenate the OF-Net and AU-Net. The OF-Net is designed to estimate facial appearance movement from a single input image using unsupervised learning. In addition, the AU-Net utilizes the estimated optical flows as input and detects the AU occurrence.

We evaluated our proposed method on BP4D and DISFA datasets. The results showed that our approach achieved better performance on AU detection than the state-of-the-art methods. We also compared different settings: training the OF-Net and AU-Net separately and jointly. We found that by training both networks jointly, the OF-Net learned to generate more meaningful optical flow by the semantic constraints provided by AU loss, and in return, the AU-Net achieved better performance. To prove the usefulness of the learned optical flow from a single image, we fed both the estimated optical flow from a single image and calculated optical flow from image pairs into the same CNN model, and the results proved that the OF-Net is able to capture the motion information from a single input image, and even works better than the traditional image pair based optical flow methods.

We believe that the learned optical flow is not only for detection of AU occurrence but also for estimation of AU intensity. We will extend this work to AU intensity estimation (e.g. with BP4D+[32]) in the future.

VI. ACKNOWLEDGMENTS

The material is based upon the work supported in part by the National Science Foundation under grants CNS-1629898 and CNS-1205664.

REFERENCES

[1] A. Ahmadi and I. Patras. Unsupervised convolutional neural networks for motion estimation. In *Image Processing (ICIP), 2016 IEEE International Conference on*, pages 1629–1633. IEEE, 2016.

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

[3] 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.

[4] 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.

[5] W. Chu, F. De la Torre, and J. F. Cohn. Learning spatial and temporal cues for multi-label facial action unit detection. In *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, pages 25–32, May 2017.

[6] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2758–2766, 2015.

[7] P. Ekman and W. V. Friesen. *Facial action coding system: Investigator's guide*. Consulting Psychologists Press, 1978.

[8] I. A. Essa and A. P. Pentland. Coding, analysis, interpretation, and recognition of facial expressions. *IEEE transactions on pattern analysis and machine intelligence*, 19(7):757–763, 1997.

[9] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *IEEE conference on computer vision and pattern recognition (CVPR)*, volume 2, page 6, 2017.

[10] S. Jaiswal and M. Valstar. Deep learning the dynamic appearance and shape of facial action units. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–8. IEEE, 2016.

[11] J. Y. Jason, A. W. Harley, and K. G. Derpanis. Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness. In *European Conference on Computer Vision*, pages 3–10. Springer, 2016.

[12] 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.

[13] W. Li, F. Abtahi, and Z. Zhu. Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 6766–6775. IEEE, 2017.

[14] W. Li, F. Abtahi, Z. Zhu, and L. Yin. Eac-net: A region-based deep enhancing and cropping approach for facial action unit detection. In *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, pages 103–110. IEEE, 2017.

[15] J. J.-J. Lien, T. Kanade, J. F. Cohn, and C.-C. Li. Detection, tracking, and classification of action units in facial expression. *Robotics and Autonomous Systems*, 31(3):131–146, 2000.

[16] A. C. Lints-Martindale, T. Hadjistavropoulos, B. Barber, and S. J. Gibson. A psychophysical investigation of the facial action coding system as an index of pain variability among older adults with and without alzheimer's disease. *Pain Medicine*, 8(8):678–689, 2007.

[17] C. Liu et al. *Beyond pixels: exploring new representations and applications for motion analysis*. PhD thesis, Massachusetts Institute of Technology, 2009.

[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] S. Meister, J. Hur, and S. Roth. Unflow: Unsupervised learning of optical flow with a bidirectional census loss. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[20] M. Pantic and M. S. Bartlett. Machine analysis of facial expressions. In *Face recognition*. InTech, 2007.

[21] M. Pantic and L. J. Rothkrantz. Facial action recognition for facial expression analysis from static face images. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 34(3):1449–1461, 2004.

[22] A. Ranjan and M. J. Black. Optical flow estimation using a spatial pyramid network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, page 2. IEEE, 2017.

[23] L. I. Reed, M. A. Sayette, and J. F. Cohn. Impact of depression on response to comedy: A dynamic facial coding analysis. *Journal of abnormal psychology*, 116(4):804, 2007.

[24] 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.

[25] 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.

[26] Y. Wang, Y. Yang, Z. Yang, L. Zhao, and W. Xu. Occlusion aware unsupervised learning of optical flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4884–4893, 2018.

[27] Z. Wang, Y. Li, S. Wang, and Q. Ji. Capturing global semantic relationships for facial action unit recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3304–3311, 2013.

[28] Y.-T. Wu, T. Kanade, J. Cohn, and C.-C. Li. Optical flow estimation using wavelet motion model. In *Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271)*, pages 992–998. IEEE, 1998.

[29] Y. Yacoob and L. S. Davis. Recognizing human facial expressions from long image sequences using optical flow. *IEEE Transactions on pattern analysis and machine intelligence*, 18(6):636–642, 1996.

[30] 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.

[31] 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.

[32] Z. Zhang, J. M. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang, J. F. Cohn, Q. Ji, and L. Yin. Multimodal spontaneous emotion corpus for human behavior analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3438–3446, 2016.

[33] Z. Zhang, S. Zhai, and L. Yin. Identity-based adversarial training of deep cnns for facial action unit recognition. *British Machine Vision Conference*, 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.

[36] Y. Zhu, Z. Lan, S. Newsam, and A. G. Hauptmann. Hidden two-stream convolutional networks for action recognition. *ACCV*, 2018.