FG-Net: Facial Action Unit Detection with Generalizable Pyramidal Features

Yufeng Yin[1]*, Di Chang[1], Guoxian Song[2], Shen Sang[2], Tiancheng Zhi[2],
Jing Liu[2], Linjie Luo[2], Mohammad Soleymani[1]
[1] University of Southern California, [2] ByteDance
{yufengy, dichang, msoleyma}@usc.edu,
{guoxiansong, shen.sang, tiancheng.zhi, jing.liu, linjie.luo}@bytedance.com

Abstract

Automatic detection of facial Action Units (AUs) allows for objective facial expression analysis. Due to the high cost of AU labeling and the limited size of existing benchmarks, previous AU detection methods tend to overfit the dataset, resulting in a significant performance loss when evaluated across corpora. To address this problem, we propose ***FG-Net*** *for generalizable facial action unit detection. Specifically, FG-Net extracts feature maps from a StyleGAN2 model pre-trained on a large and diverse face image dataset. Then, these features are used to detect AUs with a Pyramid CNN Interpreter, making the training efficient and capturing essential local features. The proposed FG-Net achieves a strong generalization ability for heatmap-based AU detection thanks to the generalizable and semantic-rich features extracted from the pre-trained generative model. Extensive experiments are conducted to evaluate within- and cross-corpus AU detection with the widely-used DISFA and BP4D datasets. Compared with the state-of-the-art, the proposed method achieves superior cross-domain performance while maintaining competitive within-domain performance. In addition, FG-Net is data-efficient and achieves competitive performance even when trained on 1000 samples. Our code will be released at* https://github.com/ihp-lab/FG-Net

1. Introduction

Automatic detection of facial action units is a fundamental block for objective facial expression analysis [6]. Manual annotations for facial action units are cumbersome and costly, as they require trained coders to label each frame individually. Common AU datasets, *i.e.*, DISFA [22] and BP4D [33], only contain a limited number of subjects (27 and 41 subjects respectively). Recent methods for AU detection [21, 25, 36] focus on deep representation learning, requiring a large number of samples. Existing AU detection methods are often evaluated with within-domain cross-validation, with training and testing data from the same dataset, and the generalization to other datasets (model trained and tested on different datasets) has not been widely investigated. As within-domain performance can be due to **overfitting**, cross-corpus performance can suffer **a considerable loss** (see comparisons in Figure 1).

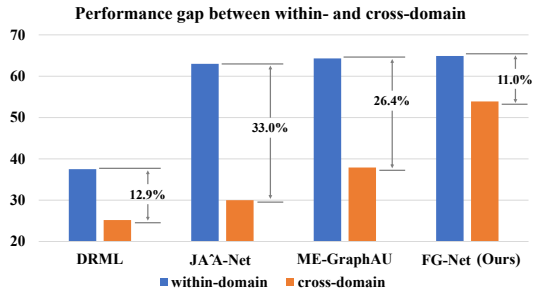


Figure 1. Performance (F1 score ↑) gap between the within- and cross-domain AU detection for DRML [36], JÂA-Net [25], ME-GraphAU [21], and the proposed FG-Net. The within-domain performance is averaged between DISFA and BP4D, while the cross-domain performance is averaged between BP4D to DISFA and DISFA to BP4D. The proposed FG-Net has the highest cross-domain performance, thus, superior generalization ability.

In the field of semantic segmentation, recent studies [2, 34] leverage a well-trained generative model to synthesize image-annotation pairs from only a few labeled examples (around 30 training samples). They show that the intermediate features of generative models exhibit semantic-rich representations that are well-suited for pixel-wise segmentation tasks in a few-shot manner. Li *et al.* [16] showcase extreme out-of-domain generalization ability from such approaches. However, to the best of our knowledge, no existing work adapts such architectures to AU detection, potentially due to the following limitations: (i) the high dimensionality of the pixel-wise features results in inefficient training, and (ii) inference with pixel-level features lacks

*Work done during internship at ByteDance.

the information from the nearby regions, which is crucial to AU detection [25, 36].

In this paper, inspired by the success of GAN features in semantic segmentation, we propose **FG-Net**, a facial action unit detection method that can better generalize across domains. The general idea of FG-Net is to extract the generalizable and semantic-rich deep representations from a well-trained generative model (see Figure 2). Specifically, FG-Net first encodes and decodes the input image with a StyleGAN2 encoder (pSp) [24], and a StyleGAN2 generator [14], trained on the FFHQ dataset [13]. Then, FG-Net extracts feature maps from the generator during decoding. To take advantage of the informative pixel-wise representations from the generator, FG-Net detects the AUs through heatmap regression. We propose a Pyramid CNN Interpreter which incorporates the multi-resolution feature maps in a hierarchical manner. The proposed module makes the training efficient and captures essential information from nearby regions. Thanks to the powerful features from the generative model pre-trained on a large and diverse facial image dataset, the proposed FG-Net obtains a strong generalization ability and data efficiency for AU detection.

To demonstrate the effectiveness of our proposed method, we conduct extensive experiments with the widely-used DISFA [22] and BP4D [33] for AU detection. The results show that the proposed FG-Net method has a strong generalization ability and achieves state-of-the-art cross-domain performance (see Figure 1). In addition, FG-Net achieves comparable or superior within-domain performance to the existing methods. Finally, we showcase that FG-Net is a data-efficient approach. With only 100 training samples, it can achieve decent performance.

Our major contributions are as follows. (i) We propose FG-Net, a data-efficient method for generalizable facial action unit detection. To the best of our knowledge, we are the first to utilize StyleGAN model features for AU detection. (ii) Extensive experiments on the widely-used DISFA and BP4D datasets show that FG-Net has a strong generalization ability for heatmap-based AU detection achieving superior cross-domain performance and maintaining competitive within-domain performance compared to the state-of-the-art. (iii) FG-Net is data-efficient. The performance of FG-Net trained on 1k samples is close to the whole set ($\sim$100k).

2. Related Work

Facial Action Unit Detection. A facial action unit is an indicator of activation of an individual or a group of muscles, *e.g.*, cheek raiser (AU6). AUs are formalized by Paul Ekman in Facial Action Coding System (FACS) [6]. Previous studies explore attention mechanism [12, 25, 28] or self-supervised learning [4] to get discriminative representations for AU detection.

Shao *et al*. [25] propose JÂA-Net for joint AU detection and face alignment. JÂA-Net uses adaptive attention learning to refine the attention map for each AU. Jacob *et al*. [12] combine transformer-based architectures with region of interest (ROI) attention module, per-AU embeddings, and correlation module to capture relationships between different AUs. Chang *et al*. [4] propose a knowledge-driven self-supervised representation learning framework. AU labeling rules are leveraged to design facial partition manners and determine correlations between facial regions.

Recent work on AU detection use graph neural networks [21, 27, 35]. Zhang *et al*. [35] utilize a heatmap regression-based approach for AU detection. The ground-truth heatmaps are defined based on the ROI for each AU. Besides, the authors utilize graph convolution for feature refinement. Luo *et al*. [21] propose an AU relationship modeling approach that learns a unique graph to explicitly describe the relationship between each pair of AUs of the target facial display. Previous studies on AU detection achieve promising within-domain performance. However, the generalization ability, *i.e.*, cross-domain performance, for AU detection has not been widely investigated.

Ertugrul *et al*. [7, 8] demonstrate that the deep-learning-based AU detectors achieve poor cross-domain performance due to the variations in the cameras, environments, and subjects. Tu *et al*. [29] propose Identity-Aware Facial Action Unit Detection (IdenNet). IdenNet is jointly trained by AU detection and face clustering datasets that contain numerous subjects to improve the model's generalization ability. Yin *et al*. [32] propose to use domain adaptation and self-supervised patch localization to improve the cross-corpora performance for AU detection. However, this method requires data from the target domain for domain adaptation. Hernandez *et al*. [10] conduct an in-depth analysis of performance differences across subjects, genders, skin types, and databases. To address this gap, they propose deep face normalization (DeepFN) that transfers the facial expressions of different people onto a common facial template.

In this paper, without using any data from the target domain, we improve the cross-corpus AU detection with the semantic-rich features from a generative model trained on a large-scale and diverse dataset.

Face Understanding with Generative Models. Generative models provide an estimate of the distribution of training samples [3]. Prior work utilizing generative models for face understanding has mainly focused on semantic segmentation [2, 16, 34] and landmark detection [31, 34].

Zhang *et al*. [34] introduce DatasetGAN, an automatic procedure to generate massive datasets of high-quality semantically segmented images requiring minimal human effort. The authors show how the GAN latent code can be decoded to produce a semantic segmentation of the image and allow the decoder to be trained with only a few labeled ex-

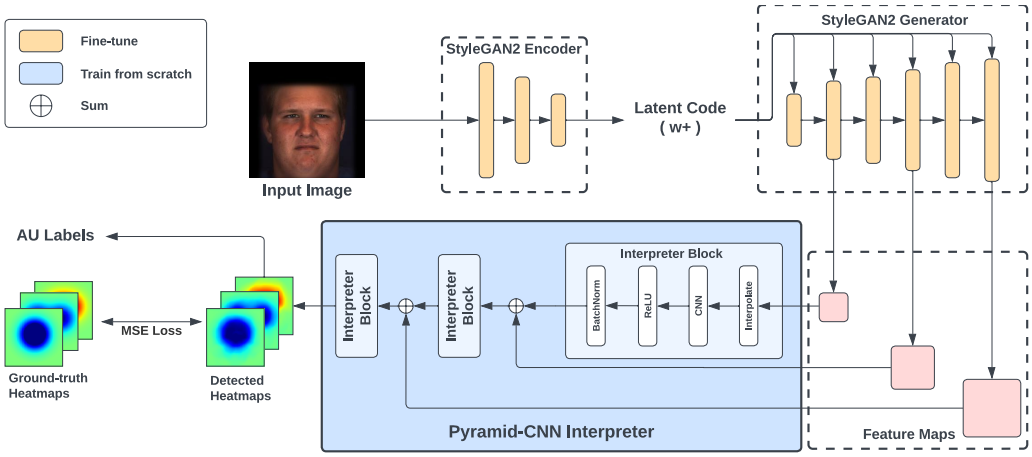


Figure 2. Overview of our proposed pipeline. FG-Net first encodes the input image into a latent code using a StyleGAN2 encoder (e.g. pSp [24] here). In the decoding stage [14], we extract the intermediate multi-resolution feature maps and pass them through our Pyramid CNN Interpreter to detect AUs coded in the form of heatmaps. Mean Squared Error (MSE) loss is used for optimization between the ground truth and predicted heatmaps.

amples. Baranchuk *et al*. [2] demonstrate that feature maps from diffusion models [5] can capture the semantic information and appear to be excellent pixel-wise representations. Li *et al*. [16] propose semanticGAN, a generative adversarial network that captures the joint image-label distribution. The proposed semanticGAN showcases an extreme out-of-domain generalization ability, such as transferring from real faces to paintings, sculptures, and even cartoons and animal faces. Xu *et al*. [31] consider the pre-trained StyleGAN generator as a learned loss function and train a hierarchical encoder to get visual representations, namely GH-Feat, for input images. GH-Feat has strong transferability to both generative and discriminative tasks.

Previous studies show that the hidden states from the generative models are powerful representations for face understanding. However, to the best of our knowledge, no existing work adapts such architectures to AU detection. Zhang *et al*. [34] and Baranchuk *et al*. [2] extract pixel-wise features and treat each pixel as a training sample, leading to extreme inefficiency due to the per-sample computational overhead. More importantly, inference with singe-pixel features lacks the inductive bias (local features), crucial to AU detection shown in the previous studies [25,36]. In addition, semanticGAN [16] has to encode the input image to the latent space in an optimization-based manner for inference, which is extremely time-consuming. Thus semanticGAN can be only tested with a few samples. The limitation of this approach does not allow for training or testing with larger datasets. In this paper, we propose a Pyramid CNN Interpreter to detect the heatmaps, representing activated AUs, for AU detection, which is more efficient and can capture both local and global information. GH-Feat [31] is the closest method to ours that extracts latent code representations

Figure 3. Visualizations of the ROI centers for DISFA (left) and BP4D (right). AU indices are labeled above or below.

from generative models. The major differences between GH-Feat and our method are: (i) GH-Feat extracts the 1D latent code features while FG-Net further decodes the latent codes to images and gets the 2D feature maps. (ii) GH-Feat is trained in a multi-stage manner while our method is end-to-end. GH-Feat utilizes the StyleGAN generator as a learned loss function and trains a hierarchical encoder and then uses this encoder to extract visual representations for downstream tasks. The whole pipeline requires more than 700 GPU hours due to the complicated training process while FG-Net only needs 10 GPU hours for training.

3. Methods

3.1. Problem Formulation

Facial Action Unit Detection. Given a video set S, for each frame $x \in S$, the goal is to detect the occurrence for each AU a_i $(i = 1, 2, ..., n)$ using function $\mathrm{F}(\cdot)$.

$$a_1, a_2, ..., a_n = \mathrm{F}(x), \tag{1}$$

where n is the number of AUs. $a_i = 1$ if the AU is active otherwise $a_i = 0$.

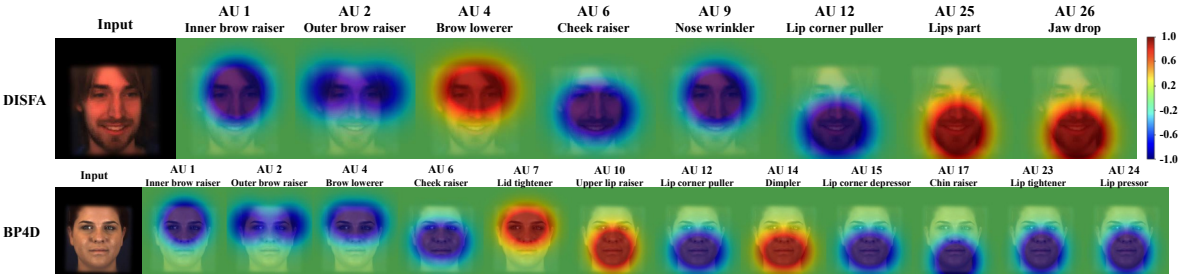


Figure 4. Visualization of the generated ground-truth heatmaps on DISFA (first row) and BP4D (second row). We generate one heatmap for every AU which has two Gaussian windows with the maximum values at the two ROI centers (see Figure 3). The peak value is either 1 (red, AU is active) or -1 (blue, AU is inactive).

3.2. Overview

Figure 2 illustrates an overview of the proposed FG-Net. FG-Net first encodes and decodes the input image with the pSp encoder [24] and the StyleGAN2 generator [14] pre-trained on the FFHQ dataset [13].

During the decoding, FG-Net extracts feature maps from the generator. Leveraging the features extracted from a generative model trained on a large-scale and diverse dataset, FG-Net offers a higher generalizability for AU detection.

To take advantage of the pixel-wise representations from the generator, FG-Net is designed to detect the AUs using a heatmap regression. To keep the training efficient and capture both local and global information, a Pyramid-CNN Interpreter is proposed to incorporate the multi-resolution feature maps in a hierarchical manner and detect the heatmaps representing facial action units.

3.3. Model

Prerequisites. Our proposed method is built on top of the StyleGAN2 generator [14]. The StyleGAN2 generator decodes a latent code $z \in \mathcal{Z}$ sampled from $\mathcal{N}(0, I)$ to an image. The latent code z is first mapped to a style code $w \in \mathcal{W}$ by a mapping function. Both z and w have 512 dimensions. There are k synthesis blocks (in practice $k = 9$) and each block has two convolution layers and one upsampling layer. Each convolution layer is followed by an adaptive instance normalization (AdaIN) layer [11] which is controlled by the style code w. However, for image inversion which encodes the images into the latent space, $\mathcal{W}$-space has limited expressiveness and thus can not fully reconstruct the input [30]. Therefore, prior works [1, 24] extend $\mathcal{W}$-space to $\mathcal{W}^+$-space where a different style code w is fed to each AdaIN layer. $\mathcal{W}^+$-space alleviates the reconstruction distortion. The dimension of $w^+ \in \mathcal{W}^+$ is 18×512.

Feature Extraction from StyleGAN2. To extract features from the StyleGAN2 generator, we first encode the input image to the latent space and then decode the latent code.

Prior work [16] encodes the input image in an optimization-based manner. Optimization-based methods iteratively optimize a reconstruction objective which is extremely time-consuming. Instead, we utilize the pSp encoder E [24] to encode the input image $x \in \mathcal{X}$ and get the latent code $w^+ \in \mathcal{W}^+$ via $w^+ = \mathrm{E}(x)$.

Despite the efficient encoding of the pSp encoder, the generator features may not capture the key facial features for AU detection. To address the problem, we fine-tune the encoder and the generator during training. Then, we decode the latent code with the StyleGAN2 generator G [14] to obtain image $x' = \mathrm{G}(w^+)$. During decoding, we extract the intermediate activations from the generator. To keep the training efficient, unlike the previous work [34] which extracts the outputs from all the AdaIN layers [11], we only extract the hidden states after the second AdaIN layer in each block. We denote the feature maps we get from the k blocks as $\{f_1, f_2, ..., f_k\} = \mathrm{G}'(w^+) = \mathrm{G}'(\mathrm{E}(x))$.

Heatmap Detection. The proposed method detects the AU occurrences in a heatmap regression-based approach. We generate the ground-truth heatmaps following the previous work [35]. We first define the Region of Interest (ROI) for each AU. We select two points on the face based on the most representative landmarks (see Figure 3, detailed positions are provided in the supplementary).

Then, we generate the ground-truth heatmaps with the definition of ROI. Figure 4 gives the visualization of the ground-truth heatmaps on DISFA and BP4D. Formally, given a face image $x \in \mathbb{R}^{w \times h \times 3}$, we generate n ground-truth heatmaps $m_1, m_2, ..., m_n \in \mathbb{R}^{w \times h}$ with the AU labels, where n is the number of AUs. Specifically, for heatmap m_i, we add two Gaussian windows g_i^1 and g_i^2 with the maximum value at the two ROI centers c_i^1 and c_i^2 following [35].

$$g_i^j(p) = \lambda_i \exp\left(-\frac{||p - c_i^j||_2^2}{2\sigma^2}\right), \quad j = 1, 2, \tag{2}$$

$$m_i(p) = g_i^1(p) + g_i^2(p). \tag{3}$$

where p is the pixel location ($p \in [1, w] \times [1, h]$). λ_i is the indicator denoting whether the i-th AU is active. $\lambda_i = 1$ if $a_i = 1$ otherwise $\lambda_i = -1$. σ is the standard derivation. We

clip the heatmaps into the range of $[-1, 1]$ to make sure the peak value is either 1 or -1.

After feature extraction from the generative models, prior work [2, 34] upsamples the features to the input resolution and concatenates them according to the channel dimension. Then, each pixel is treated as a training sample and a multi-layer perceptron (MLP) is trained to detect the semantic class. Simply upsampling and concatenating all the feature maps results in redundant and high dimensional features (in practice the number of channels is 6080), thus leading to inefficient training and inference. More importantly, using single-pixel features for inference lacks the spatial context from nearby regions, which is crucial to AU detection, as demonstrated in the previous studies [25, 36].

To address these problems, we propose a multi-scale Pyramid-CNN Interpreter H for heatmap-based AU detection which incorporates the multi-resolution feature maps in a hierarchical manner (see Figure 2). Specifically, the Pyramid-CNN Interpreter H contains k pyramid levels, where k is the number of feature maps extracted from the generator. In each pyramid level, the hidden states from the last pyramid level c_{i-1} are first summed with the feature map from the generator f_i and then passed through an interpreter block C_i. Each interpreter block consists of one Interpolate, one Convolution, one ReLU, and one BatchNorm layer. $m = c_k$ is the ultimate AU heatmap. specifically,

$$c_0 = 0, c_i = \mathrm{C}_i(c_{i-1} + f_i), i = 1, 2, ..., k, \quad (4)$$

$$m = c_k = \mathrm{H}(f_1, f_2, ..., f_k). \quad (5)$$

3.4. Training and Inference

Training. The learning objective is the Mean Squared Error (MSE) loss between the ground-truth heatmap m and the detected heatmap $\hat{m}$: $\mathcal{L} = ||m - \hat{m}||_2^2$.

Inference. For each detected heatmap $\hat{m}_i$, we sum up the whole heatmap. If the sum is greater than 0, the corresponding AU is active otherwise the AU is inactive.

4. Experiments and Discussions

4.1. Datasets

We select two publicly available datasets, *i.e.*, DISFA [22] and BP4D [33]. These two datasets are widely used for AU detection and are captured from different subjects with different backgrounds and lighting conditions.

DISFA [22] includes videos from 27 subjects, with around 130,000 frames. Each frame has labels for eight AU intensities (1, 2, 4, 6, 9, 12, 25, and 26). Following the settings of previous studies [21, 25, 36], we map the AU intensity greater than 1 to the positive class.

Table 1. Within-domain evaluation in terms of F1 score (↑). Except for GH-Feat and ME-GraphAU + FFHQ pre-train, all the baseline numbers are from the original papers. Our method achieves competitive performance compared to the state-of-the-art.

Methods	DISFA	BP4D
DRML [36]	26.7	48.3
IdenNet [29]	52.6	59.3
SRERL [17]	55.9	62.9
UGN-B [26]	60.0	63.3
HMP-PS [27]	61.0	63.4
FAT [12]	61.5	64.2
Zhang *et al.* [35]	62.0	63.5
JÂA-Net [25]	63.5	62.4
PIAP [28]	63.8	64.1
Chang *et al.* [4]	64.5	64.5
ME-GraphAU [21]	63.1	**65.5**
ME-GraphAU + FFHQ pre-train	59.5	61.1
GH-Feat [31]	36.9	56.7
Ours	**65.4**	64.3

BP4D [33] consists of videos from 41 subjects with around 146,000 frames. Each frame has labels for 12 AU occurrences (1, 2, 4, 6, 7, 10, 12, 14, 15, 17, 23, and 24).

We use dlib [15] to detect the 68 facial landmarks for all the frames and FFHQ-alignment to align them. The detected landmarks are also used for generating the ground-truth heatmaps (see Figure 4).

4.2. Implementation and Training Details

All methods are implemented in PyTorch [23]. Code and model weights are available, for the sake of reproducibility.[1] We use a machine with two Intel(R) Xeon(R) Gold 5218 (2.30GHz) CPUs with eight NVIDIA Quadro RTX8000 GPUs for all the experiments. Each image is resized into 128×128. We train the proposed model with the AdamW optimizer [20] for 15 epochs with a batch size of 8 on a single GPU. The learning rate is $5e-5$. The weight decay is $5e-4$. The gradient clipping is set to 0.1. σ for the heatmaps (Equation 2) is 20.0. The dropout rate is 0.1.

4.3. Experimental Results

The models are evaluated for within-domain and cross-domain performance in addition to data efficiency. Cross-domain evaluation enables us to measure the generalization ability of our AU detection method. For all the experiments, F1 score (↑) is the evaluation metric.

Within-domain Evaluation. We perform within-domain evaluation on widely used DISFA and BP4D datasets. We follow the same evaluation protocols as the previous studies [21, 25, 27]. Both datasets are evaluated with subject-independent 3-fold cross-validation. We use two folds

[1] https://github.com/ihp-lab/FG-Net

Table 2. Cross-domain evaluation between DISFA and BP4D in terms of F1 scores (↑). Our model achieves superior performance compared to the baselines. * The numbers are reported from the original paper.

Direction	DISFA → BP4D						BP4D → DISFA					
AU	1	2	4	6	12	**Avg.**	1	2	4	6	12	**Avg.**
DRML [36]	19.4	16.9	22.4	58.0	64.5	36.3	10.4	7.0	16.9	14.4	22.0	14.1
JÂA-Net [25]	10.9	6.7	**42.4**	52.9	68.3	36.2	12.5	13.2	27.6	19.2	46.7	23.8
ME-GraphAU [21]	36.5	30.3	35.8	48.8	62.2	42.7	43.3	22.5	41.7	23.0	34.9	33.1
ME-GraphAU + FFHQ pre-train	20.1	32.9	38.0	**64.0**	73.0	45.6	51.2	14.4	**54.4**	17.7	30.6	33.7
GH-Feat [31]	29.4	30.0	37.1	**64.0**	**73.5**	46.8	18.9	15.2	27.5	**52.7**	50.1	32.9
Patch-MCD* [32]	-	-	-	-	-	-	34.3	16.6	52.1	33.5	50.4	37.4
IdenNet* [29]	-	-	-	-	-	-	20.1	25.5	37.3	49.6	**66.1**	39.7
Ours	**51.4**	**46.0**	36.0	49.6	61.8	**49.0**	**61.3**	**70.5**	36.3	42.2	61.5	**54.4**

for training and one fold for validation and iterate three times. We compare FG-Net to the state-of-the-art AU detection methods, including DRML [36], IdenNet [29], SRERL [17], UGN-B [26], HMP-PS [27], FAT [12], Zhang *et al*. [35], JÂA-Net [25], PIAP [28], Chang *et al*. [4], and ME-GraphAU [21]. These baseline numbers are reported from the original papers.

Previous methods do not use the FFHQ dataset for training. Thus, to make the comparison fair, two baselines are implemented and compared, *e.g*., ME-GraphAU + FFHQ pre-train and GH-Feat [31]. Specifically, we first pre-train the ME-GraphAU's backbones (ResNet and Swin Transformer) with the FFHQ dataset and its facial expression labels. Then we train the ME-GraphAU with the pre-trained backbones. GH-Feat extracts features from generative models and it is pre-trained on the FFHQ in a self-supervised manner. For both baselines, we implement with the officially released source codes.

Table 1 reports the within-domain results regarding the average performance of AUs. We provide detailed results for every individual AU in the supplementary material. On DISFA, FG-Net outperforms all the baseline methods and achieves an average F1 score of 65.4. The major improvement comes from AU1 and AU2. On BP4D, FG-Net achieves competitive performance. These results demonstrate that the pixel-wise features extracted from StyleGAN2 are beneficial for heatmap-based AU detection.

Cross-domain Evaluation. We perform two directions of cross-domain evaluation, *i.e*., DISFA to BP4D and BP4D to DISFA. For each direction, we use two folds and one fold of data from the source domain as the training and validation set and use the target data as the testing set. We compare the proposed method with DRML [36], JÂA-Net [25], ME-GraphAU [21], ME-GraphAU + FFHQ pre-train, and GH-Feat [31] since they are open-source, and we can use the officially released source codes and model weights to conduct the experiments. In addition, we compare with Patch-MCD [32] and IdenNet [29]. The numbers are reported from the original paper. Ertugrul *et al*. [7, 8] and Hernandez *et al*. [10] do not report F1 scores for the cross-domain performance of the aforementioned directions.

We report the cross-domain results in Table 2. As expected, compared to the within-domain performance, all the baseline methods suffer a considerable performance loss when evaluated across corpora. In particular, when evaluated from BP4D to DISFA, the baseline methods' performance (average F1) drops by more than 30%, which demonstrates the challenging nature of cross-domain AU detection and the importance of developing generalizable AU detection.

Compared with DRML and JÂA-Net, ME-GraphAU achieves higher cross-domain performance. We suspect it is because it utilizes the pre-trained models (ResNet [9] and Swin Transformer [18]) as the backbones. In addition, when we continue pre-training ME-GraphAU with the FFHQ dataset, we observe a further performance boost in both directions of cross-domain evaluations. Similarly, GH-Feat, which is trained on the FFHQ dataset, also obtains superior performance than DRML and JÂA-Net. The experimental results show the effectiveness of pre-training on the FFHQ dataset since it is a large and diverse facial image dataset. Moreover, Patch-MCD utilizes unsupervised domain adaptation with unlabeled target data while IdenNet is jointly trained by AU detection and face cluster datasets (CelebA [19]). Thus, with additional face data, these two methods have better cross-domain performance than the aforementioned baselines.

For both directions of cross-domain evaluation, our proposed method achieves superior performance compared to the baselines. Specifically, when evaluated from BP4D to DISFA, FG-Net can outperform the baselines by 15% in terms of the average F1 score. The major improvement comes from AU1 and AU2, which is consistent with the findings in within-domain evaluation. Overall, the results showcase that features extracted from the StyleGAN2 generator are generalizable and thus improve the performance for cross-domain AU detection, showing its potential to solve AU detection in a real-life scenario.

We present two qualitative examples of cross-domain prediction in Figure 5. The models are trained on the BP4D

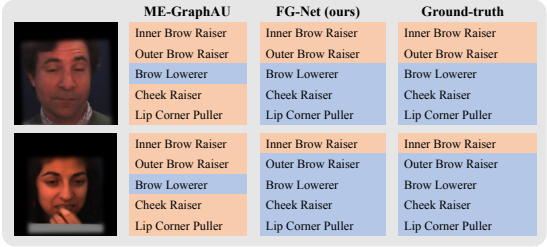
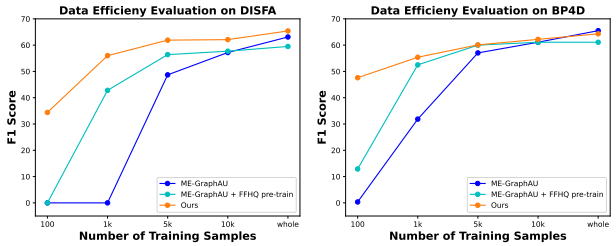


Figure 5. Case analysis on ME-GraphAU [21] and FG-Net. Models are trained on BP4D and tested on DISFA. Orange means active AU while blue means inactive AU. FG-Net is more accurate than ME-GraphAU.



(a) Evaluation on DISFA. (b) Evaluation on BP4D.

Figure 6. Data efficiency evaluation with different numbers of samples. Our method is data-efficient and its performance trained on 1k samples is close to the whole set.

dataset and tested on the DISFA dataset. ME-GraphAU fails in those two cases while the proposed FG-Net method accurately predicts the action units.

Data Efficiency Evaluation. To further evaluate the generalization capacity of the proposed approach, an investigation of its learning capability with limited samples is conducted through within-domain evaluation. In this evaluation, a subset of the training data is randomly selected, while the testing data remains unchanged to facilitate assessment. The model is trained using four different sample sizes: 100, 1k, 5k, and 10k. A comparative analysis is performed between our method and two other approaches, namely ME-GraphAU [21] and ME-GraphAU + FFHQ pre-train. It is noteworthy that ME-GraphAU + FFHQ pre-train and our method employ the same pre-training dataset.

The efficiency evaluation results, depicted in Figure 6, demonstrate the impact of data scarcity on performance for both datasets. Notably, ME-GraphAU [21] exhibits remarkably low F1 scores when trained with 100 and 1k samples on the DISFA dataset, as well as with 100 samples on the BP4D dataset. This outcome can be attributed to the limited and sparse nature of the training set, causing ME-GraphAU to predict inactive AUs predominantly. By contrast, the performance of ME-GraphAU improves when pre-trained on the FFHQ dataset, underscoring the effectiveness of utilizing this extensive and diverse facial dataset for pre-training.

Table 3. Ablation study for FG-Net. F1 score (↑) is the metric. D and B stand for DISFA and BP4D. D → B means the model is trained on DISFA and tested on BP4D and similar to B → D. (i) Our method gets better performance than GH-Feat [31]. (ii) With every component, our method achieves the highest within-domain performance while removing late features gets the best cross-domain performance.

	D	B	**Avg.**	D → B	B → D	**Avg.**
Upscale & concat	64.2	62.7	63.4	42.5	35.9	39.2
Latent code	68.4	58.8	63.6	46.4	47.3	46.9
- Early	68.1	61.7	64.9	37.9	47.2	42.6
- Middle	67.4	63.1	65.3	48.5	38.0	43.3
- Late	67.4	62.8	65.1	**51.2**	**56.6**	**53.9**
FG-Net	**68.9**	**63.6**	**66.3**	49.0	54.4	51.7

However, even with 100 samples from the DISFA dataset, the performance of ME-GraphAU remains at 0.

In comparison, FG-Net outperforms ME-GraphAU + FFHQ pre-train when trained with partial training data for both datasets. Notably, FG-Net trained on 1k samples achieves performance levels approaching those of the full training set. Furthermore, even with a mere 100 training samples, FG-Net manages to achieve commendable performance. These results serve as evidence of the robust generalization ability exhibited by our proposed method when confronted with limited data.

Ablation Study. We conduct three ablation experiments: (i) We compare to the existing upscaling and concatenating features proposed in [2, 34] (upscale & concat). (ii) We directly compare to using latent code to predict the activations of AUs (latent code). (iii) We explore the best blocks for extracting feature maps. Specifically, we divide the features extracted from the nine synthesis blocks into three groups, where each group has three feature maps, and denote them as the early, middle, and late groups. Each time, we remove one group. We perform both within- and cross-domain evaluations for the ablation study. Note that for within-domain evaluation, we use two folds for training and one fold for validation.

Table 3 shows the within- and cross-domain performance on DISFA and BP4D. (i) We observe that FG-Net outperforms Upscale & concat for both within- and cross-domain settings. We believe inference with singe-pixel features lacks the inductive bias, considering local features, necessary for AU detection. (ii) FG-Net outperforms latent code for predicting AU activations for both within- and cross-domain experiments. We think using the heatmap regression allows the model to localize where the AUs occur and improves the model's capacity. In addition, compared with the 2D feature maps, the latent codes lose the semantic-rich representations. (iii) For the contributions of different feature maps, we observe that removing any one of the feature groups lowers the within-domain performance. Sur-

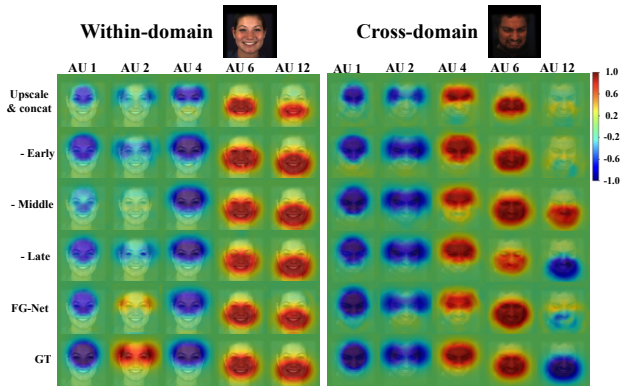


Figure 7. Visualization of the detected heatmaps for ablation study. With all the components, FG-Net detects the most similar heatmaps to the ground-truth (GT) for within-domain evaluation. Removing late features results in the best cross-domain evaluation.

prisingly, removing late features achieves the highest cross-domain performance. We suspect it is because the late features contain more high-frequency and domain-specific information which reduces the model's generalization ability.

We visualize the ground-truth and detected heatmaps for ablation study in Figure 7. For the within-domain evaluation, models are trained and tested with BP4D; For the cross-domain evaluation, models are trained with BP4D and tested with DISFA. For latent code, we directly use it to predict the AU activations, thus, we do not have the detected heatmaps for latent code. For within-domain evaluation, FG-Net detects all AUs correctly, whereas the other methods output the wrong prediction for AU2 (outer brow raiser), showing that FG-Net achieves the best within-domain performance with every component. For cross-domain evaluation, both using all features and removing late features detect all AUs correctly. However, removing late features results in a more accurate heatmap for AU12 than using all features.

4.4. Limitations

In the within-domain evaluation, FG-Net achieves inferior results on AU9 (nose wrinkler), AU15 (lip corner depressor), and AU26 (jaw drop). Failure cases are shown in Figure 8. We suspect it is because the FFHQ dataset lacks faces with such active AUs, and thus the StyleGAN2 features can not capture the corresponding information well. In addition, these failure AUs are not common in DISFA and BP4D thus they do not appear in the cross-domain evaluations and we can not evaluate the generalization for them.

FG-Net addresses the AU detection problem using a heatmap regression. Though our method can be extended to AU intensity estimation, there are only three common AUs for intensity estimation between BP4D and DISFA (6, 12, and 17) with no AU on the eyebrows. Thus, we can not properly evaluate the generalization ability of FG-Net for AU intensity estimation.

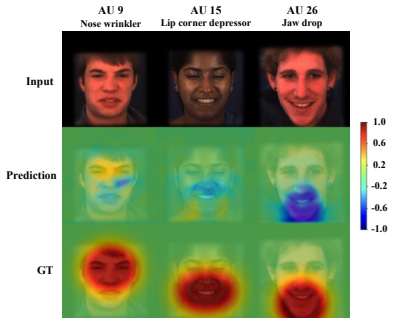


Figure 8. Visualization of the failure cases. FG-Net achieves inferior performance on AU9, AU15, and AU26.

5. Conclusion

In this paper, we propose FG-Net, a data-efficient method for generalizable facial action unit detection. FG-Net extracts the generalizable and semantic-rich features from the generative model. A Pyramid CNN-Interpreter is proposed to detect AUs coded as heatmaps which makes the training efficient and captures essential information from the nearby regions. The experimental results demonstrate the challenging nature of cross-domain AU detection and the importance of developing generalizable AU detection. We show that the proposed FG-Net method has a strong generalization ability when evaluated across corpora or trained with limited data, demonstrating its strong potential to solve action unit detection in a real-life scenario.

Social Implications. Our work falls within the broad domain of facial expression analysis. Despite potential benefits, any surveillance technology can be misused, and sensitive private information may be revealed by malicious actors. Mitigation strategies for such misuses include restrictive licensing and government regulations.

6. Acknowledgement

The work of Soleymani, Yin and Chang was sponsored by the Army Research Office and was accomplished under Cooperative Agreement Number W911NF-20-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4432–4441, 2019. 4

[2] Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*, 2021. 1, 2, 3, 5, 7

[3] Sam Bond-Taylor, Adam Leach, Yang Long, and Chris G Willcocks. Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models. *arXiv preprint arXiv:2103.04922*, 2021. 2

[4] Yanan Chang and Shangfei Wang. Knowledge-driven self-supervised representation learning for facial action unit recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20417–20426, 2022. 2, 5, 6

[5] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 3

[6] Paul Ekman. Facial action coding system, 1977. 1, 2

[7] Itir Onal Ertugrul, Jeffrey F Cohn, László A Jeni, Zheng Zhang, Lijun Yin, and Qiang Ji. Cross-domain au detection: Domains, learning approaches, and measures. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–8. IEEE, 2019. 2, 6

[8] Itir Onal Ertugrul, Jeffrey F Cohn, László A Jeni, Zheng Zhang, Lijun Yin, and Qiang Ji. Crossing domains for au coding: Perspectives, approaches, and measures. *IEEE transactions on biometrics, behavior, and identity science*, 2(2):158–171, 2020. 2, 6

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[10] Javier Hernandez, Daniel McDuff, Alberto Fung, Mary Czerwinski, et al. Deepfn: towards generalizable facial action unit recognition with deep face normalization. *arXiv preprint arXiv:2103.02484*, 2021. 2, 6

[11] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017. 4

[12] Geethu Miriam Jacob and Bjorn Stenger. Facial action unit detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7680–7689, 2021. 2, 5, 6

[13] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2, 4

[14] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 2, 3, 4

[15] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009. 5

[16] Daiqing Li, Junlin Yang, Karsten Kreis, Antonio Torralba, and Sanja Fidler. Semantic segmentation with generative models: Semi-supervised learning and strong out-of-domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8300–8311, 2021. 1, 2, 3, 4

[17] Guanbin Li, Xin Zhu, Yirui Zeng, Qing Wang, and Liang Lin. Semantic relationships guided representation learning for facial action unit recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8594–8601, 2019. 5, 6

[18] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 6

[19] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 6

[20] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[21] Cheng Luo, Siyang Song, Weicheng Xie, Linlin Shen, and Hatice Gunes. Learning multi-dimensional edge feature-based au relation graph for facial action unit recognition. *arXiv preprint arXiv:2205.01782*, 2022. 1, 2, 5, 6, 7

[22] S Mohammad Mavadati, Mohammad H Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F Cohn. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013. 1, 2, 5

[23] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. In *NeurIPS Autodiff Workshop*, 2017. 5

[24] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021. 2, 3, 4

[25] Zhiwen Shao, Zhilei Liu, Jianfei Cai, and Lizhuang Ma. Jaanet: Joint facial action unit detection and face alignment via adaptive attention. *International Journal of Computer Vision*, 129(2):321–340, 2021. 1, 2, 3, 5, 6

[26] Tengfei Song, Lisha Chen, Wenming Zheng, and Qiang Ji. Uncertain graph neural networks for facial action unit detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 5993–6001, 2021. 5, 6

[27] Tengfei Song, Zijun Cui, Wenming Zheng, and Qiang Ji. Hybrid message passing with performance-driven structures for facial action unit detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6267–6276, 2021. 2, 5, 6

[28] Yang Tang, Wangding Zeng, Dafei Zhao, and Honggang Zhang. Piap-df: Pixel-interested and anti person-specific facial action unit detection net with discrete feedback learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12899–12908, 2021. 2, 5, 6

[29] Cheng-Hao Tu, Chih-Yuan Yang, and Jane Yung-jen Hsu. Idennet: Identity-aware facial action unit detection. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–8. IEEE, 2019. 2, 5, 6

[30] Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. Gan inversion: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 4

[31] Yinghao Xu, Yujun Shen, Jiapeng Zhu, Ceyuan Yang, and Bolei Zhou. Generative hierarchical features from synthesizing images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4432–4442, 2021. 2, 3, 5, 6, 7

[32] Yufeng Yin, Liupei Lu, Yizhen Wu, and Mohammad Soleymani. Self-supervised patch localization for cross-domain facial action unit detection. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–8. IEEE, 2021. 2, 6

[33] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014. 1, 2, 5

[34] Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. Datasetgan: Efficient labeled data factory with minimal human effort. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10145–10155, 2021. 1, 2, 3, 4, 5, 7

[35] Zheng Zhang, Taoyue Wang, and Lijun Yin. Region of interest based graph convolution: A heatmap regression approach for action unit detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2890–2898, 2020. 2, 4, 5, 6

[36] Kaili Zhao, Wen-Sheng Chu, and Honggang 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. 1, 2, 3, 5, 6