G2RL: Geometry-Guided Representation Learning for Facial Action Unit Intensity Estimation

Yingruo Fan[1] and **Zhaojiang Lin**[2]

[1]Department of Electrical and Electronic Engineering, University of Hong Kong
[2]Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology
yingruo@hku.hk, zlinao@connect.ust.hk

Abstract

Facial action unit (AU) intensity estimation aims to measure the intensity of different facial muscle movements. The external knowledge such as AU co-occurrence relationship is typically leveraged to improve performance. However, the AU characteristics may vary among individuals due to different physiological structures of human faces. To this end, we propose a novel geometry-guided representation learning (G2RL) method for facial AU intensity estimation. Specifically, our backbone model is based on a heatmap regression framework, where the produced heatmaps reflect rich information associated with AU intensities and their spatial distributions. Besides, we incorporate the external geometric knowledge into the backbone model to guide the training process via a learned projection matrix. The experimental results on two benchmark datasets demonstrate that our method is comparable with the state-of-the-art approaches, and validate the effectiveness of incorporating external geometric knowledge for facial AU intensity estimation.

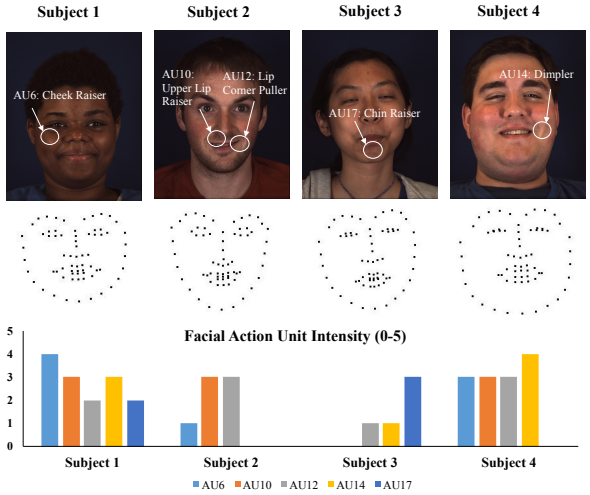


Figure 1: Examples of facial displays (row 1) and the corresponding facial landmark locations (row 2) and AU intensities (row 3). Different human faces have different morphological aspects and different ways of expressing emotions. Hence, we expect the facial geometric information of different facial expressions could help to enhance the feature representation learning, thus facilitating distinguishing the subtle differences among facial AU intensities.

1 Introduction

Most facial expression recognition (FER) systems have achieved high accuracy in recognizing a set of prototypical expressions, e.g., angry, happy, sad, etc., following the pioneer work of Ekman [1993]. However, they cannot provide detailed descriptions of the fine-grained physical appearance changes in human facial expressions. To describe emotions more systematically, the Facial Action Coding System (FACS) [Friesen and Ekman, 1978] encodes the movements of some specific facial muscles, named Action Units (AUs), and quantifies AU intensities into six discrete levels. Based on FACS, human coders can manually deconstruct any possible facial expression into several AUs and their corresponding intensities. Therefore, estimating the intensities of AUs is important for interpreting and analyzing facial expressions.

Facial AU intensity estimation has attracted increasing attention in affective computing community. It has potential applications in human-computer interaction [Lin *et al.*, 2019b; Lin *et al.*, 2019a], health care, games, marketing, etc. Nevertheless, AU intensity estimation remains a difficult problem given the differences among individuals and large variations in facial geometry. As shown in Figure 1, the locations of facial points characterize the face shape patterns of different subjects. Intuitively, the facial geometric information may influence the performance of AU intensity estimation. For example, the intensity of AU12 (Lip Corner Puller) is highly correlated with the height of the mouth opening and the distance between lip corners. Given that AUs appear as the movements of facial muscles, we propose to use geometric features as the external knowledge in learning better AU representations. Moreover, facial geometric information can be obtained reliably and quickly with the progress of facial landmark detection algorithms.

Meanwhile, prior knowledge can be used to guide and facilitate the model learning. For AU intensity estimation, researchers have begun to investigate how to leverage exter-

nal knowledge such as co-occurrence relationship between AUs [Walecki *et al.*, 2017], facial symmetry [Zhang *et al.*, 2018a], and temporal label smoothness [Zhang *et al.*, 2019]. The facial geometric features have been shown to be effective for FER [Jung *et al.*, 2015] and facial AU detection [Niu *et al.*, 2019]. However, they are rarely utilized in previous studies to improve AU intensity estimation. Considering the low computational cost, geometric features offer much room for injecting external knowledge into the deep neural network during its training stage. In addition, there is a strong correlation between the geometric positions of facial points due to the physiological structure of the human face, which can be used to enrich the geometric information via relationship learning. In the era of geometric deep learning, existing studies such as PointNet [Qi *et al.*, 2017] and its various extensions have explored flexible geometric representation suitable for modeling the geometric data. Motivated by such applications in computer graphics, we propose to capture geometric properties of facial points by constructing a graph convolutional neural network (Graph CNN) [Kipf and Welling, 2016] to learn more robust AU-related representation. The Graph CNN is an ideal model for modeling spatial vector data since it can learn complex relationships and interdependency between vectors. In our case, for each person, a set of facial landmarks is represented as a graph, where vertices correspond to the 2D coordinates and edges represent their spatial relations.

On the other hand, the heatmap can provide a per-pixel likelihood for keypoint locations in an image, e.g., human pose estimation [Xiao *et al.*, 2018] and face alignment, and has shown its effectiveness in AU localization and intensity estimation [Sánchez-Lozano *et al.*, 2018]. Thus, in our experiments, we incorporate both the intensity information and spatial configuration of AUs into the heatmap regression-based network. Most AU-related methods usually train a model to output a vector representing AU intensities. In contrast, the heatmap regression-based framework can jointly regress AU intensities and their locations. The most common method of illustrating a heatmap is representing the values stored in a matrix with gradually-changed colors. Specifically, the heatmap in our framework is utilized to represent the varying intensity of facial muscles, where the lighter color indicates a lower AU intensity while the darker color reflects a higher AU intensity. In our experimental setting, we define the locations of AUs using a set of facial points, as shown in Figure 2.

In this work, we also explore how to incorporate the geometric knowledge into the heatmap regression framework. Inspired by the work of [Ning *et al.*, 2017], which employs knowledge projection for guided learning, we inject geometric features that are inferred from facial landmarks as the external knowledge to enrich the representation power of the predicted heatmaps. In our framework, as illustrated in Figure 3, the external geometric knowledge is encoded into a latent representation that characterizes the constraints of facial points and their interdependencies. Besides, we introduce an auxiliary loss function for the knowledge projection module, enforcing the geometry-guided representation learning during the training process. During the testing stage, the external knowledge representation and its associated modules are removed without increasing computational cost.

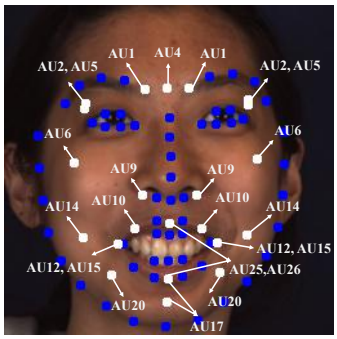


Figure 2: Facial AU locations defined by a set of facial landmark coordinates. Each white point denotes the location of a specific AU. Note that most AU locations are in pairs.

In summary, our key contributions are three folds:

- We present a new approach that incorporates the external geometric knowledge to guide the training of the heatmap regression network for facial AU intensity estimation.
- We propose to capture the facial geometric constraints and relationships among facial points by constructing a graph convolutional neural network to learn more robust AU-related representation.
- The experimental results on BP4D and DISFA datasets show that G2RL achieves better or comparable performance than current state-of-the-art methods in estimating facial AU intensities.

2 Related Work

2.1 Facial Action Unit Intensity Estimation

The focus of most existing studies have been on facial expression recognition or facial action unit detection, whereas relatively few works have investigated the intensity estimation of facial AUs. As an earlier work, [Li *et al.*, 2013] considered the temporal information and the dependencies among multiple AU intensities in a unified probabilistic framework. Similarly, [Sandbach *et al.*, 2013] integrated traditional handcrafted features with AU intensity combination priors in Markov Random Field (MRF) Structures. They demonstrated that MRF structures were able to model the interdependencies between AUs. Later on, [Kaltwang *et al.*, 2015] built a latent tree model for structure learning based on the Bayesian Expectation Maximization (EM) algorithm to capture higher-order relationships between the observable features and AU intensities. To model the non-linear dependencies among multiple AUs, [Walecki *et al.*, 2016] proposed a copula ordinal regression framework by considering the ordinal structure in output AU intensities. However, these approaches depend heavily on the quality of the extracted features and the choice of the probabilistic models.

To ameliorate this issue, several methods began to leverage the discriminative capabilities of deep neural networks

for AU intensity estimation. For instance, [Walecki *et al.*, 2017] utilized a Copula CNN deep learning model for joint learning of multiple AU intensity outputs. [Sánchez-Lozano *et al.*, 2018] jointly performed facial AU localization and intensity estimation via heatmap regression. In the work of Li *et al.* [2018], a new network architecture, named Edge Convolutional Network (ECN), was designed to learn edge-like detectors for capturing subtle facial muscle changes.

2.2 Knowledge-based Methods for AU Intensity Estimation

Recently, researchers recognize that using external knowledge representation can bring significant advantages in estimating AU intensities. [Zhang *et al.*, 2018a] proposed a knowledge-based semi-supervised deep model, which could identify four types of domain knowledge including facial symmetry, ordinal intensity, etc. Meanwhile, [Zhang *et al.*, 2018b] emphasized domain knowledge on the relevance in sequential data to learn a feasible frame-level AU intensity regressor. More recently, the follow-up work of Zhang *et al.* [2019] incorporated different types of human knowledge, e.g., temporal label smoothness, label ranking, etc., as regularization terms or other constraints to learn AU representations and estimator simultaneously.

Different from previous methods, we propose a new approach that explores the geometric knowledge associated with facial point locations as well as their dependencies. We expect this external knowledge would help in analyzing spontaneous facial AUs and boosting the performance of AU intensity estimation. To our knowledge, the approaches introduced in [Niu *et al.*, 2019; Wu and Ji, 2017] also take into account the geometric constraints in the face shape. However, they ignore the underlying latent relationship between different facial landmarks, which can provide more robustness than only using the basic face shape and coordinate information.

3 Methodology

3.1 Backbone Model

The backbone model illustrated in Figure 3 is a heatmap regression-based network, where feature maps contain rich semantic information of AU intensities and locations. In our framework, we adopt the simplest network structure [Xiao *et al.*, 2018] to generate heatmaps. Let us denote the set of training images as $\mathcal{I} \subseteq \mathbb{R}^{C\times W\times H}$ (C: image channels, W: image width, H: image height). Given an image $X \in \mathcal{I}$, according to its corresponding facial landmark annotations, we calculate its coordinates of AU locations and denote them as $L \in \mathbb{R}^{K\times 2}$, where K is the total number of AUs; On the other hand, the corresponding AU intensities are represented by $I \in \mathbb{R}^{K\times 1}$. The network $\Phi_{\phi_\mathcal{I}}$ is trained to predict the set of AU locations and intensities, where $\phi_\mathcal{I}$ denotes the set of weight parameters of Φ. For each AU location $L_k = (i_k, j_k)(k = 1, ..., K)$, the ground-truth heatmap is produced by applying a Gaussian function as follows

$$S_k(i, j; X) = \frac{I_k}{2\pi\sigma^2} \exp(-\frac{\|(i, j) - (i_k, j_k)\|_2^2}{2\sigma^2}), \quad (1)$$

where σ is the standard deviation. Hence, each image is labeled with a set of ground-truth heatmaps $S = \{S_1, ..., S_K\} \subseteq \mathbb{R}^{W^{'}\times H^{'}}$ ($W^{'}$: heatmap width, $S^{'}$: heatmap height). From Equation 1, it can be observed that the pixel location (i, j) that is farther away from the AU location (i_k, j_k) would result in a lower value in $S_k(i, j; X)$.

To optimize the network parameters $\phi_\mathcal{I}$, we utilize the mean squared error (MSE) loss to minimize the difference between the heatmaps predicted by $\Phi_{\phi_\mathcal{I}}$ and the ground-truth. Therefore, the optimization process is formulated as

$$L_S = \min_{\phi_\mathcal{I}} \sum_{X\in\mathcal{I}} \|\Phi_{\phi_\mathcal{I}}(X) - S(X)\|_2^2. \quad (2)$$

During the inference stage, the estimated AU locations are given by $\hat{L} = \arg\max \Phi_{\phi_\mathcal{I}}(X)$ while the corresponding AU intensities are obtained from the highest values of the produced heatmaps, i.e., $\hat{I} = \max \Phi_{\phi_\mathcal{I}}(X)$.

3.2 External Geometric Knowledge Module

The facial geometry is represented as a set of facial landmarks $\{p|i = 1, ..., n\}$, where each point p_i is a 2-D vector of its position coordinate. Similarly to principles in geometry modeling [Qi *et al.*, 2017], those facial points should be correlated with each other and the interaction among them is expected to enhance the learned representation. The goal of the external geometric knowledge module is to summarize the face shape pattern and the interdependencies of facial points into a latent vector G. Therefore, we design a model derived from graph convolutional network (GCN) to aggregate the geometric information in the graph structure. The position coordinates of facial landmarks are obtained through the Dlib[1] library. Based on the position coordinates of n points, denoted by $P \in \mathbb{R}^{n\times 2}$, the external geometric knowledge module builds a graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, where $\mathcal{N}$ is the vertices and each edge $e_{i,j} \in \mathcal{E}$ indicates the relationship between two nodes. In our case, we use three GCN layers to extract geometric features and the l_{th} GCN layer can be represented as

$$F^l = ReLU(A^{l-1}F^{l-1}W^{l-1}), \quad (3)$$

where W^{l-1} represents a trainable weight matrix for the specific layer, A^{l-1} is the adjacency matrix determined by the Euclidean distances between the nodes, and $ReLU(\cdot)$ is used as the activation function. By integrating the information of each node and its neighbors, the representation for each node would be updated after each GCN layer.

Accordingly, three GCN layers execute the transformation of Equation 3, after which their features are concatenated into the following feature representation

$$F^{'} = \|_{k=1}^{3}\ ReLU(A^{l-1}F^{l-1}W^{l-1}), \quad (4)$$

where $\|$ represents concatenation, and $F^0 = P$. Finally, three fully-connected layers are applied to obtain the latent vector G, which can be formulated as

$$G = Fc(Fc(Fc(Conv(F^{'})))), \quad (5)$$

where Fc is the fully connected layer and $Conv$ means the 1×1 convolution operation to aggregate the concatenated

[1] dlib.net

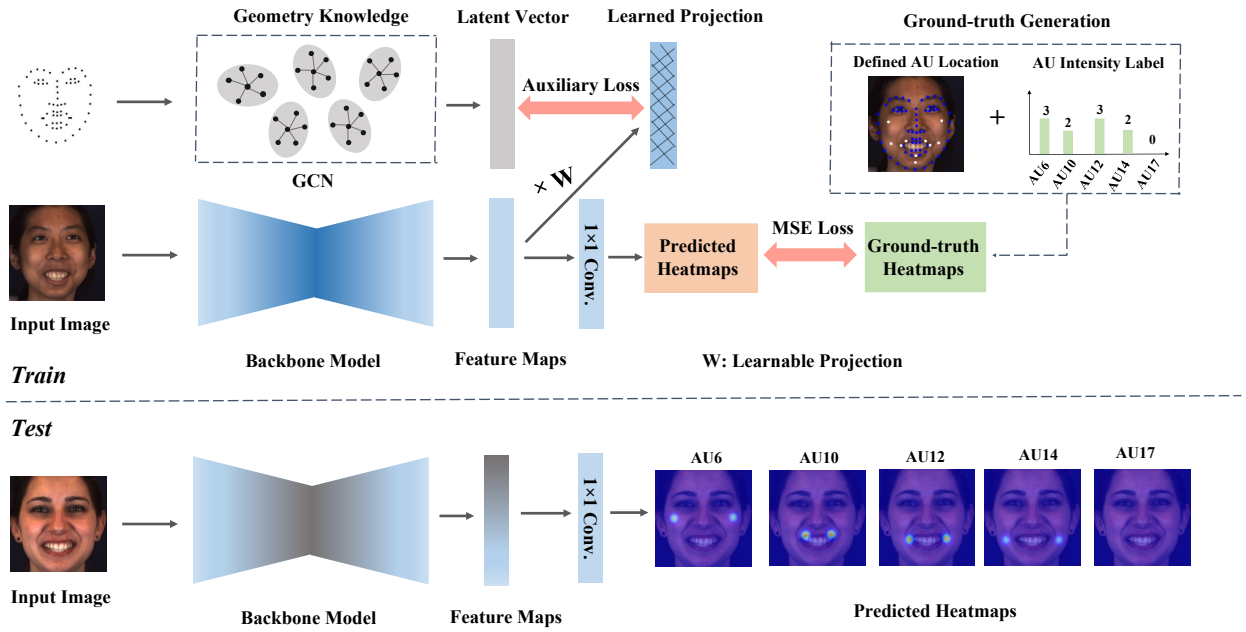


Figure 3: Illustration of our framework. During the training process, a Graph CNN is utilized to capture knowledge associated with the facial geometric constraints and interdependencies among facial points. Besides, an auxiliary loss function is introduced to enforce geometry-guided representation learning via a projection matrix. The gradients generated from the auxiliary loss are propagated back to the backbone model. During the test process, the knowledge representation and its associated modules removed and the predicted heatmaps are directly inferred from the learned backbone model.

features. Hence, in the external geometric knowledge module, we first obtain a multi-resolution feature representation integrating both low- and high-level geometric information via Equation 4, and then transform it into a latent vector using Equation 5 for the following guided learning.

3.3 Knowledge Projection Layer

As shown in Figure 3, during the training process, we aim to inject geometric knowledge to guide the learning of the basic framework, so as to enrich the representation power of the predicted heatmaps. To this end, a linear projection W is applied between the geometric knowledge representation G and the feature maps M before the predicted heatmaps. Inspired by [Ning *et al.*, 2017], we apply an auxiliary loss L_G to generate gradients enforcing the backbone model to learn the external knowledge

$$L_G = \min_{\phi_{\mathcal{I},\mathcal{P}}} \|G - W \times M\|_2^2, \tag{6}$$

where $\mathcal{P}$ refers to facial landmark annotations of the training set. The learned projection (Figure 3) can affect the gradients propagated back to the backbone model. We inject the geometric knowledge into the feature maps M as they are highly correlated with the predicted heatmaps. Moreover, they share the same parameters on former layers and can influence the parameter learning of these layers.

With the geometric knowledge introduced, the MSE loss in Equation 2 is then reformulated as

$$L_S^{'} = \min_{\phi_{\mathcal{I},\mathcal{P}}} \sum_{X \in \mathcal{I}, P \in \mathcal{P}} \left\| \Phi_{\phi_{\mathcal{I},\mathcal{P}}}(X, P) - S(X) \right\|_2^2. \tag{7}$$

Finally, the overall loss for joint training is a weighted combination of L_G and $L_H^{'}$

$$L = \min_{\phi_{\mathcal{I},\mathcal{P}}} (\lambda \times L_G + (1 - \lambda) \times L_S^{'}), \tag{8}$$

where λ is a weight parameter that controls how much guidance is imposed during the training. During the test stage, the external geometric knowledge module is removed, with only the backbone model retained. For inference, the heatmaps are directly produced by the backbone model that has learned external geometric knowledge.

4 Experiments

4.1 Datasets and Evaluation Methods

We adopt two benchmark datasets, BP4D [Zhang *et al.*, 2014] and DISFA [Mavadati *et al.*, 2013], for our experiments. The BP4D database contains 328 videos of 41 subjects performing 8 emotion-related tasks, each of which elicits the specific emotion. Each frame is annotated with intensity information of five AUs, i.e., AU6, AU10, AU12, AU14, and AU17. The AU intensity is indicated by the value in the range of 0-5, with 0 indicating the absence whereas 5 indicating the maximum. The DISFA database has 27 videos of 27 subjects, and

Database		BP4D						DISFA												
AU		6	10	12	14	17	Avg.	1	2	4	5	6	9	12	15	17	20	25	26	Avg.
ICC	KJRE	.71	.61	**.87**	.39	.42	.60	.27	**.35**	.25	**.33**	**.51**	.31	.67	.14	.17	.20	.74	.25	.35
	BORMIR	.73	.68	.86	.37	.47	.62	.20	.25	.30	.17	.39	.18	.58	.16	.23	.09	.71	.15	.28
	CCNN-IT	.75	.69	.86	.40	.45	.63	.20	.12	.46	.08	.48	.44	**.73**	**.29**	.45	**.21**	.60	.46	.38
	KBSS	**.76**	.75	.85	.49	.51	.67	.23	.11	.48	.25	.50	.25	.71	.22	.25	.06	.83	.41	.36
	Baseline	.70	.77	.78	.59	.49	.67	.46	.16	.74	.02	.32	.38	.71	.02	.35	.02	.93	.74	.40
	G2RL	.70	**.81**	.83	**.59**	**.51**	**.69**	**.71**	.31	**.82**	.06	.48	**.67**	.68	.21	**.47**	.17	**.95**	**.75**	**.52**
MAE	KJRE	.82	.95	.64	1.08	.85	.87	1.02	.92	1.86	.70	.79	.87	.77	.60	.80	.72	.96	.94	.91
	BORMIR	.85	.90	.68	1.05	.79	.85	.88	.78	1.24	.59	.77	.78	.76	.56	.72	.63	.90	.88	.79
	CCNN-IT	1.17	1.43	.97	1.65	1.08	1.26	.73	.72	1.03	.21	.72	.51	.72	.43	.50	.44	1.16	.79	.66
	KBSS	**.56**	.65	**.48**	.98	.63	.66	.48	.49	.57	.08	.26	.22	.33	.15	.44	.22	.43	.36	.33
	Baseline	.61	.60	.55	**.83**	.39	.60	.20	.16	.29	**.03**	.29	.16	.33	.15	.20	**.08**	.31	**.36**	.21
	G2RL	.62	**.55**	.50	.87	**.38**	**.58**	**.16**	**.16**	**.28**	.04	**.25**	**.15**	**.31**	**.14**	**.20**	.11	**.25**	.39	**.20**

Table 1: Performance comparison with related work on two spontaneous AU intensity datasets. The best results are shown in bold.

it also provides the six-point scale intensity labels for 12 AUs, i.e., AU1, AU2, AU4, AU5, AU6, AU9, AU12, AU15, AU17, AU20, AU25, and AU26.

In our experiments, we evaluate our method on BP4D using the official training/development partitions. While for DISFA, the 3-fold subject independent cross-validation is adopted for evaluation. To compare the performance with the state-of-the-art approaches, we empirically use the intra-class correlation coefficient (ICC) and the mean absolute error (MAE) as evaluation metrics for measuring the performance of AU intensity estimation.

4.2 Implementation Details

To pre-process the data, the landmark detector based on a C++ library Dlib is firstly employed to locate the 68 facial landmarks. Next, the face images are cropped and aligned to size 256×256 as the input of the network. Similarly to [Xiao *et al.*, 2018], in Figure 3, our backbone network is initialized by the ResNet-50 [He *et al.*, 2016] model pre-trained on ImageNet, and combines the upsampling and convolution operations into three deconvolutional layers to produce a set of 64×64 heatmaps. For the external knowledge module, the network used for extracting geometric features mainly consists of three GCN layers and three fully-connected (fc) layers. Besides, a concatenation layer is included to summarize the GCN outputs as a new feature descriptor. The multi-scale geometric features from three previous GCN layers are transformed into a 256-D latent vector after three fc layers.

The framework is implemented in Tensorflow[2] and NVIDIA GeForce GTX 1080Ti GPUs are used. In the training phase, we use the Adam optimizer [Kingma and Ba, 2014], with the base learning rate of 5e-4. For parameter setting, we set the value of the standard deviation σ to 2 in the heatmap ground-truth generation (Equation 1), and assign 0.05 to λ in the overall loss function (Equation 8) according to the performance of G2RL.

[2]https://www.tensorflow.org/

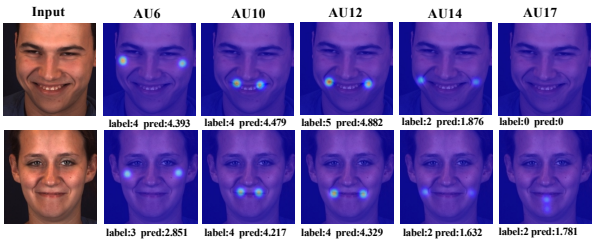


Figure 4: Visualizations of the predicted AU heatmaps from two testing samples in the BP4D dataset. The redder colors represent the higher AU intensity.

4.3 Experimental Results and Analyses

Quantitative Evaluation

We evaluate our G2RL method on two benchmark datasets, and the comparative results are summarized in Table 1. To verify the effectiveness of the external geometric knowledge, we also report the performance of the backbone model (Baseline) without any knowledge incorporated. Furthermore, we compare with the state-of-the-art approaches (KJRE [Zhang *et al.*, 2019], BORMIR [Zhang *et al.*, 2018b], CCNN-IT [Walecki *et al.*, 2017], and KBSS [Zhang *et al.*, 2018a]).

Table 1 provides the performance comparison on the BP4D and DISFA datasets. The results show that G2RL is able to estimate AU intensities with higher overall performance. The baseline model has achieved comparable performance with KBSS, which indicates the superiority of using the heatmap regression-based framework. Furthermore, the proposed G2RL outperforms the baseline version in terms of ICC and MAE on average, highlighting the importance of considering the geometric knowledge. Comparing to other existing approaches, G2RL achieves promising performances with an ICC score of 0.69 and a MAE score of 0.58 on BP4D, as well as an ICC score of 0.52 and a MAE score of 0.2 on DISFA. More importantly, G2RL does not introduce additional parameters during the inference stage since the external knowledge module has been removed. The experimental results well demonstrate that G2RL can enhance the AU representa-

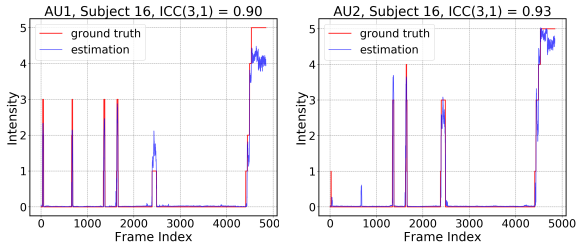
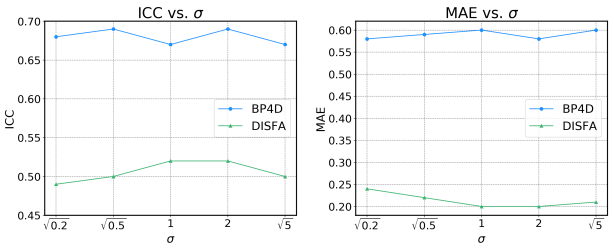
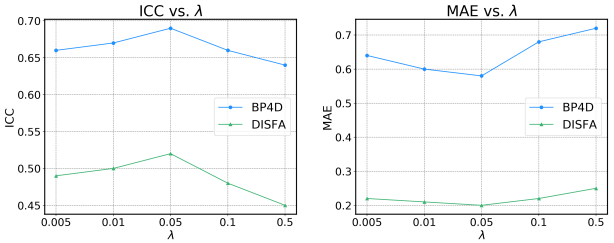


Figure 5: Predicted AU1 and AU2 intensities by G2RL for one randomly selected subject from DISFA, as compared to the corresponding ground-truths.



(a) ICC w.r.t σ (b) MAE w.r.t σ

Figure 6: Parameter sensitivity study for G2RL on BP4D and DISFA datasets with varying values of σ.



(a) ICC w.r.t λ (b) MAE w.r.t λ

Figure 7: Parameter sensitivity study for G2RL on BP4D and DISFA datasets with varying values of λ.

tions by incorporating the external geometric knowledge into the heatmap regression-based framework.

From the results in Table 1, it is notable that all the methods perform relatively better for some specific AUs, e.g., AU6, AU10, and AU12 for the BP4D database, and AU6, AU12, and AU25 for the DISFA database. One possible reason is that these AUs contain more discriminative features for distinguishing the subtle differences among their intensities. Second, the six-level intensities data of most AUs in two datasets are highly imbalanced, while "Cheek Raiser" (AU6), "Upper Lip Raiser" (AU10), "Lip Corner Puller" (AU12), and "Lips Part" (AU25) might have more data of high intensity since they occur more frequently in the real-world scenario. A more balanced distribution of AU intensities would definitely be a benefit in feature representation learning.

Qualitative Evaluation

We show further examples of the heatmaps produced by our G2RL for two subjects from the testing set of BP4D in Figure 4. It can be seen that the predicted heatmaps capture the varying intensities of different facial AUs and their corresponding locations. The redder colors represent the higher AU intensity that causes more visible facial appearance changes. The predicted AU intensities are inferred from the maximum values of the produced heatmaps. Besides, we plot the estimated AU1 and AU2 intensities of the same subject from the DISFA dataset and the ground-truth intensities in Figure 5. Note that the subject is from the validation set, following the 3-fold cross-validation protocol. We only show two correlation curves for the qualitative evaluation of our proposed method due to the page limit. It can be observed that the estimated intensities by G2RL are generally close to the ground-truth, validating the effectiveness of G2RL for frame-level AU intensity estimation.

4.4 Parameter Sensitivity Analysis

Additional experiments are conducted to explore the sensitivity of G2RL to various parameters. In Equation 1, the standard deviation σ defines the peak width of the Gaussian function, which means σ influences the pattern of the generated ground-truth heatmaps. To achieve optimal performance under a range of typical values w.r.t. σ, we report the results on BP4D and DISFA by varying σ in $\{\sqrt{0.2}, \sqrt{0.5}, 1, 2, \sqrt{5}\}$. We plot ICC and MAE w.r.t. different values of σ in Figure 6. Studying the results, G2RL performs well and steadily when σ varies and both the ICC and MAE are not overly sensitive to σ. Additionally, we investigate the effect of the parameter λ in Equation 6, which determines the degree of guidance on the joint training. Intuitively, a good trade-off between the knowledge-guided learning and the heatmap regression learning can enrich the representation power of the final predicted heatmaps. Figure 7 provides an illustration of the variation of estimation performance as $\lambda \in \{0.005, 0.01, 0.05, 0.1, 0.5\}$. We can observe that similar trends on BP4D and DISFA are shown with ICC first increasing and then decreasing. According to the curves, G2RL reaches the best performance when $\lambda = 0.05$. This validates that injecting the geometric knowledge can indeed help improve the performance provided that a proper trade-off is achieved.

5 Conclusion

In this work, we have proposed a novel geometry-guided representation learning (G2RL) framework for facial AU intensity estimation. Based on the heatmap regression framework, a Graph CNN is utilized to encode the external geometric knowledge associated with facial geometric constraints and relationships among facial points. Particularly, an auxiliary loss is tailored to generate gradients enforcing the backbone model to learn the external knowledge. Our method has considered the differences between individuals for learning more robust AU-related representations. Finally, the empirical evaluation on two benchmark datasets demonstrates the efficacy of the proposed G2RL against previous approaches and the potential of using the external geometric knowledge.

References

[Ekman, 1993] Paul Ekman. Facial expression and emotion. *American Psychologist*, 48(4):384, 1993.

[Friesen and Ekman, 1978] Wallace Friesen and Paul Ekman. Facial action coding system: a technique for the measurement of facial movement. *Palo Alto*, 3, 1978.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[Jung *et al.*, 2015] Heechul Jung, Sihaeng Lee, Junho Yim, Sunjeong Park, and Junmo Kim. Joint fine-tuning in deep neural networks for facial expression recognition. In *IEEE International Conference on Computer Vision*, pages 2983–2991, 2015.

[Kaltwang *et al.*, 2015] Sebastian Kaltwang, Sinisa Todorovic, and Maja Pantic. Latent trees for estimating intensity of facial action units. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 296–304, 2015.

[Kingma and Ba, 2014] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Kipf and Welling, 2016] Thomas Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[Li *et al.*, 2013] Yongqiang Li, Mohammad Mavadati, Mohammad Mahoor, and Qiang Ji. A unified probabilistic framework for measuring the intensity of spontaneous facial action units. In *IEEE International Conference on Automatic Face and Gesture Recognition Workshps*, pages 1–7. IEEE, 2013.

[Li *et al.*, 2018] Liandong Li, Tadas Baltrusaitis, Bo Sun, and Louis-Philippe Morency. Edge convolutional network for facial action intensity estimation. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 171–178. IEEE, 2018.

[Lin *et al.*, 2019a] Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. Moel: Mixture of empathetic listeners. *arXiv preprint arXiv:1908.07687*, 2019.

[Lin *et al.*, 2019b] Zhaojiang Lin, Peng Xu, Genta Indra Winata, Zihan Liu, and Pascale Fung. Caire: An end-to-end empathetic chatbot. *arXiv preprint arXiv:1907.12108*, 2019.

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

[Ning *et al.*, 2017] Guanghan Ning, Zhi Zhang, and Zhiquan He. Knowledge-guided deep fractal neural networks for human pose estimation. *IEEE Transactions on Multimedia*, 20(5):1246–1259, 2017.

[Niu *et al.*, 2019] Xuesong Niu, Hu Han, Songfan Yang, Yan Huang, and Shiguang Shan. Local relationship learning with person-specific shape regularization for facial action unit detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 11917–11926, 2019.

[Qi *et al.*, 2017] Charles Qi, Hao Su, Kaichun Mo, and Leonidas Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 652–660, 2017.

[Sánchez-Lozano *et al.*, 2018] Enrique Sánchez-Lozano, Georgios Tzimiropoulos, and Michel Valstar. Joint action unit localisation and intensity estimation through heatmap regression. *arXiv preprint arXiv:1805.03487*, 2018.

[Sandbach *et al.*, 2013] Georgia Sandbach, Stefanos Zafeiriou, and Maja Pantic. Markov random field structures for facial action unit intensity estimation. In *IEEE International Conference on Computer Vision Workshops*, pages 738–745, 2013.

[Walecki *et al.*, 2016] Robert Walecki, Ognjen Rudovic, Vladimir Pavlovic, and Maja Pantic. Copula ordinal regression for joint estimation of facial action unit intensity. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4902–4910, 2016.

[Walecki *et al.*, 2017] Robert Walecki, Vladimir Pavlovic, Björn Schuller, Maja Pantic, et al. Deep structured learning for facial action unit intensity estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3405–3414, 2017.

[Wu and Ji, 2017] Yue Wu and Qiang Ji. Constrained joint cascade regression framework for simultaneous facial action unit recognition and facial landmark detection. *arXiv preprint arXiv:1709.08129*, 2017.

[Xiao *et al.*, 2018] Bin Xiao, Haiping Wu, and Yichen Wei. In *European Conference on Computer Vision*, pages 466–481, 2018.

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

[Zhang *et al.*, 2018a] Yong Zhang, Weiming Dong, Bao-Gang Hu, and Qiang Ji. Weakly-supervised deep convolutional neural network learning for facial action unit intensity estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2314–2323, 2018.

[Zhang *et al.*, 2018b] Yong Zhang, Rui Zhao, Weiming Dong, Bao-Gang Hu, and Qiang Ji. Bilateral ordinal relevance multi-instance regression for facial action unit intensity estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7034–7043, 2018.

[Zhang *et al.*, 2019] Yong Zhang, Baoyuan Wu, Weiming Dong, Zhifeng Li, Wei Liu, Bao-Gang Hu, and Qiang Ji. Joint representation and estimator learning for facial action unit intensity estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3457–3466, 2019.