

## PROBLEM

- Facial expressions are combinations of basic basic patterns of muscular activation called Action Units (AU). Recognizing AUs is key for general facial expression analysis.
- Patch Learning.** AUs modify facial morphology locally. One could predict specific AUs from informative face regions selected depending on the facial geometry.
- Structure Learning.** Several AUs can be active at the same time and certain AU combinations are more probable than others. AU prediction performance could be improved by considering probabilistic dependencies.



**Figure 1:** Patch and structure learning are key problems in AU recognition. (a) By masking a region an expression becomes indistinguishable from neutral. (b) Multiple, correlated AUs can be active at the same time.

## CONTRIBUTIONS

- we propose a model that is capable of patch learning and structure learning end-to-end.
- we introduce a structure inference topology that replicates inference algorithm in probabilistic graphical models by using a recurrent neural network.

## REFERENCES

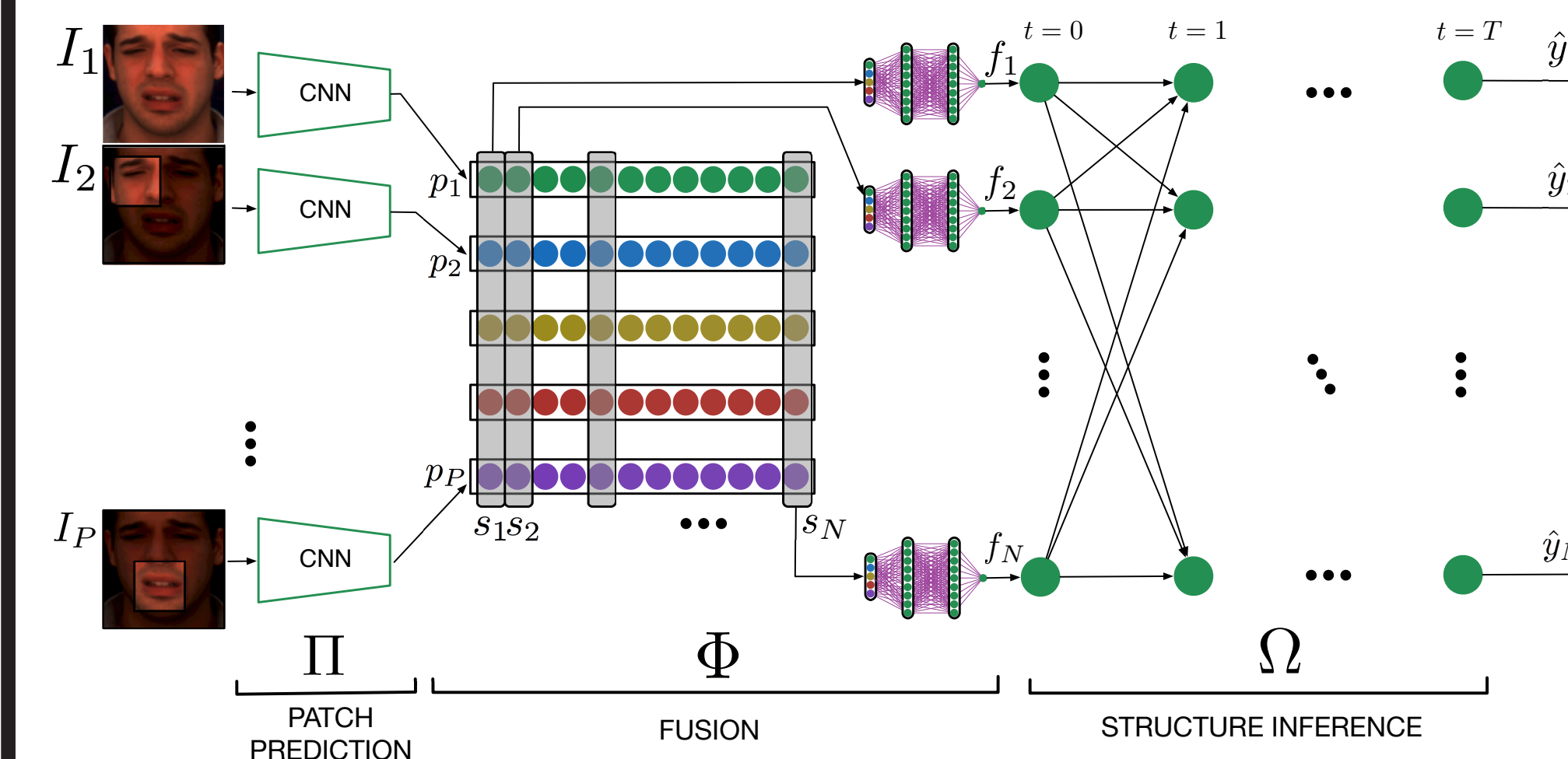
- Zhao, K., Chu, W.S., De la Torre, F., Cohn, J.F., Zhang, H.: Joint patch and multi-label learning for facial action unit detection. In: IEEE CVPR. (2015) 2207–2216
- Zhao, K., Chu, W.S., Zhang, H.: Deep region and multi-label learning for facial action unit detection. In: IEEE CVPR. (2016) 3391–3399
- Zeng, J., Chu, W.S., De la Torre, F., Cohn, J.F., Xiong, Z.: Confidence preserving machine for facial action unit detection. In: Proceedings of the IEEE ICCV. (2015) 3622–3630
- Li, W., Abtahi, F., Zhu, Z.: Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing. In: CVPR, IEEE (2017) 6766–6775

## METHOD

### Overview

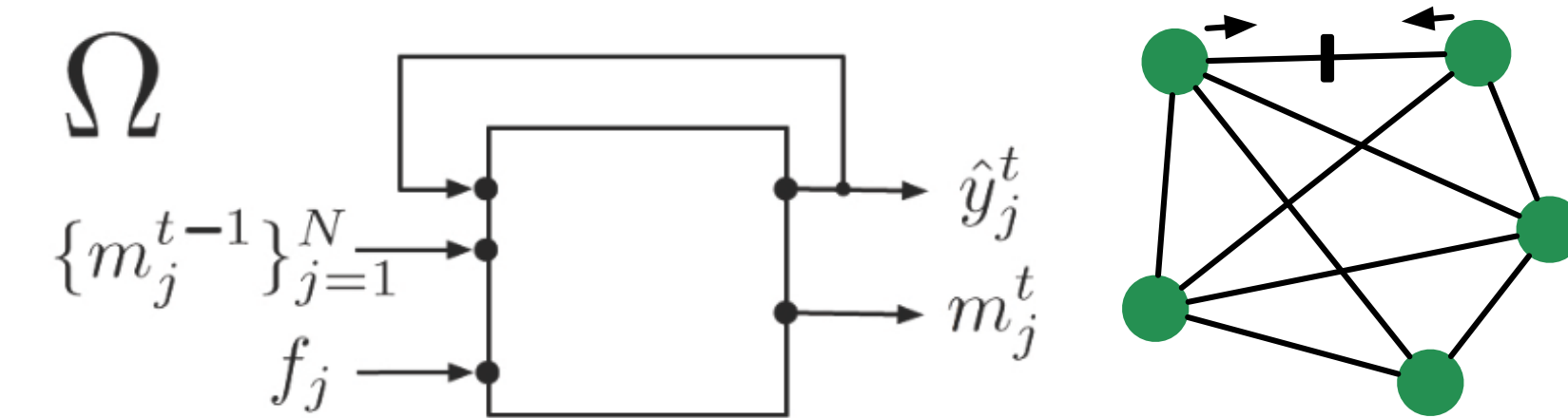
The Deep Structure Inference Network (DSIN) consists of three components:

- Patch Prediction ( $\Pi$ )** exhaustively learns deep local representations from facial patches and produce local predictions.
- Fusion ( $\Phi$ )** performs patch learning per AU.
- Structure Inference ( $\Omega$ )** refines AU prediction by capturing relationships between AUs.



**Figure 2:** The Deep Structure Inference Network.

### Structure Inference as RNN



**Figure 3:** A Structure Inference Unit (left) and a naive representation of structure inference (right).

- A Structure Inference ( $\Omega$ ) updates each AU prediction in an iterative manner.
- Mutual relationships are controlled by a gating strategy.
- $\Omega$  is a collection of interconnected recurrent structure interference units  $\Omega_j$  one for each AU, defined as:

$$\hat{y}_j^t, m_j^t = \Omega_j(f_j, m_1^{t-1}, m_2^{t-1}, \dots, m_N^{t-1}, \hat{y}_j^{t-1}; \omega_j)$$

$$m_j^t = \sigma(\omega_j^m [\mu(m_1^{t-1}, \dots, m_N^{t-1}), f_j, \hat{y}_j^{t-1}] + \beta_j^m)$$

$$\hat{y}_j^t = \sigma(\omega_j^y [\mu(m_1^t, \dots, m_N^t), f_j] + \beta_j^y)$$

## RESULTS

### Ablation Study

- Class balancing** improves performance, especially on poorly represented classes.
- Targeting subsets of AUs** On average and across patches training on groups of AUs or on all AUs is beneficial as correlation information between classes is employed by the network in the fully connected layers.
- Learning Local Representations.** Face prediction compared to patch prediction performs better on the entire output set. However, when individual AUs are considered, this is no longer the case.
- Patch Learning.** through fusion is beneficial on both tested datasets, but on DISFA benefits are higher.
- Structure Learning** is beneficial for both datasets but for DISFA, the results are even more conclusive adding more than 5% improvement over the fusion.

method	AU01	AU02	AU04	AU06	AU07	AU10	AU12	AU14	AU15	AU17	AU23	AU24	avg
VGG(face) <sup>ft</sup>	35.2	31.2	25.4	73.1	72.1	80.1	59.2	35.1	32.1	52.3	26.1	36.2	46.5
PP(face) <sup>ncb</sup>	35.1	38.1	53.9	77.2	70.7	83.1	86.2	56.1	39.8	54.5	37.2	31.4	55.3
PP(right eye) <sup>ind</sup>	46.8	40.4	45.3	68.3	69.2	-	-	-	-	-	-	-	-
PP(mouth) <sup>ind</sup>	-	-	-	-	-	78.6	82.0	54.2	38.6	54.7	[39.3]	43.3	-
PP(right eye)	38.0	[37.7]	48.3	69.5	71.0	72.4	77.4	50.7	15.0	38.9	13.8	15.3	45.7
PP(between eye)	41.7	34.8	45.9	64.9	65.5	72.1	73.9	54.9	19.7	33.9	13.9	7.0	44.0
PP(mouth)	12.4	7.3	22.4	75.5	70.5	78.9	81.3	66.2	35.8	59.6	37.6	[42.8]	49.3
PP(right cheek)	30.5	18.4	41.8	75.2	73.2	79.1	81.9	[61.9]	35.7	55.1	35.5	35.7	52.0
PP(nose)	41.6	28.4	46.4	71.1	70.5	78.8	78.0	57.1	21.3	43.7	34.0	20.3	49.3
PP(face)	43.8	37.5	[54.9]	77.4	[71.2]	[79.2]	84.0	56.6	[39.7]	[59.7]	39.2	39.5	[56.9]
PP+F	[44.8]	35.8	57.1	[76.7]	74.3	79.6	[83.7]	56.6	41.1	61.8	42.2	40.1	57.8
DSIN <sub>2</sub> <sup>ncf</sup>	46.7	34.1	62.0	76.5	74.1	[83.1]	84.9	60.9	36.0	57.1	43.3	36.1	57.9
DSIN <sub>2</sub>	47.7	36.5	55.6	76.3	[73.7]	80.1	85.0	64.0	[39.2]	60.6	[43.1]	39.9	58.2
DSIN <sub>5</sub>	[49.7]	36.3	57.3	76.8	73.4	81.6	84.5	[64.7]	38.5	[63.0]	39.0	37.3	58.5
DSIN <sub>10</sub>	51.7	[40.4]	56.0	76.1	73.5	79.9	[85.4]	62.7	37.3	62.9	38.6	[41.6]	[58.9]
DSIN <sup>tr</sup>	51.7	41.6	[58.1]	[76.6]	74.1	85.5	87.4	72.6	40.4	66.5	38.6	46.9	61.7

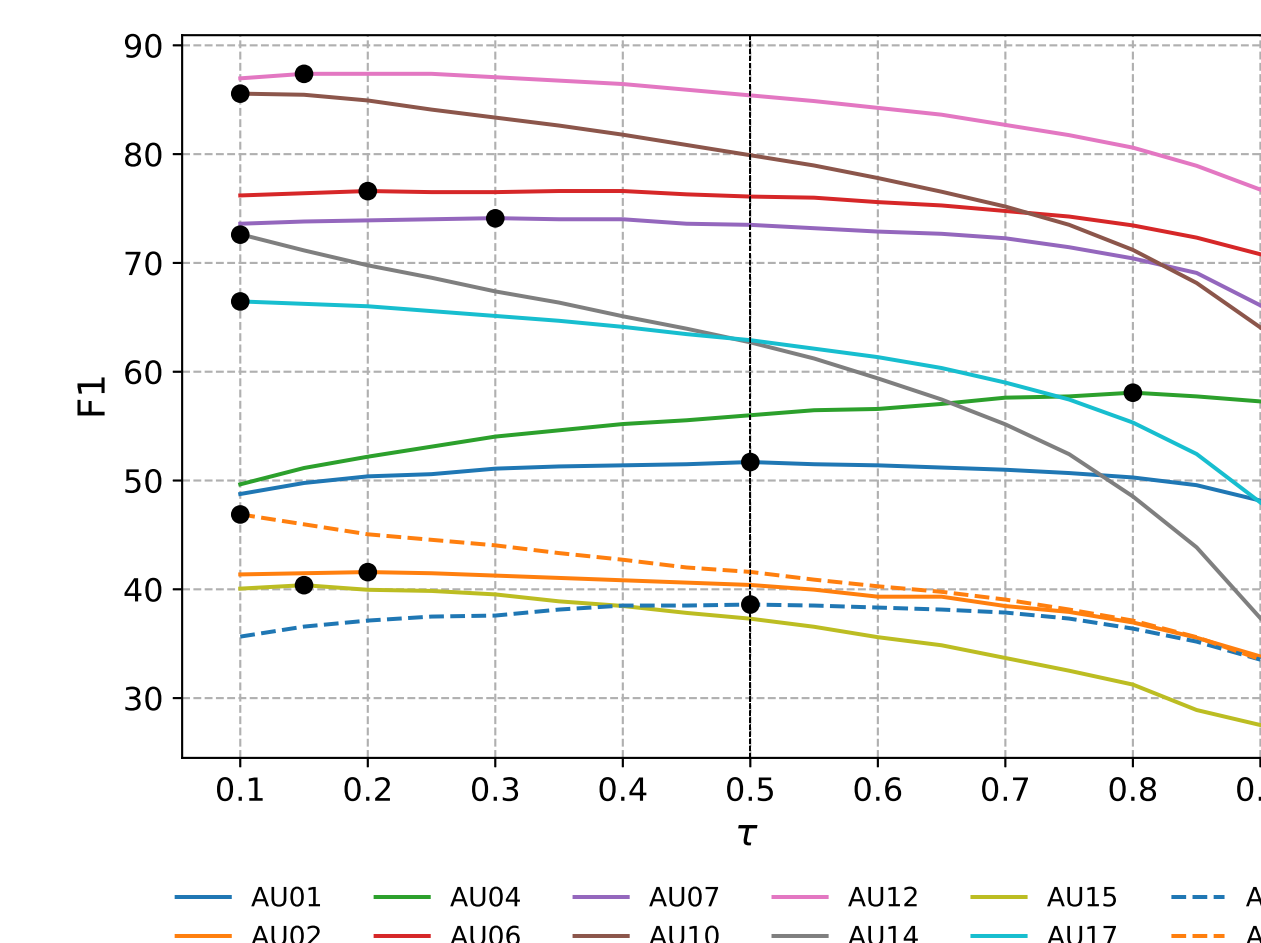
**Table 1:** Ablation study on BP4D.

### Comparison with State-of-the-Art

method	AU01	AU02	AU04	AU06	AU07	AU10	AU12	AU14	AU15	AU17	AU23	AU24	AVG
JPML [1]	32.6	25.6	37.4	42.3	50.5	72.2	74.1	[65.7]	38.1	40.0	30.4	[42.3]	45.9
DRML [2]	36.4	41.8	43.0	55.0	67.0	66.3	65.8	54.1	33.2	48.0	31.7	30.0	48.3
CPM [3]	[43.4]	40.7	43.3	59.2	61.3	62.1	68.5	52.5	36.7	54.3	39.5	37.8	50.0
ROI [4]	36.2	31.6	43.4	77.1	[73.7]	[85.0]	62.6	48.7	58.0	38.3	37.4	56.4	56.4
DSIN	51.7	40.4	[56.0]	76.1	73.5	79.9	85.4	62.7	37.3	[62.9]	[38.8]	41.6	[58.9]
DSIN <sup>tr</sup>	51.7	41.6	58.1	[76.6]	74.1	85.5	87.4	72.6	[40.4]	66.5	38.6	46.9	61.7

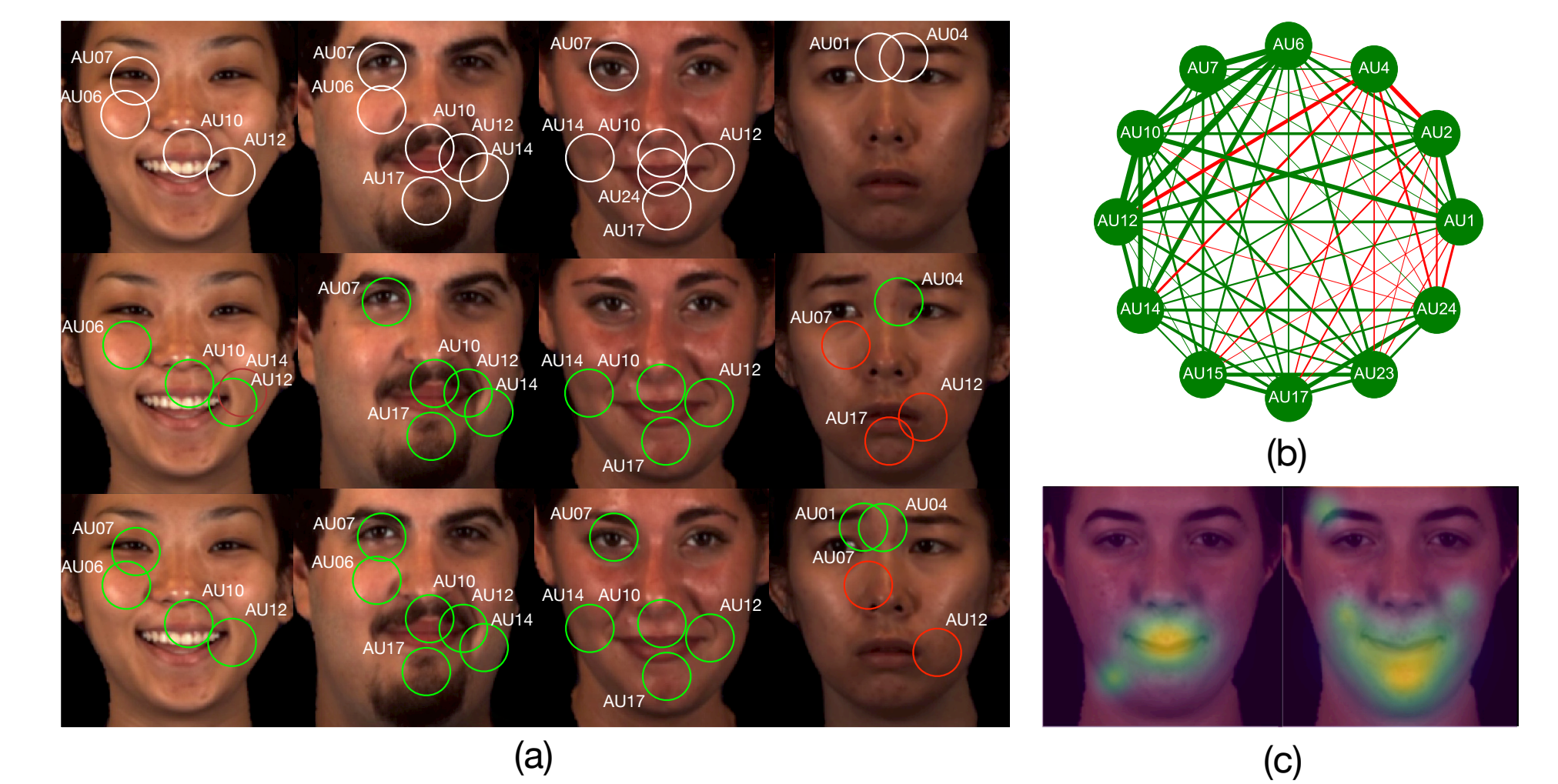
**Table 2:** AU recognition results on BP4D.

### Threshold Tuning



**Figure 4:**  $\tau$  vs AU performance on BP4D validation set. Black circles denote best score.

### Qualitative Results



**Figure 5:** (a) Examples of AU predictions: ground-truth (top), fusion module (middle) and structure inference (bottom) prediction (green dots: true positive, red dots: false positive). (b) AUs correlation in BP4D (green dots: positive, red dots: negative). Line thickness is proportional with correlation magnitude. (c) Class activation map for AU24 that shows the discriminative regions of simple patch prediction (left) and DSIN (right).