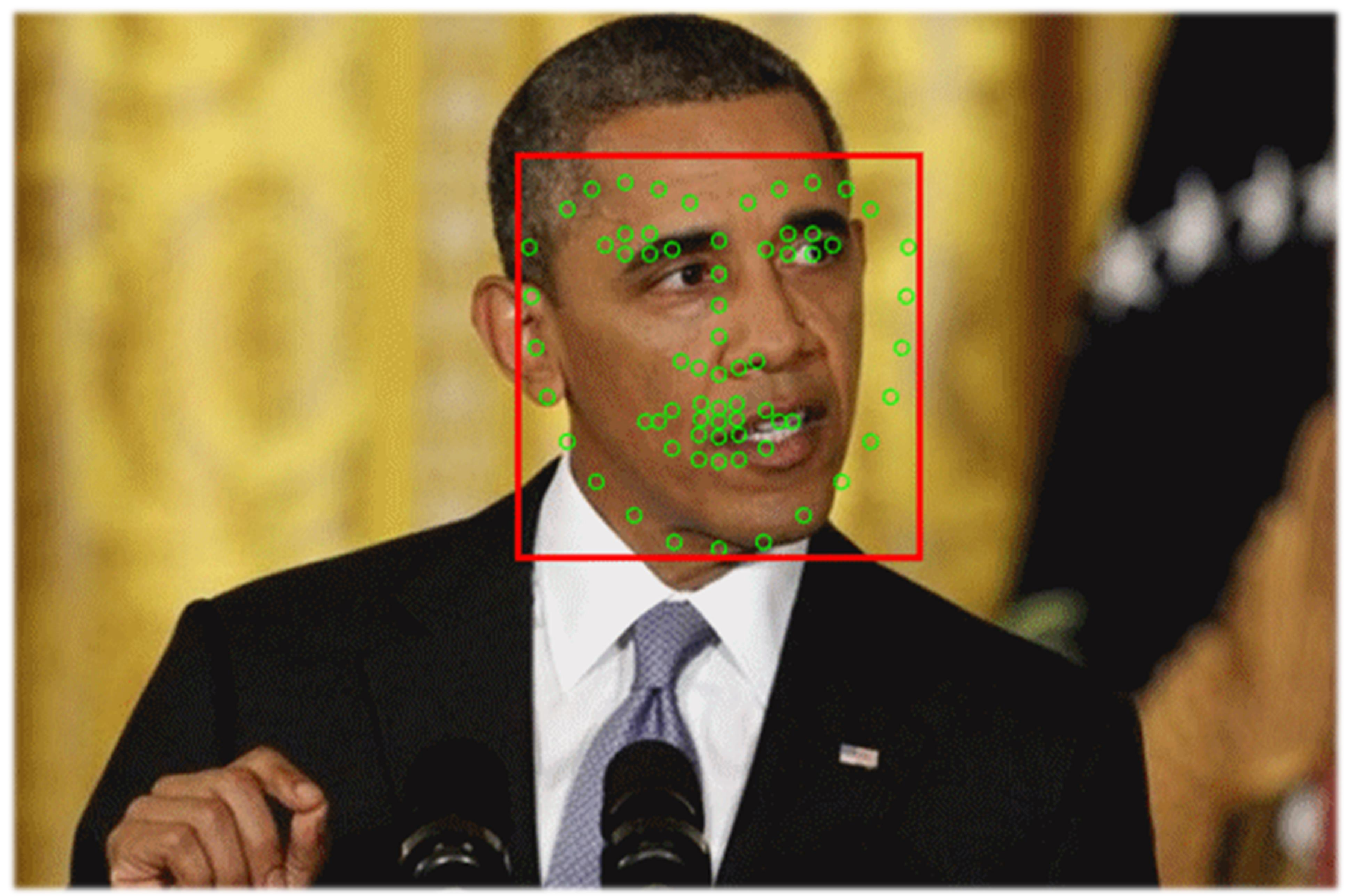
Robust Facial Landmark Detection via Recurrent Attentive-Refinement Networks

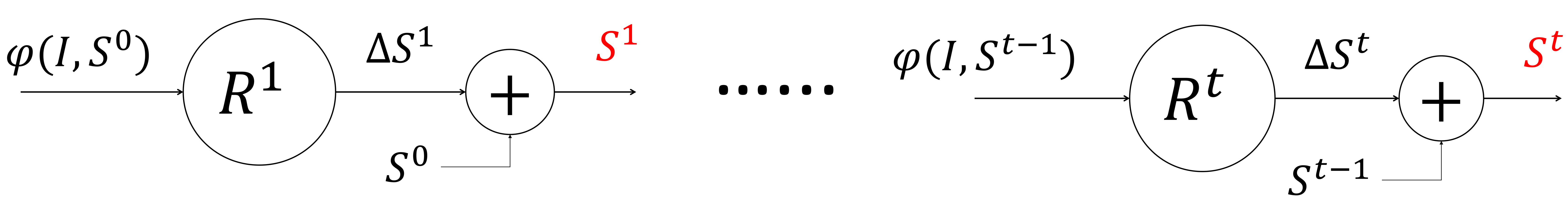
Shengtao XIAO, Jiashi FENG, **Junliang XING**, Hanjiang LAI,
Shuicheng YAN, Ashraf KASSIM

Problem Introduction

- Obtain face shape by locating pre-defined facial landmarks.

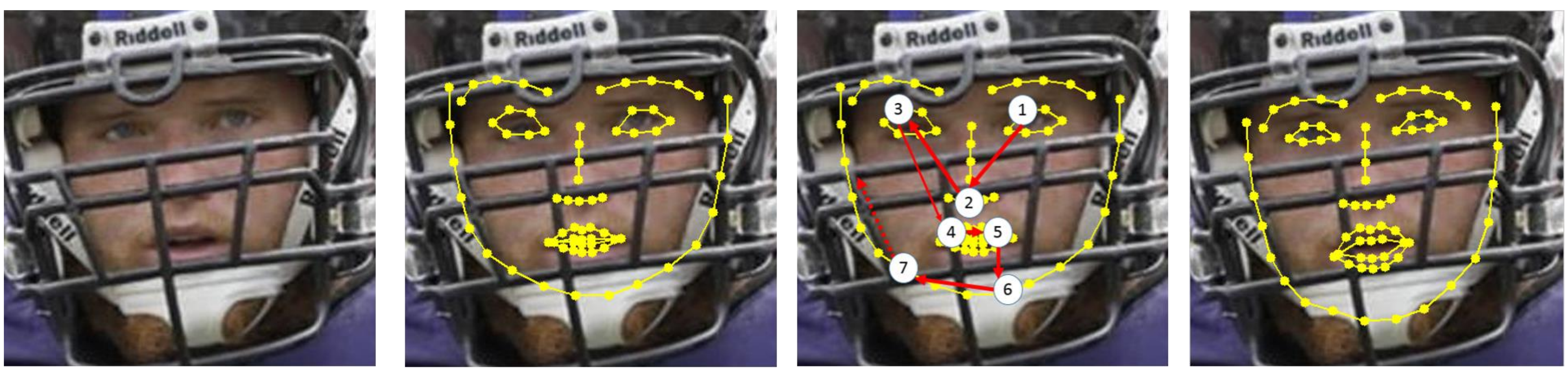


- Challenges: face occlusions, pose variations, expressions, etc.
- Solutions: cascaded face shape regression



Recurrent Attentive-Refinement (RAR) Network

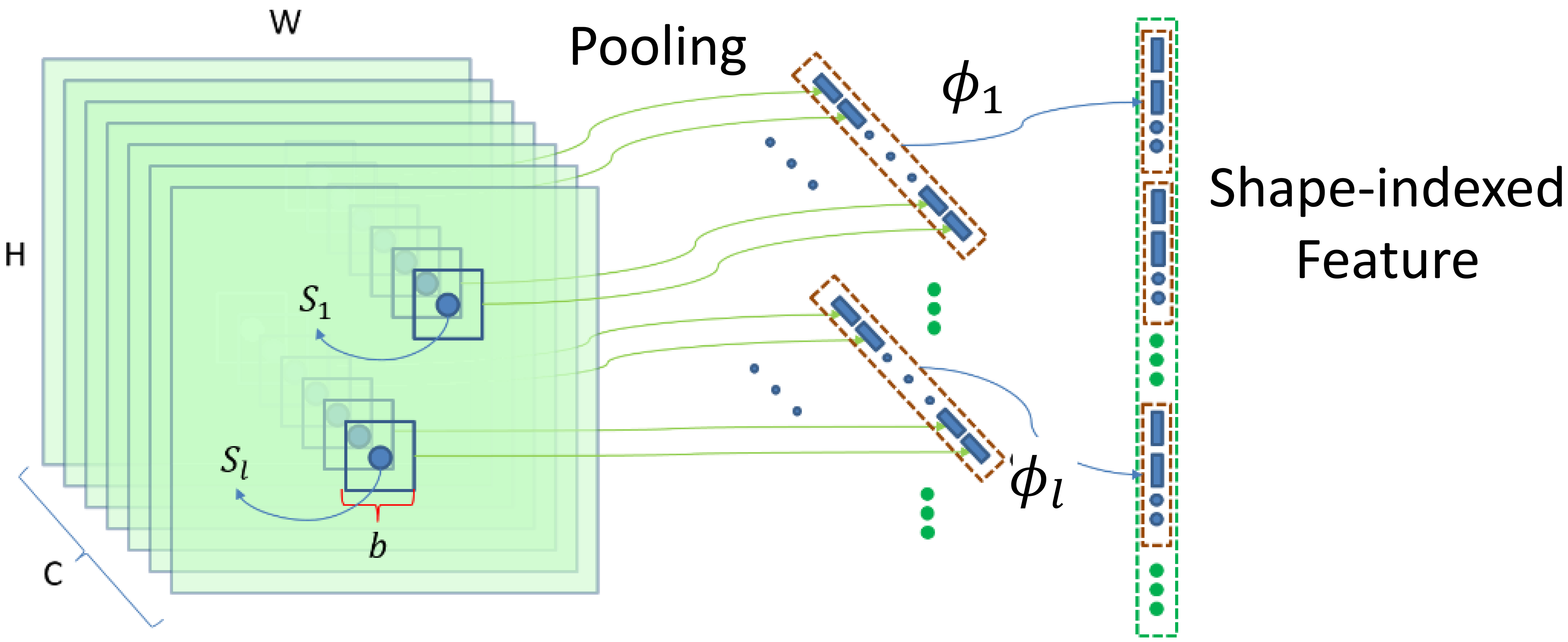
- Deep Feature Learning
- Robust Shape Initialization
- Recurrent-Attentive Refinement
  - Attention module
  - Refinement module



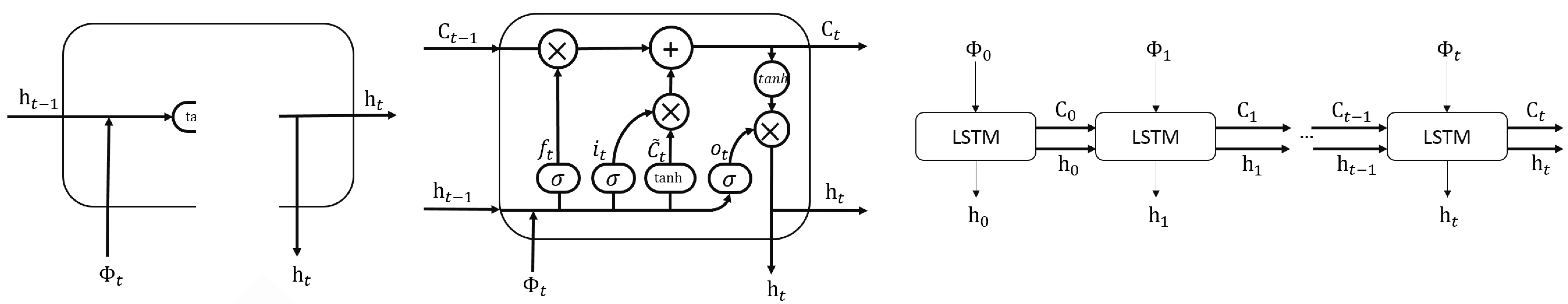
Input | Robust Initialization | Attention-driven Refinement | Results

Background

- CNN model with Shape-Indexed Pooling (SIP)

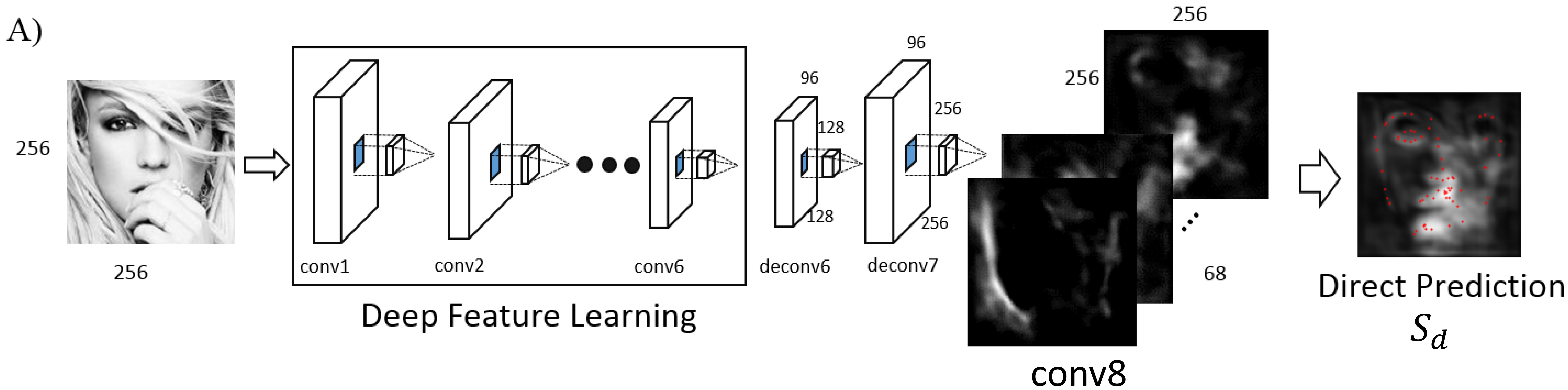


- RNN model and Long Short-Term Memory (LSTM)

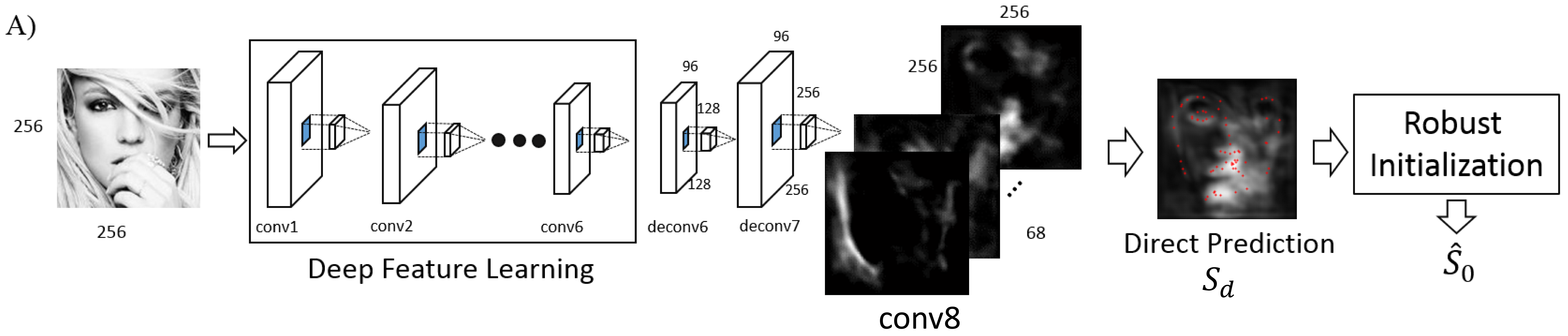


Standard LSTM Unrolled LSTM

RAR Networks

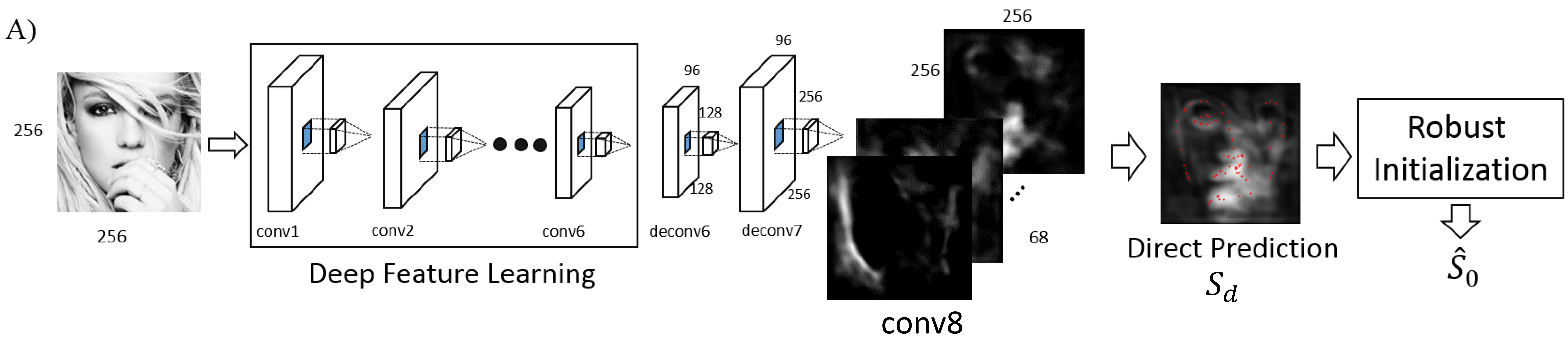


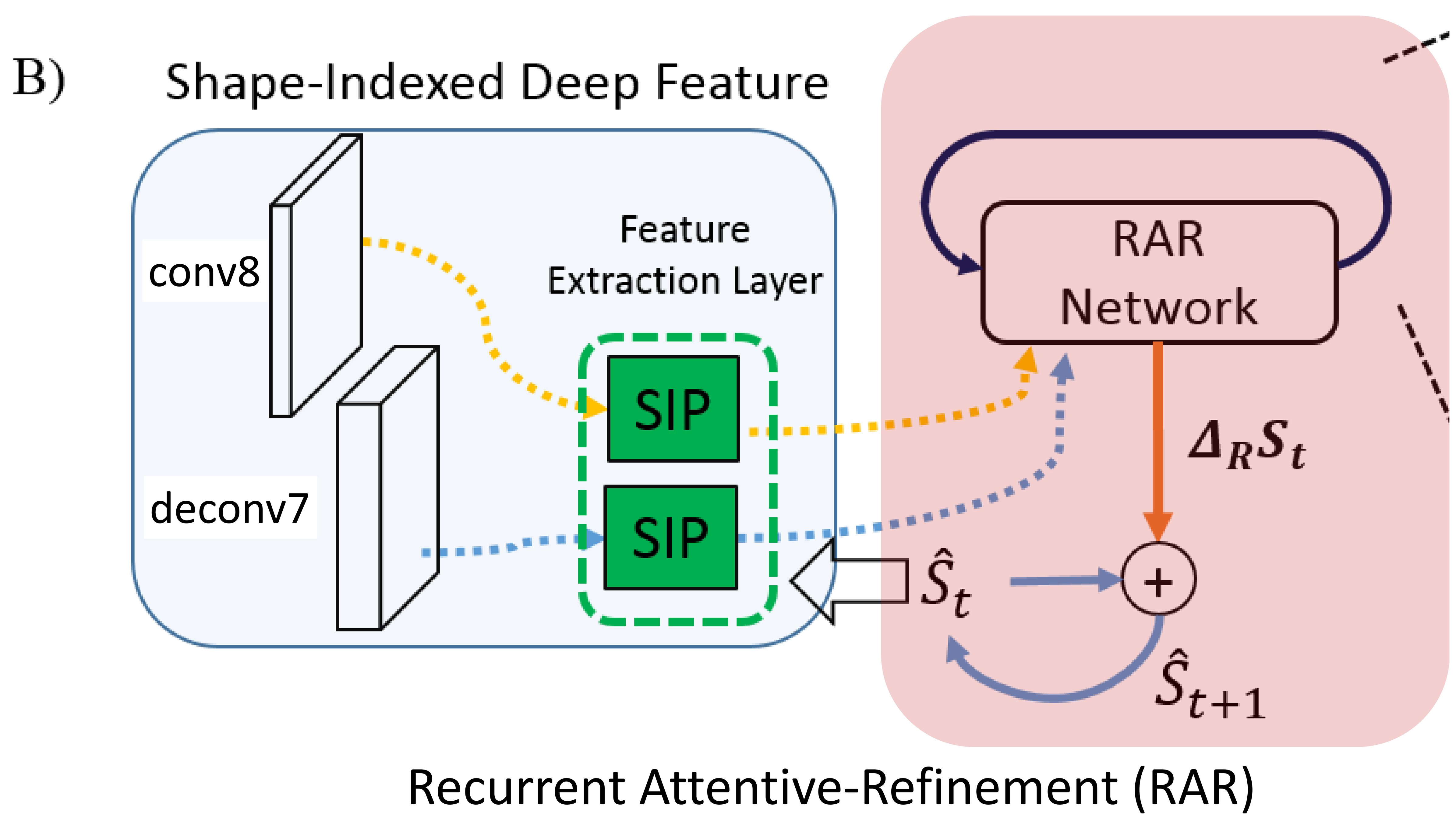
RAR Networks



A). Deep feature extraction, landmark regression and robust initialization.

RAR Networks

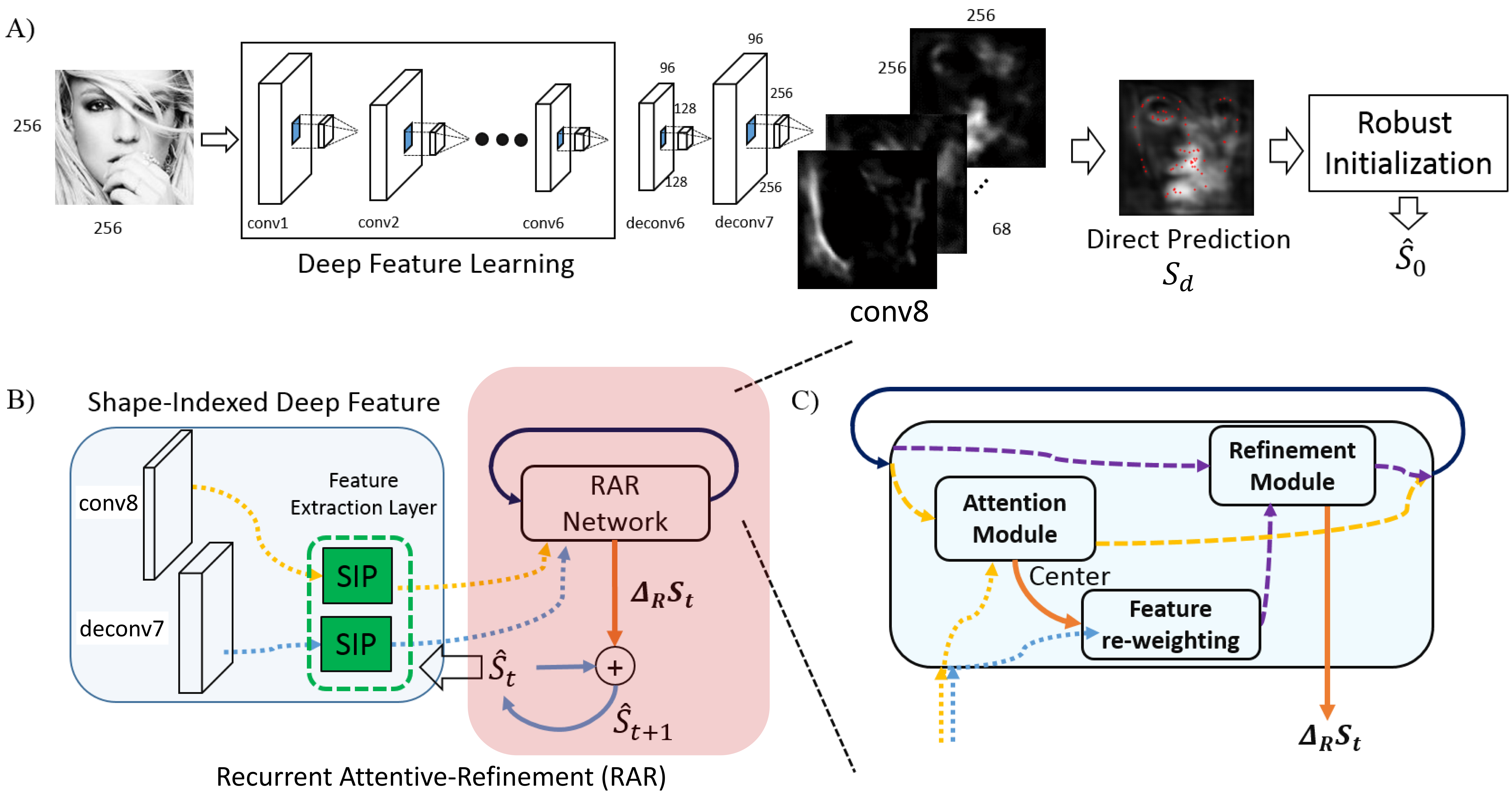




A). Deep feature extraction, landmark regression and robust initialization.

B). RAR sequentially refines the landmark estimation.

RAR Networks



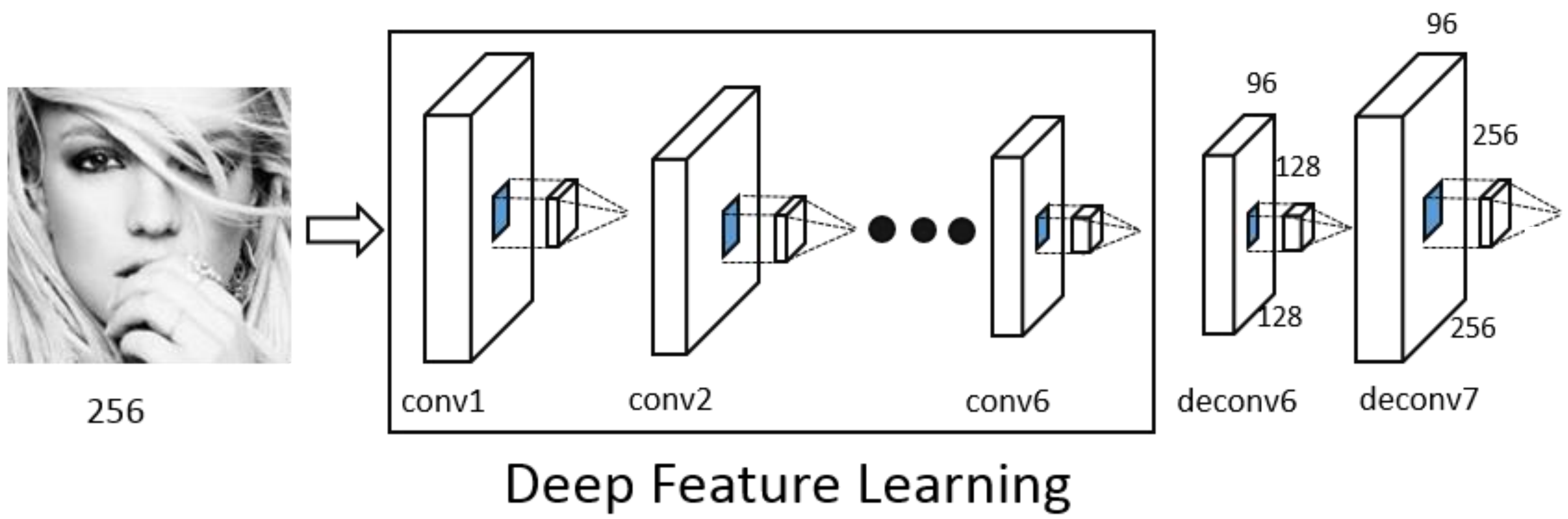
A). Deep feature extraction, landmark regression and robust initialization.

B). RAR sequentially refines the landmark estimation.

C). An attention model in RAR for adaptively selecting key landmark points.

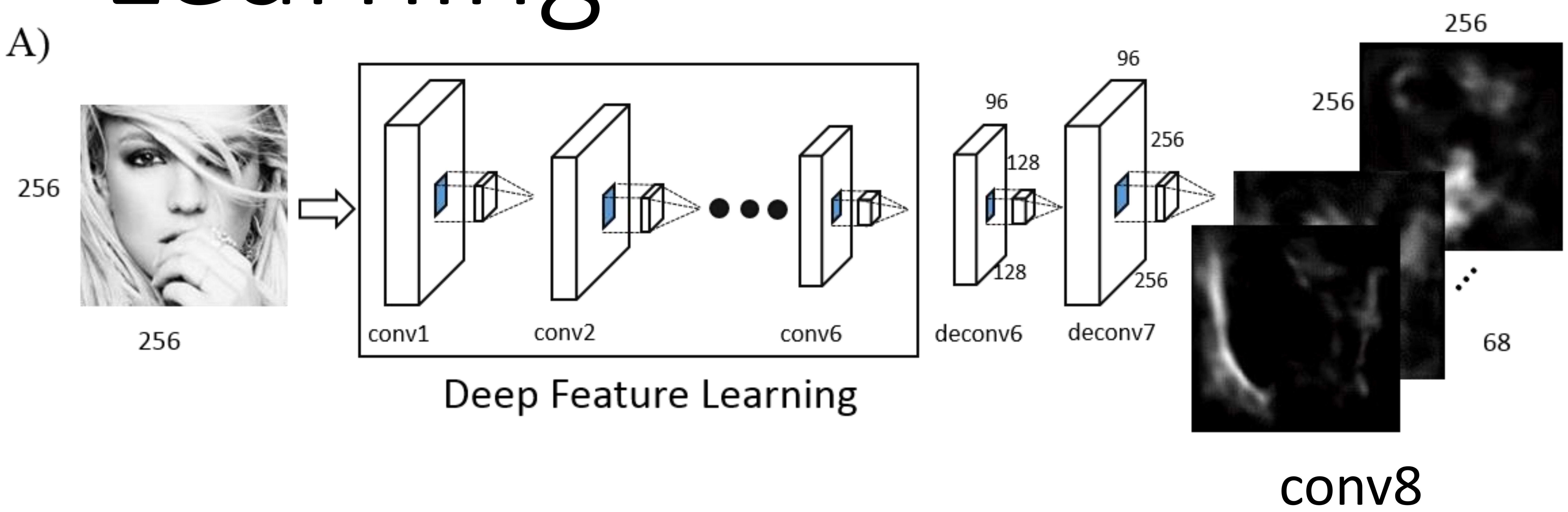
RAR Networks: Deep Feature Learning

A)



- Modified VGG19 Network + two Deconvolution layers to ensure pixel-to-pixel correspondence

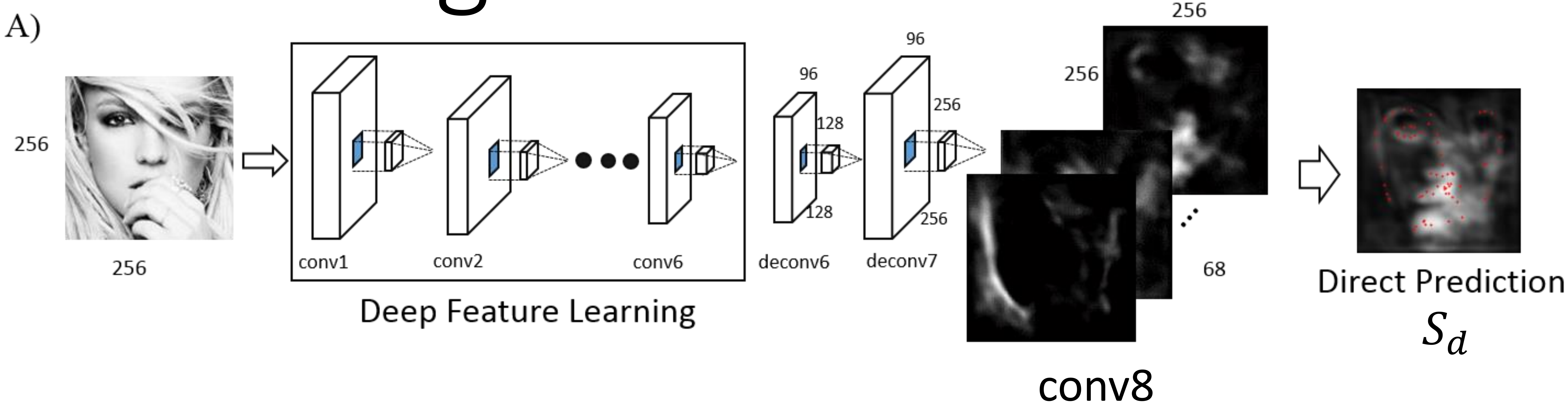
RAR Networks: Deep Feature Learning



- Modified VGG19 Network + two Deconvolution layers to ensure pixel-to-pixel correspondence

- SoftMax regression loss on **conv8**

RAR Networks: Deep Feature Learning



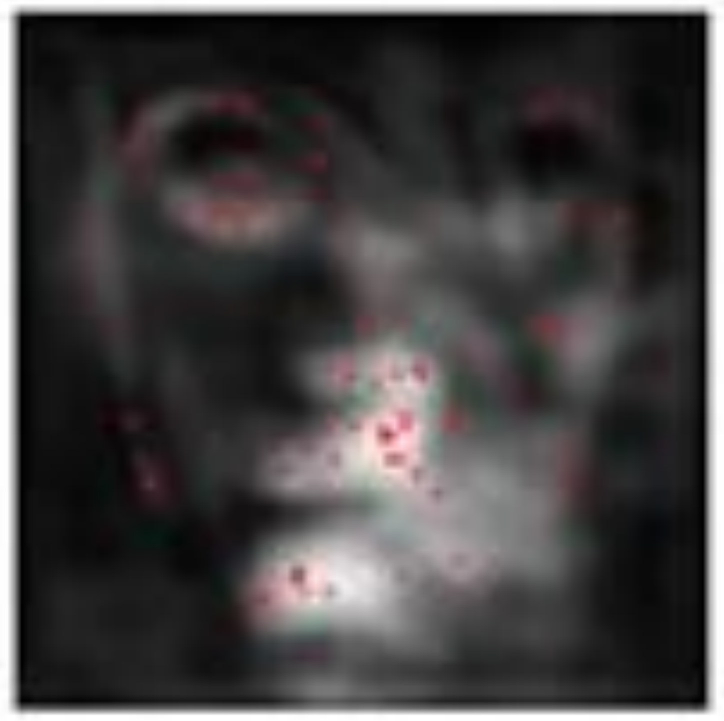
- Modified VGG19 Network + two Deconvolution lby selecting location of maximum response from v−th channel of **conv8**

- SoftMax regression loss on **conv8**

- Directly estimate landmark location $S\ d\ v\quad d\ v\quad d\ v\ d\ v\ d\ v$ by selecting location of maximum response from v−th channel of **conv8**

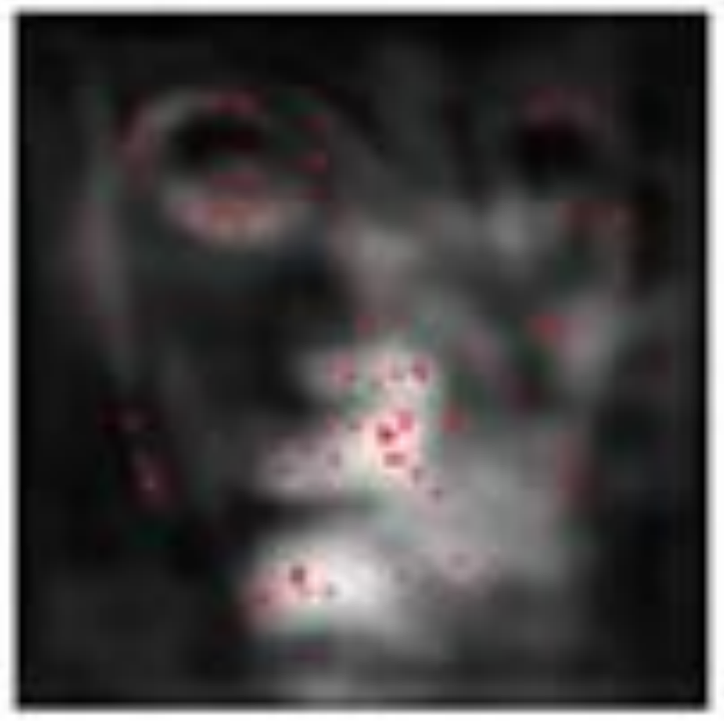
RAR Networks: Robust Shape Initialization

However, detected shape is sensitive to occlusion



RAR Networks: Robust Shape Initialization

However, detected shape is sensitive to occlusion



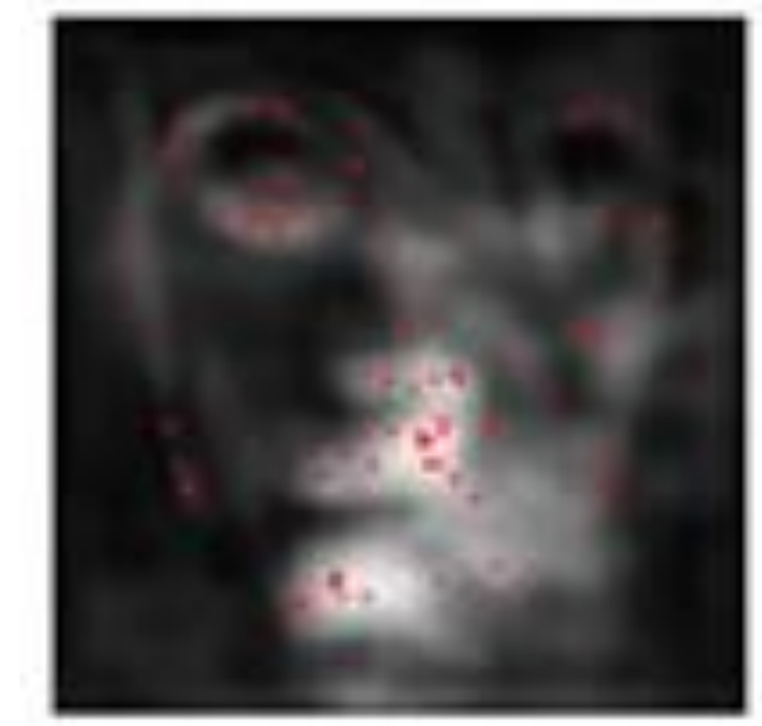
Robust Initial Shape Selection:

$$S_0 = \arg\min_{S} ||S - \boxed{S_d}||, \text{s.t. } S \in \mathcal{F}$$

Detected Shape

RAR Networks: Robust Shape Initialization

However, detected shape is sensitive to occlusion



Robust Initial Shape Selection:

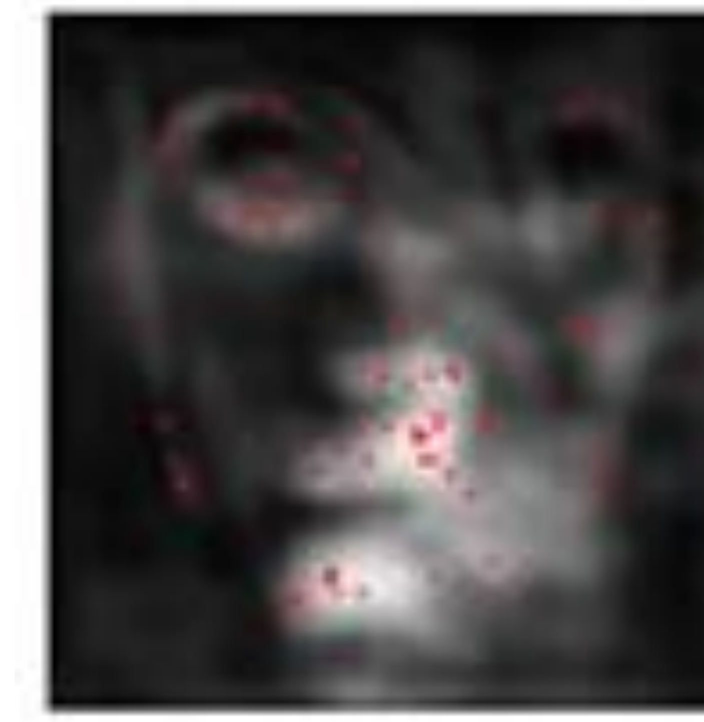
$$S_0 = \arg\min_{S} ||S - S_d||, \text{s.t. } S \in \mathcal{F}$$

$$S_0 = \arg\min_{S,\, c \doteq [c_i]} ||S - S_d||_0 + \lambda||c||_0, \text{s.t. } S = \sum_{i}^{m} c_i \boxed{S_i}$$

GT Shapes

RAR Networks: Robust Shape Initialization

However, detected shape is sensitive to occlusion



Robust Initial Shape Selection:

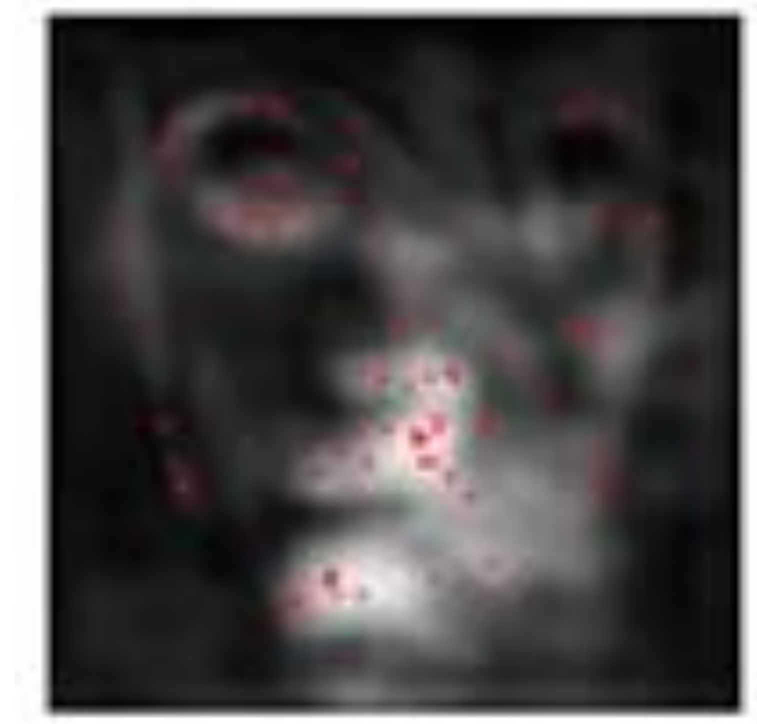
$$S_0 = \arg\min_{S} ||S - S_d||, \text{s.t. } S \in \mathcal{F}$$

$$S_0 = \arg\min_{S, c \doteq [c_i]} ||S - S_d||_0 + \lambda||c||_0, \text{s.t. } S = \sum_{i}^{m} c_i S_i$$

Solve: get K representative shapes via K-meanings clustering + RANSAC method to filter out significant outliers

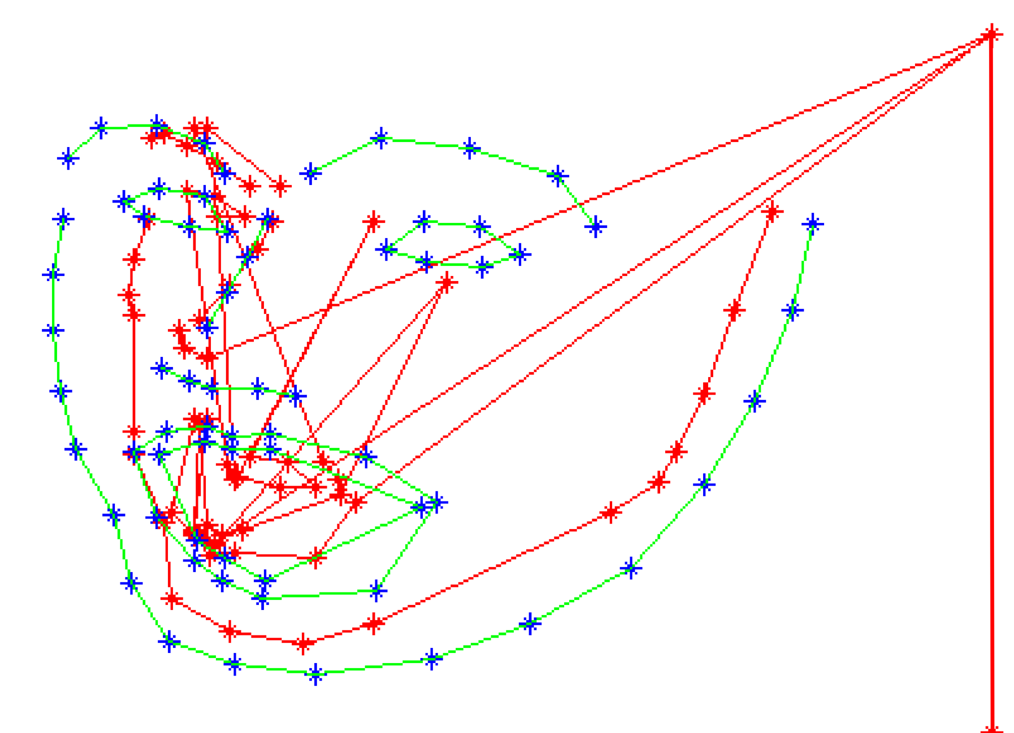
RAR Networks: Robust Shape Initialization

However, detected shape is sensitive to occlusion

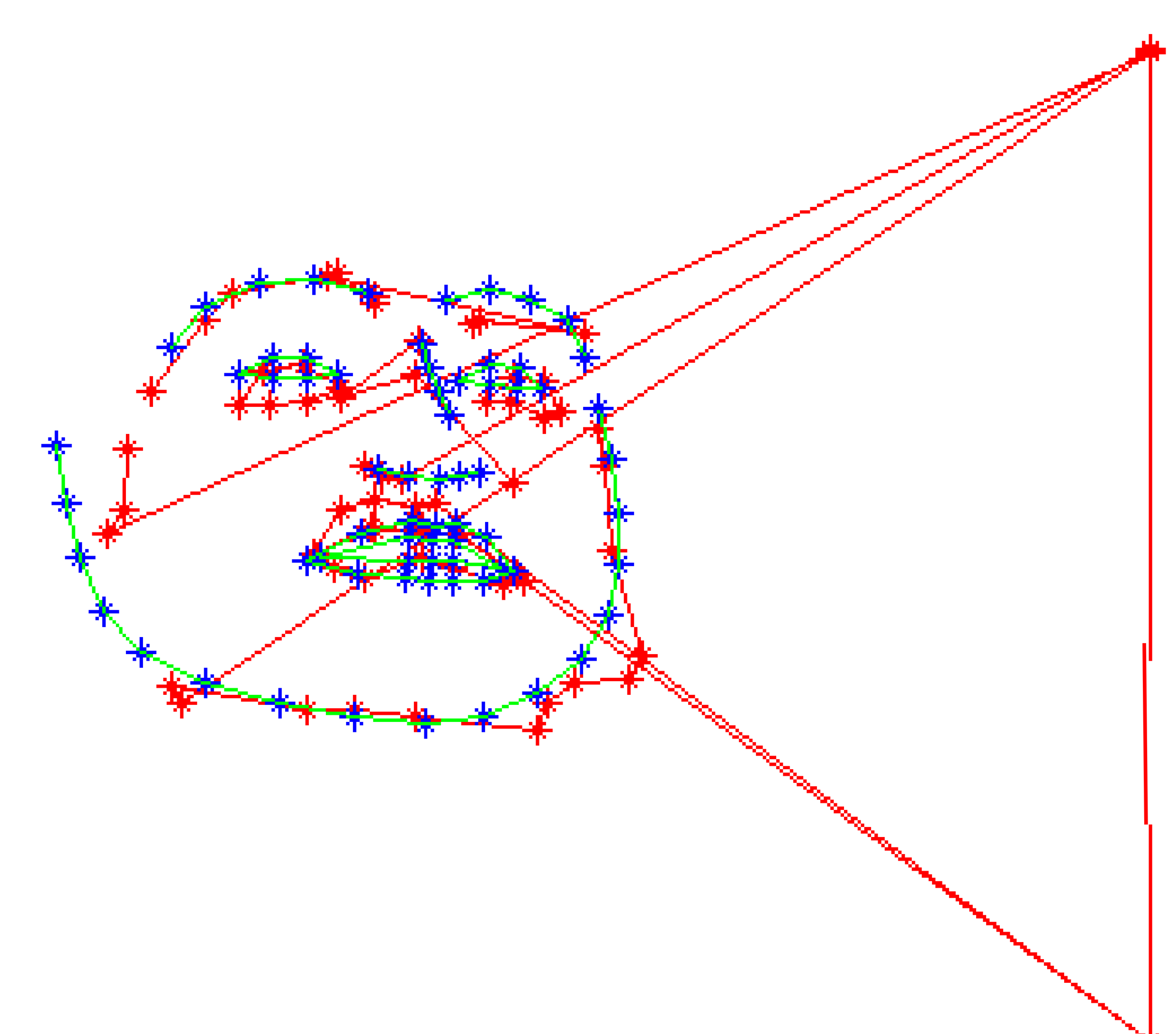


Robust Initial Shape Selection:

$$S_0 = \arg\min_{S} ||S - S_d||, \mathrm{s.t.}\ S \in \mathcal{F}$$



$$S_0 = \arg\min_{S,\, c \doteq [c_i]} ||S - S_d||_0 + \lambda||c||_0, \mathrm{s.t.}\ S = \sum_{i}^{m} c_i S_i$$



Solve: get K representative shapes via K-meanings clustering + RANSAC method to filter out significant outliers

RAR Networks: Attention Module

-

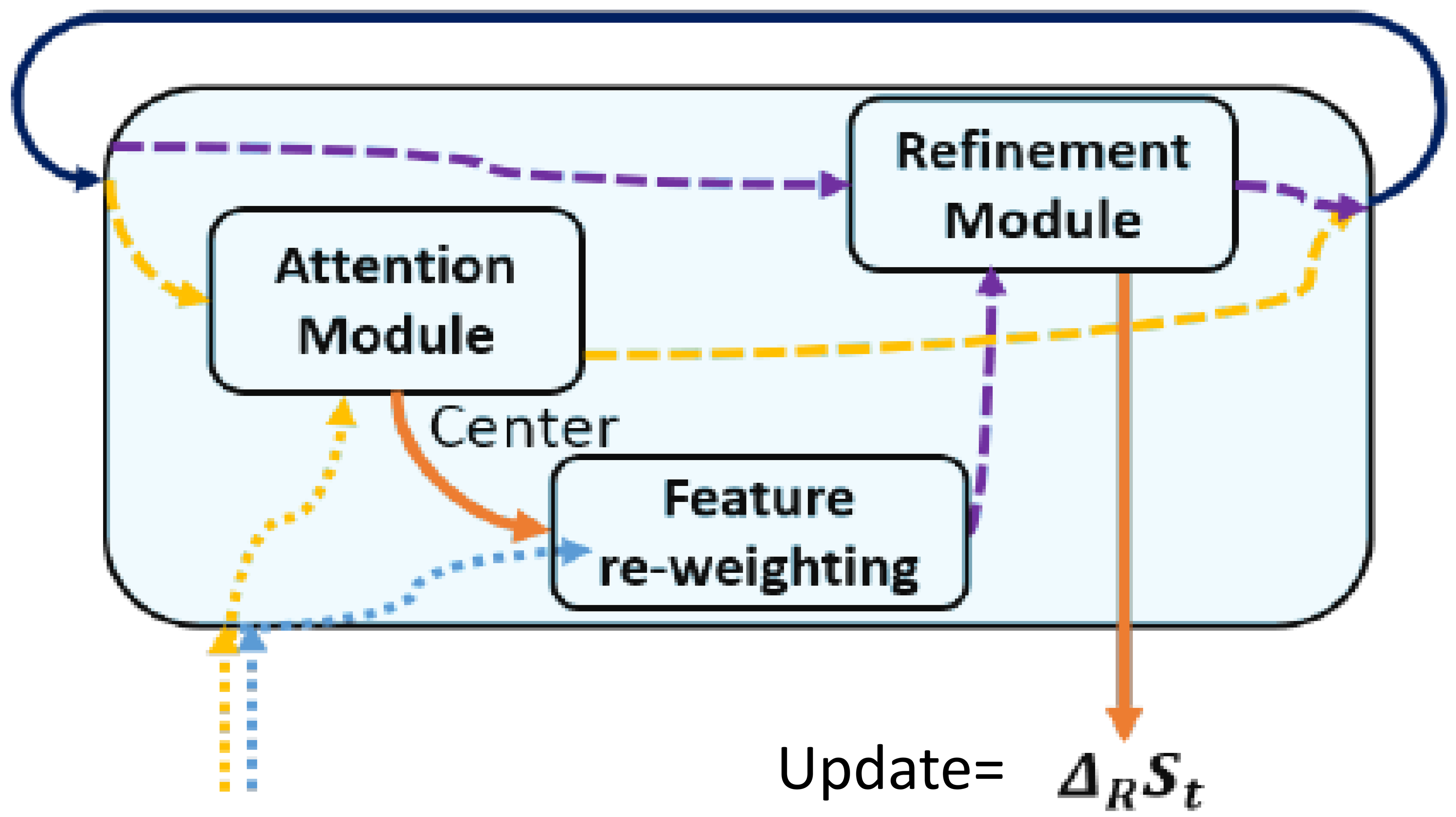
RAR Networks: Attention Module

- A-LSTM (attention module) selects attention center with top confidence at each recurrent stage

$$C^* = \underset{c \in \{1,\ldots,L\}}{\operatorname{argmax}} \text{A−LSTM}(\boxed{\Phi_a\left(I_t, \hat{S}_t\right)}; W_a, c)$$

Deep SIP Feature

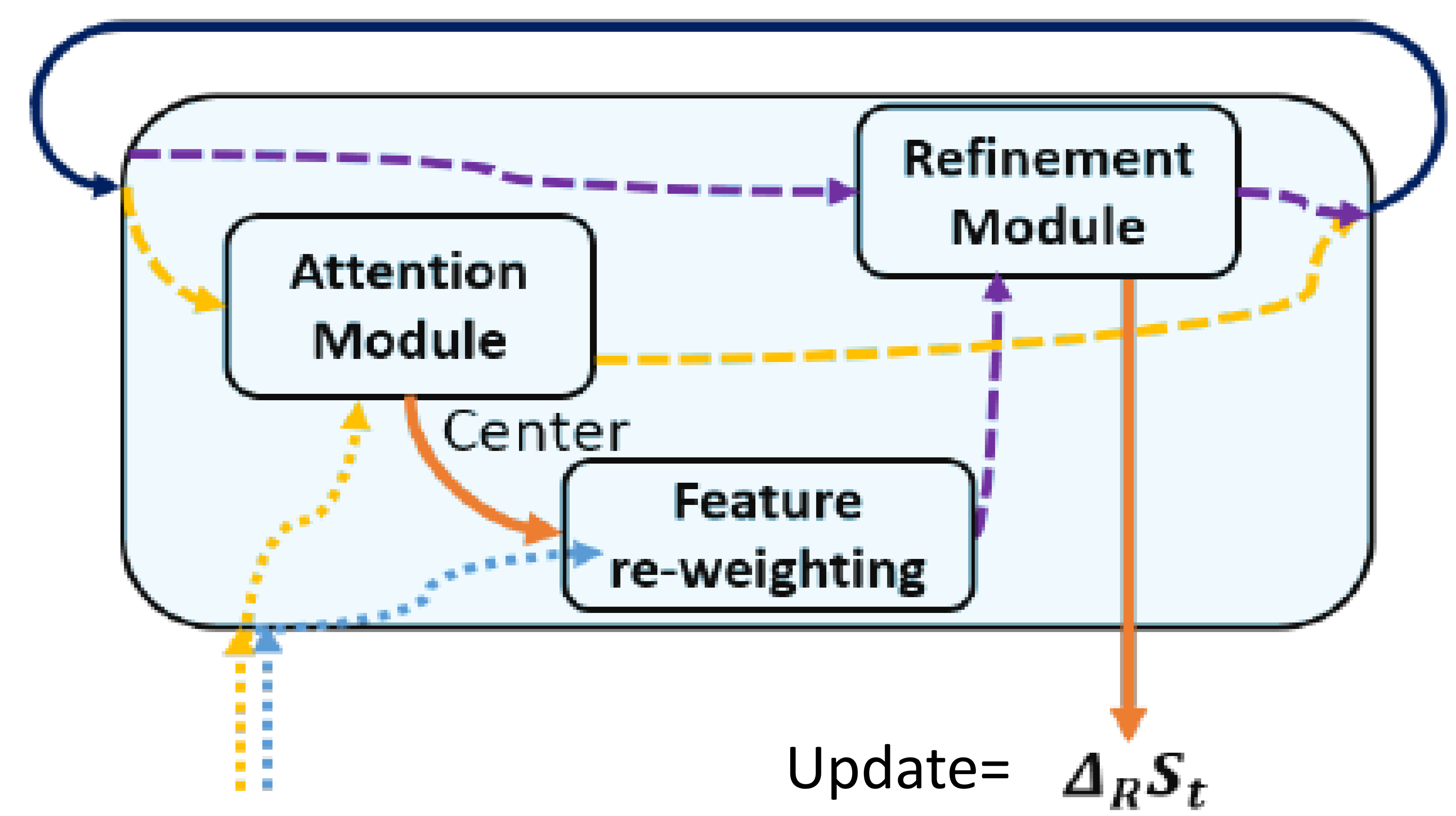
-



RAR Networks: Attention Module

- $\mathcal{R}\ a\ aa\ \mathcal{R}\ a$
- A-LSTM (attention module) selects attention center with top confidence at each recurrent stage

$$c^* = \underset{c \in \{1,\ldots,L\}}{\operatorname{argmax}} \text{A-LSTM}(\Phi_a(I_t, \hat{S}_t); W_a, c)$$



- A typical attention center is selected based on maximize reward $\mathcal{R}\ a$

$$\mathcal{R}_a = \sum_{t=1}^{\infty} \boxed{\eta^{t-1}} \boxed{R(\hat{S}_{t-1}, \hat{S}_t)}$$

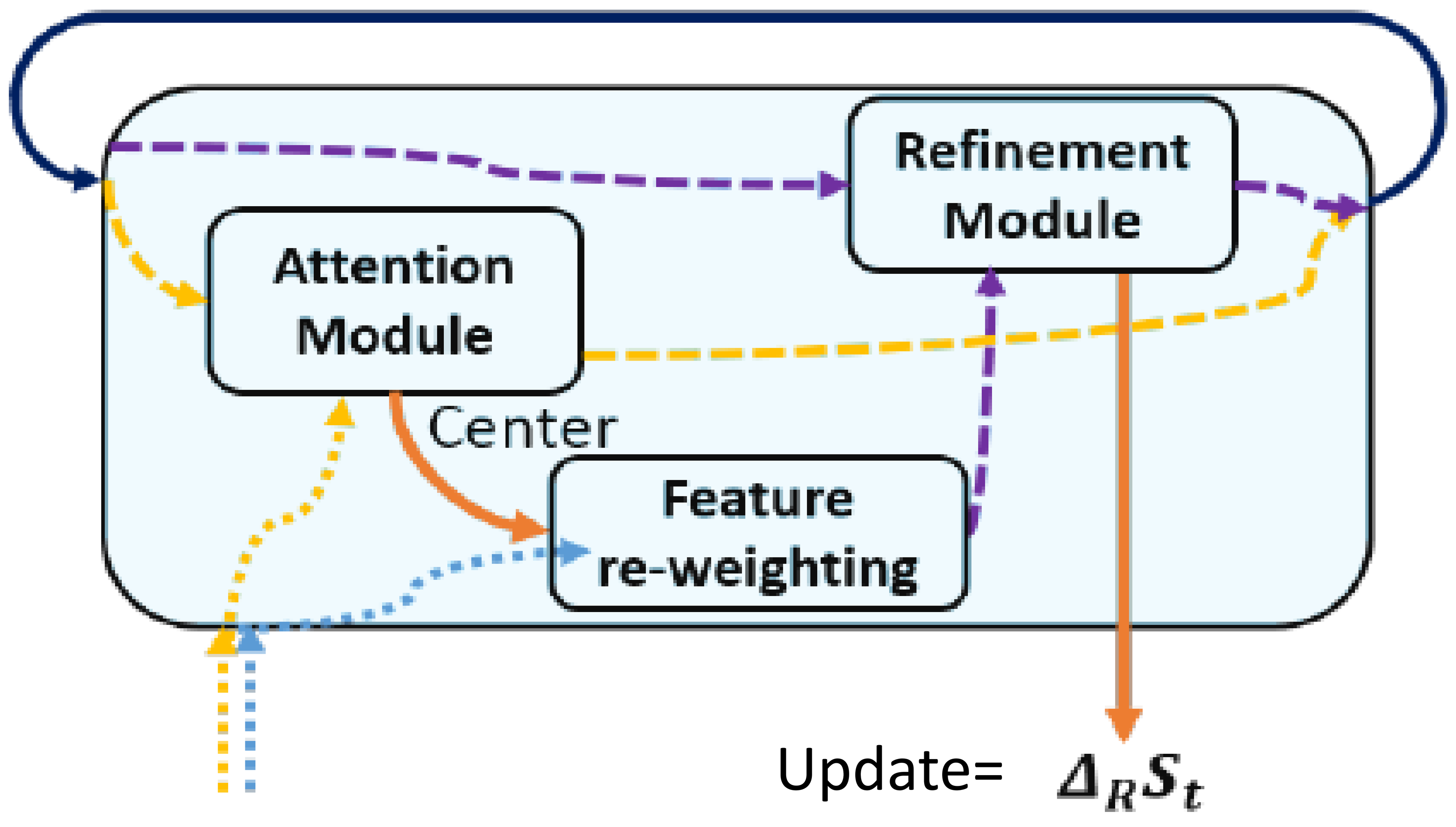
Intermediate Reward

Discount Factor

RAR Networks: Attention Module

- $\mathcal{R}\ a\ aa\ \mathcal{R}\ a$
- A-LSTM (attention module) selects attention center with top confidence at each recurrent stage

$$c^* = \operatorname*{argmax}_{c\in\{1,\ldots,L\}} \text{A-LSTM}(\Phi_a(I_t, \hat{S}_t); W_a, c)$$



- A typical attention center is selected based on maximize reward $\mathcal{R}\ a$

$$\mathcal{R}_a = \sum_{t=1}^{\infty} \eta^{t-1} R(\hat{S}_{t-1}, \hat{S}_t)$$

Weighting Factor

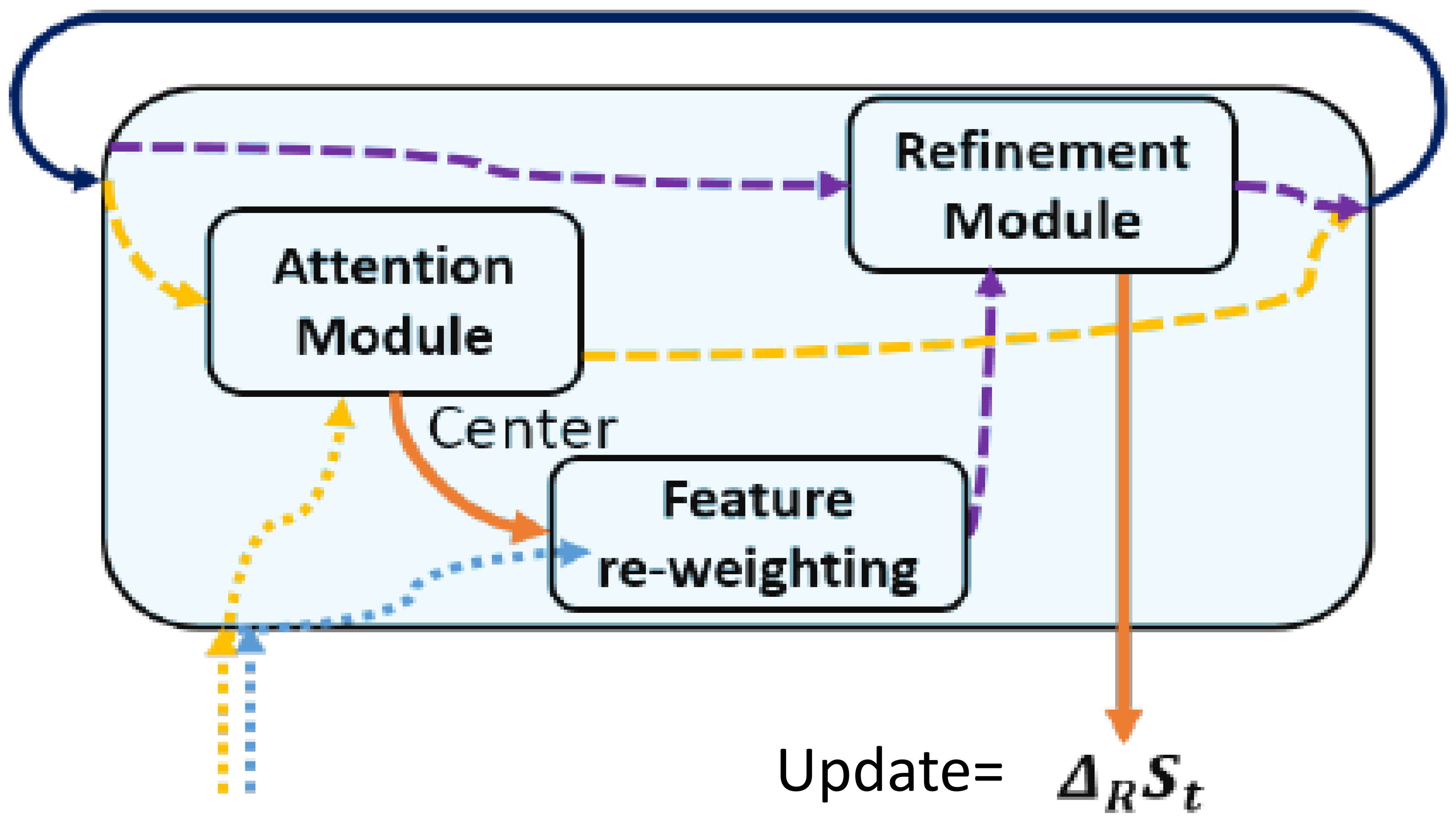
$$R(\hat{S}_{t-1}, \hat{S}_t) = ||\Gamma_t \Delta S_{t-1}||_2^2 - ||\Gamma_t \Delta S_t||_2^2$$

Ground-Truth Residue

RAR Networks: Attention Module

- $\mathcal{R}\, a\, aa\, \mathcal{R}\, a$
- A-LSTM (attention module) selects attention center with top confidence at each recurrent stage

$$c^* = \underset{c\in\{1,\ldots,L\}}{\operatorname{argmax}}\ \text{A-LSTM}(\Phi_a(I_t, \hat{S}_t); W_a, c)$$



- A typical attention center is selected based on maximize reward $\mathcal{R}\, a$

$$\mathcal{R}_a = \sum_{t=1}^{\infty} \eta^{t-1} R(\hat{S}_{t-1}, \hat{S}_t)$$

$$R(\hat{S}_{t-1}, \hat{S}_t) = ||\Gamma_t \Delta S_{t-1}||_2^2 - ||\Gamma_t \Delta S_t||_2^2$$

$$\Gamma_t = [\gamma_t^1, \gamma_t^2, \ldots, \gamma_t^L], \text{ with } \gamma_t^l = \kappa \exp\left(\frac{-||\hat{S}_t^l - \hat{S}_t^{c^*}||_{l_2}^2}{4D_t^2}\right)$$

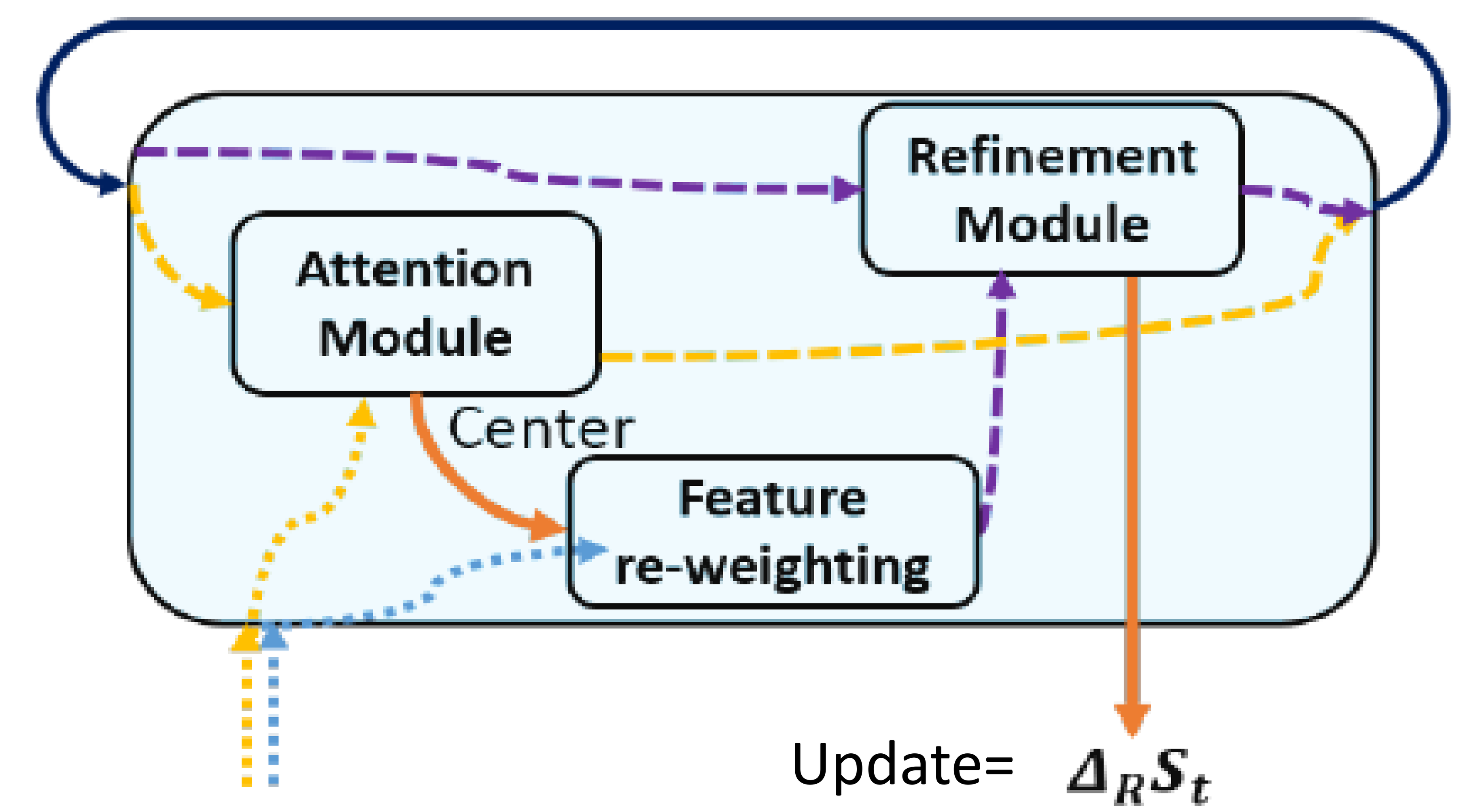
RAR Networks: Refinement Module

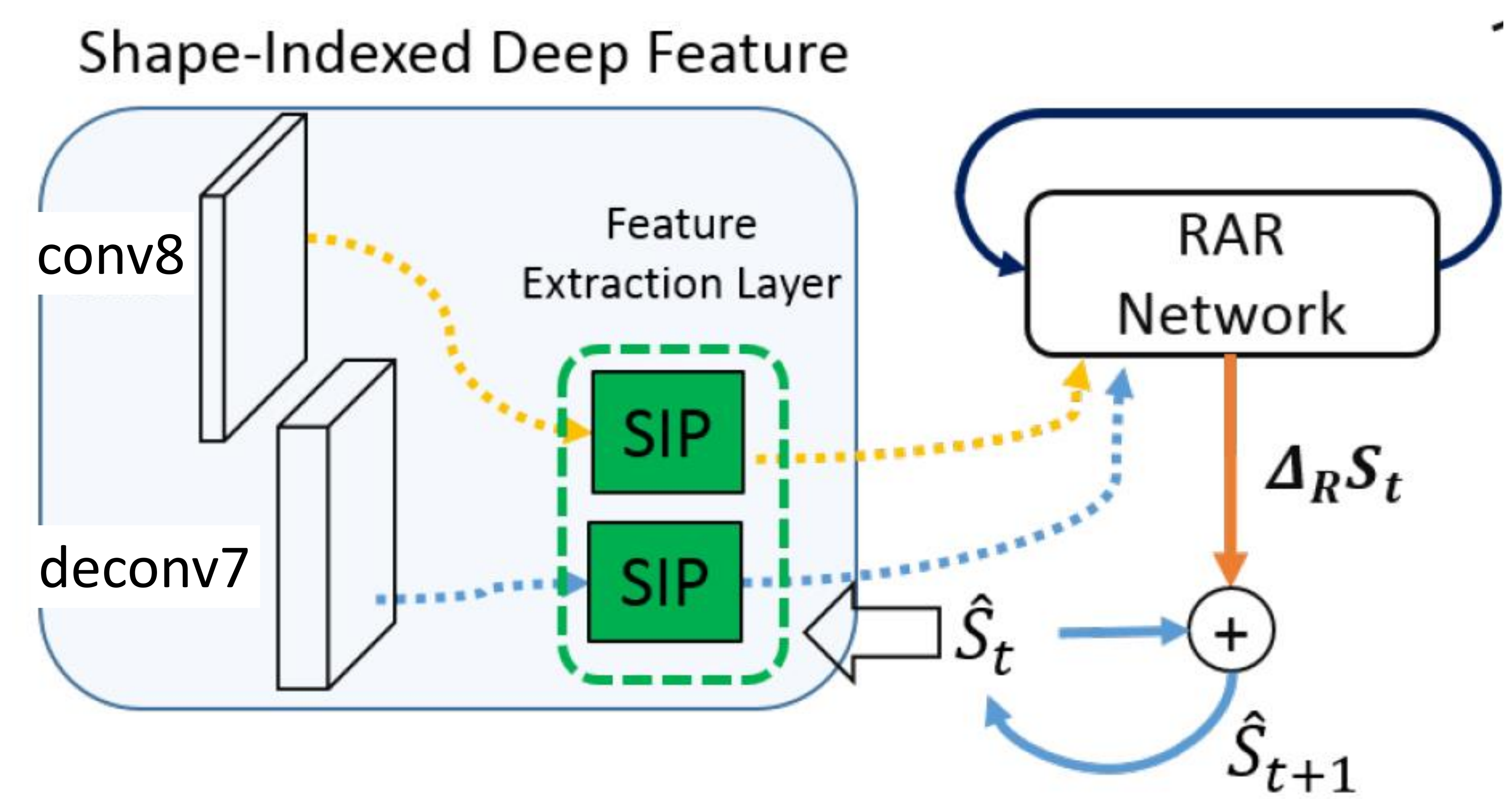
- Feature re-weighting based on distance to attention center:

Weighting Factor

Deep SIP Feature at S_t^2

$$\Phi_r(I_t, \hat{S}_t) = [\boxed{\gamma_t^1}\phi_t^1, \gamma_t^2 \boxed{\phi_t^2}, \ldots, \gamma_t^L \phi_t^L]$$





RAR Networks: Refinement Module

- Feature re-weighting based on distance to attention center:

$$\Phi_r(I_t, \hat{S}_t) = [\gamma_t^1 \phi_t^1, \gamma_t^2 \phi_t^2, \ldots, \gamma_t^L \phi_t^L]$$

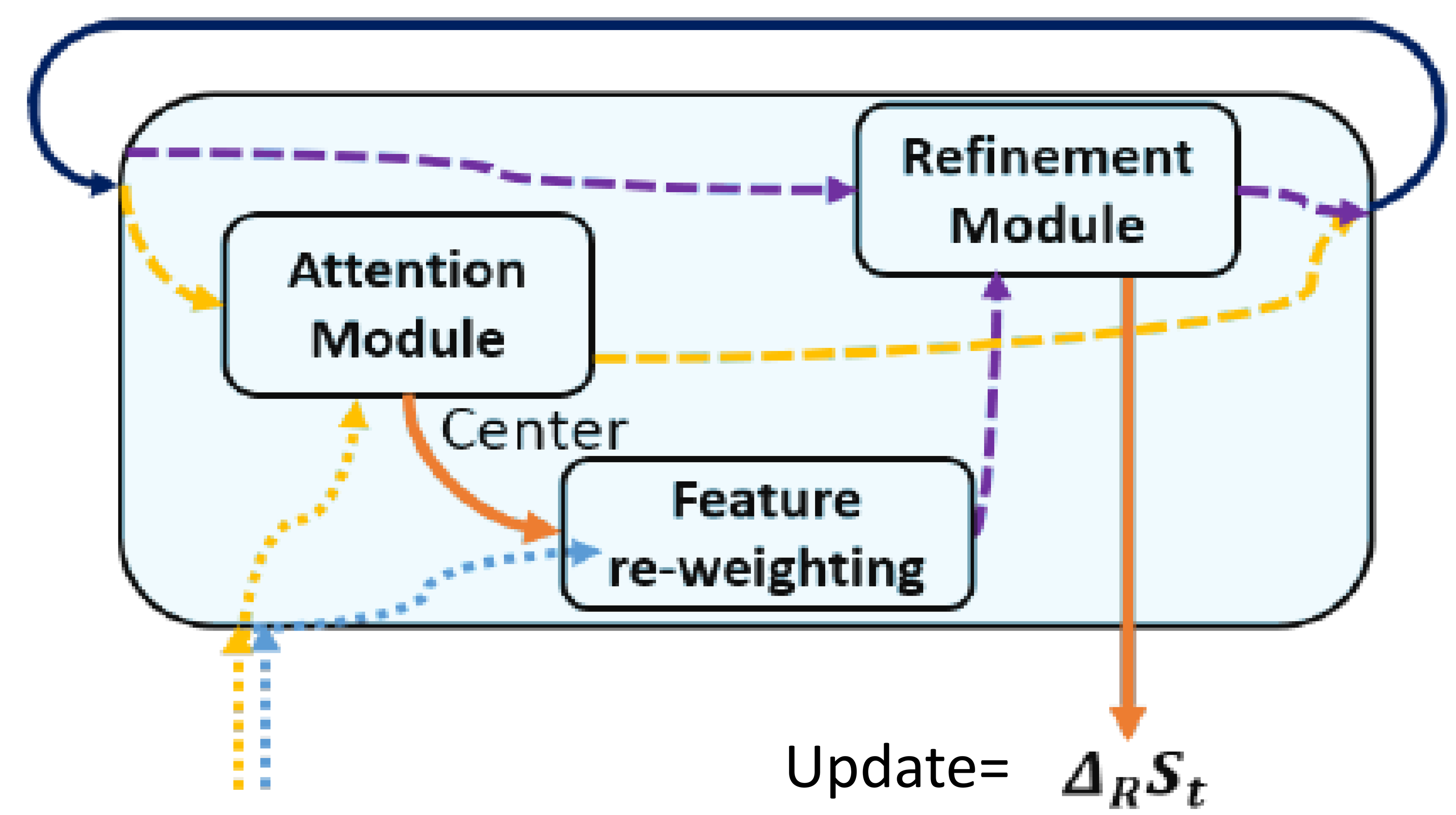
- Refinement Module to get shape update such:

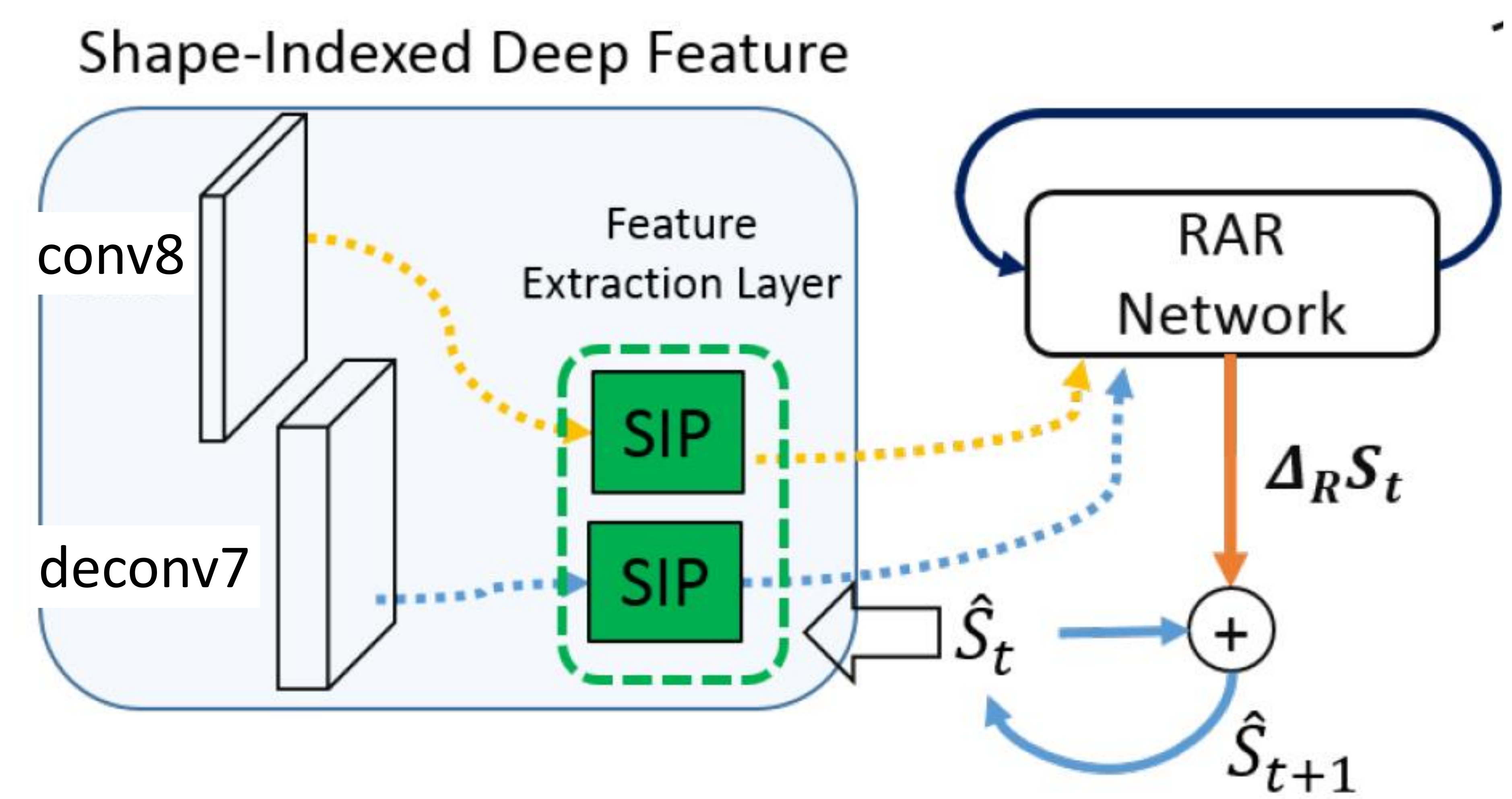
$$\mathcal{L}_R^t = || \, \Gamma_t(\boxed{\Delta_R S_t} - \boxed{\Delta S_t})||_2^2$$

Update

Ground-truth Residue

$\Delta_R S_t$ is the R-LSTM output $\Delta_R S_t = \alpha \Gamma_t \mathrm{R{-}LSTM}(\Phi_r)$





RAR Networks: Refinement Module

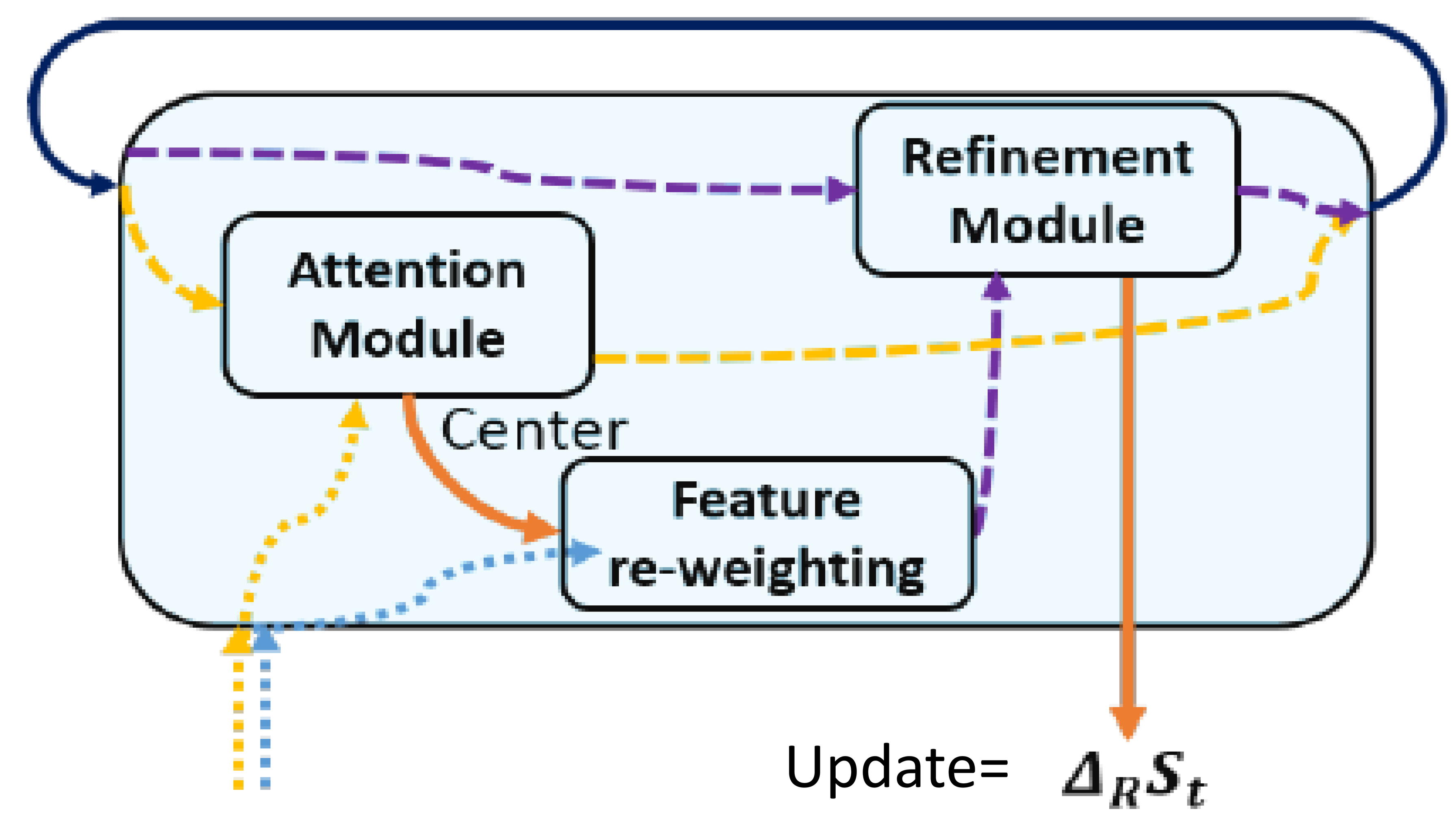
- Feature re-weighting based on distance to attention center:

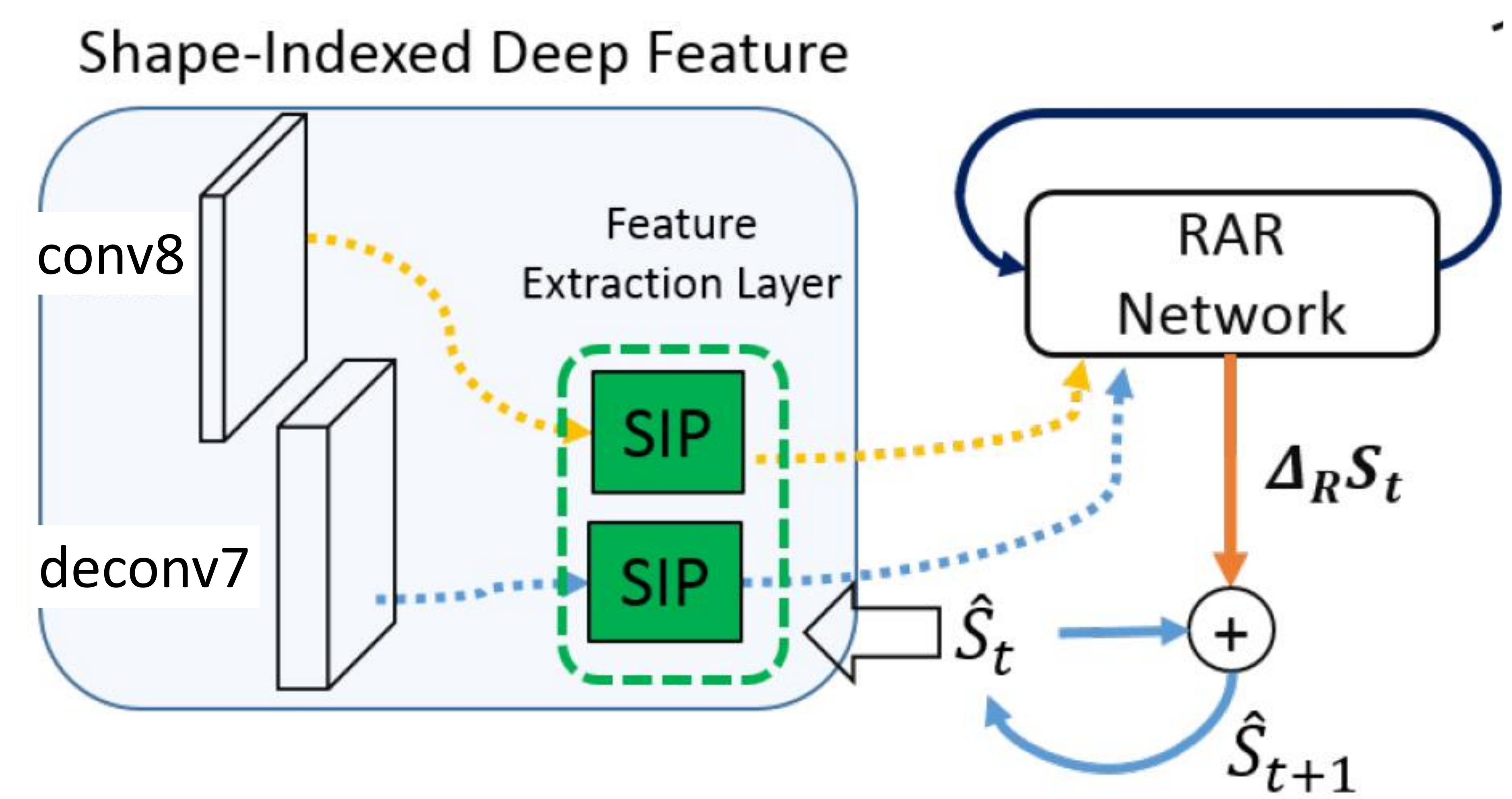
$$\Phi_r(I_t, \hat{S}_t) = [\gamma_t^1 \phi_t^1, \gamma_t^2 \phi_t^2, \ldots, \gamma_t^L \phi_t^L]$$

- Refinement Module to get shape update such:

$$\mathcal{L}_R^t = ||\, \Gamma_t(\Delta_R S_t - \Delta S_t)||_2^2$$

$\Delta_R S_t$ is the R-LSTM output $\Delta_R S_t = \alpha \Gamma_t \mathrm{R{-}LSTM}(\Phi_r)$

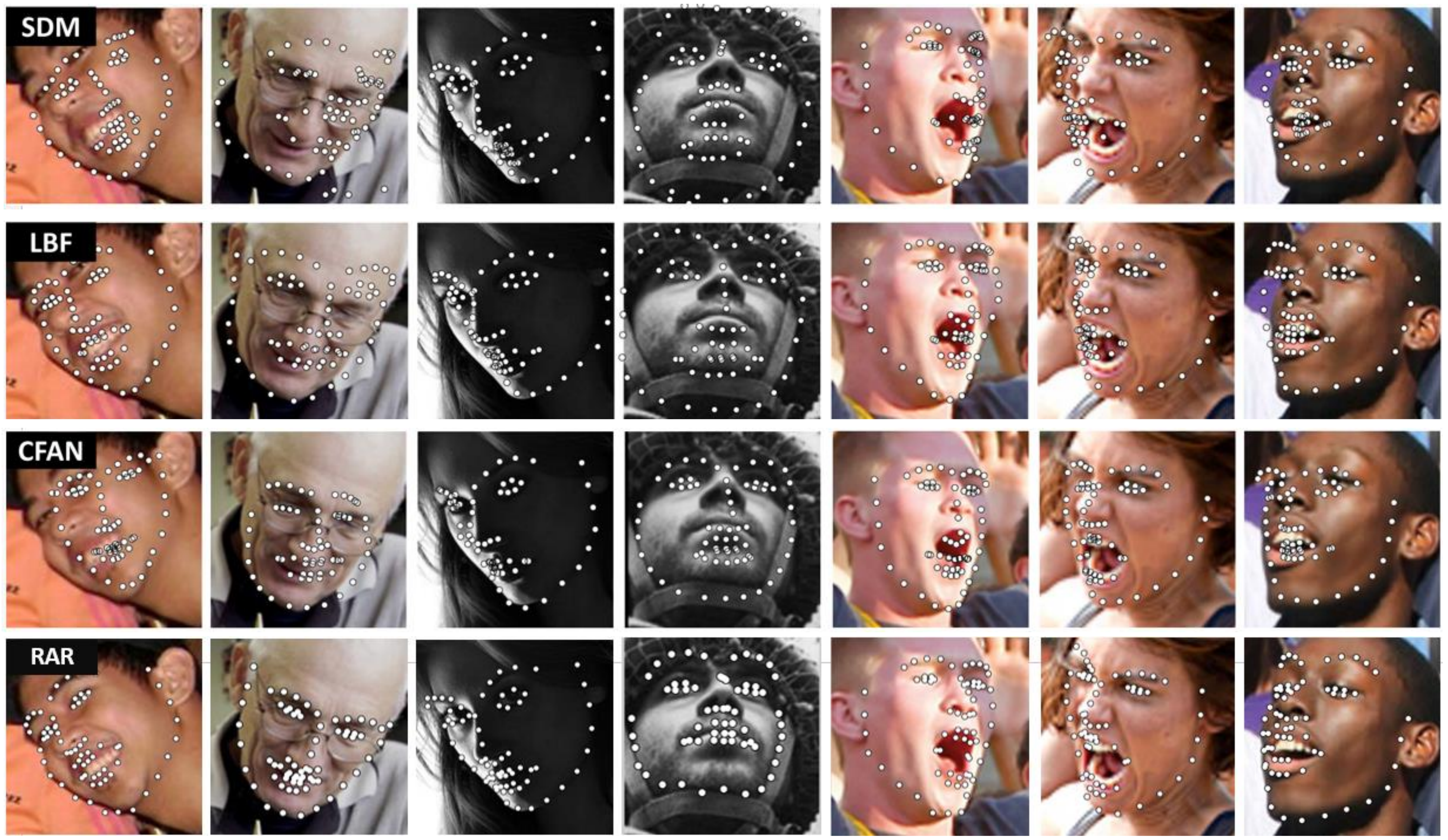




Overall Training Objective of RAR:

$$\sum_{t=1}^{T} \sum_{n=1}^{N} -\underbrace{\eta^{t-1} \mathcal{R}_a(\hat{S}_{t-1,n}, \hat{S}_{t,n})}_{\text{Attention Loss}} + \underbrace{\mathcal{L}_{R,n}^t}_{\text{Regression Loss}}$$

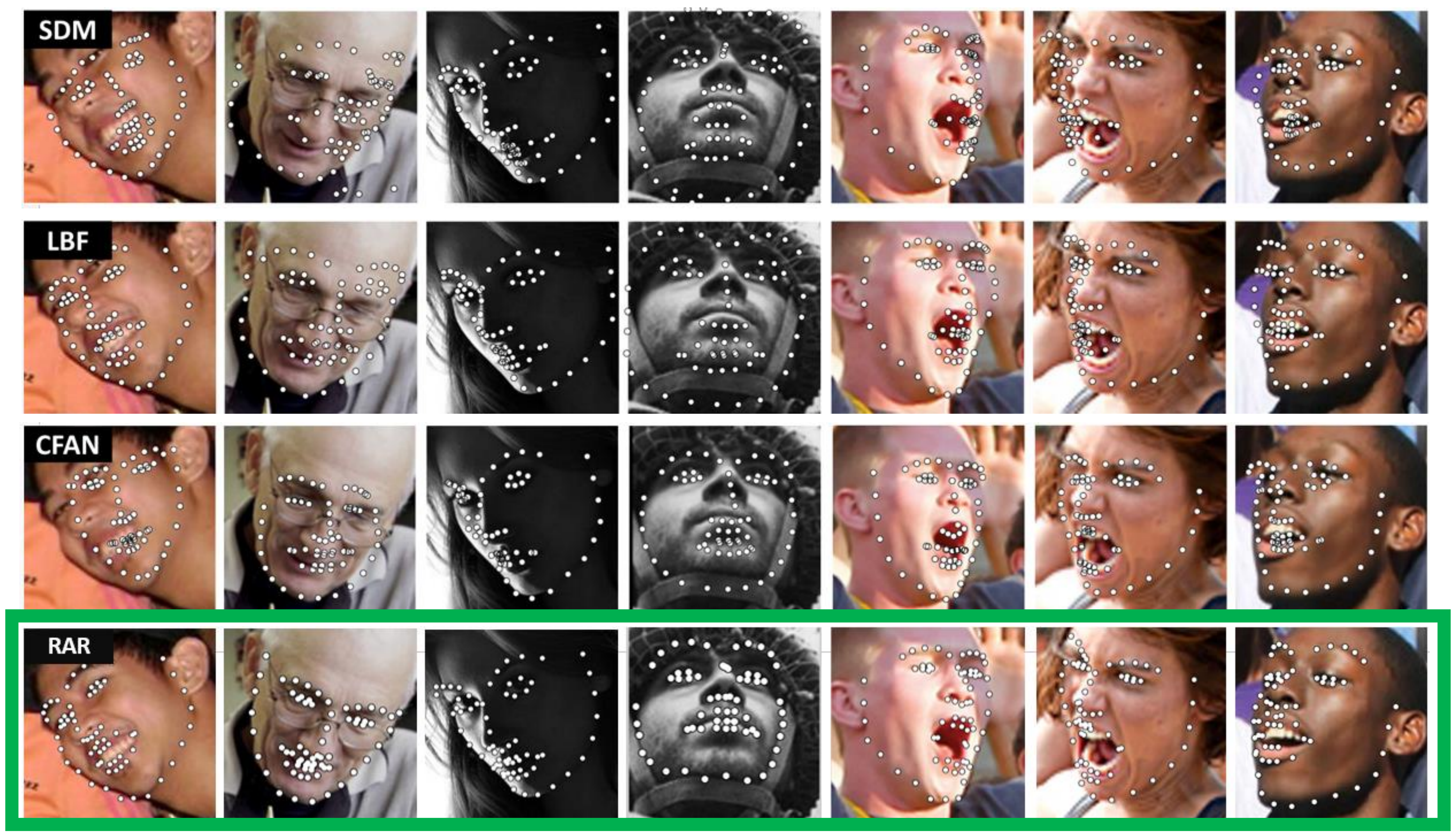
Results on 300W



Methods	300-W Dataset		
	Common	Challenging	Full
Zhu et.al [2012]	8.22	18.33	12.0
RCPR [Burgos,2013]	6.18	17.26	8.35
SDM [Xiong,2013]	5.57	15.40	7.50
LBF [Ren,2014]	4.95	11.98	6.32
LBF Fast [Ren,2014]	5.38	15.50	7.37
CFAN[Zhang, 2014]	5.50	-	-
CFSS [Zhu, 2015]	4.73	9.98	5.76
Ours (RAR)	4.12	8.35	4.94

Zhu: Face detection, pose estimation, and landmark localization in the wild. CVPR 2012
RCPR: Robust face landmark estimation under occlusion. ICCV 2013
SDM: Supervised descent method and its applications to face alignment. CVPR 2013
LBF: Face alignment at 3000 fps via regressing local binary features. ECCV 2014
CFAN: Coarse-to-ne auto-encoder networks(cfan) for real-time face alignment. ECCV2014
CFSS: Face alignment by coarse-to-fine shape searching. CVPR 2015

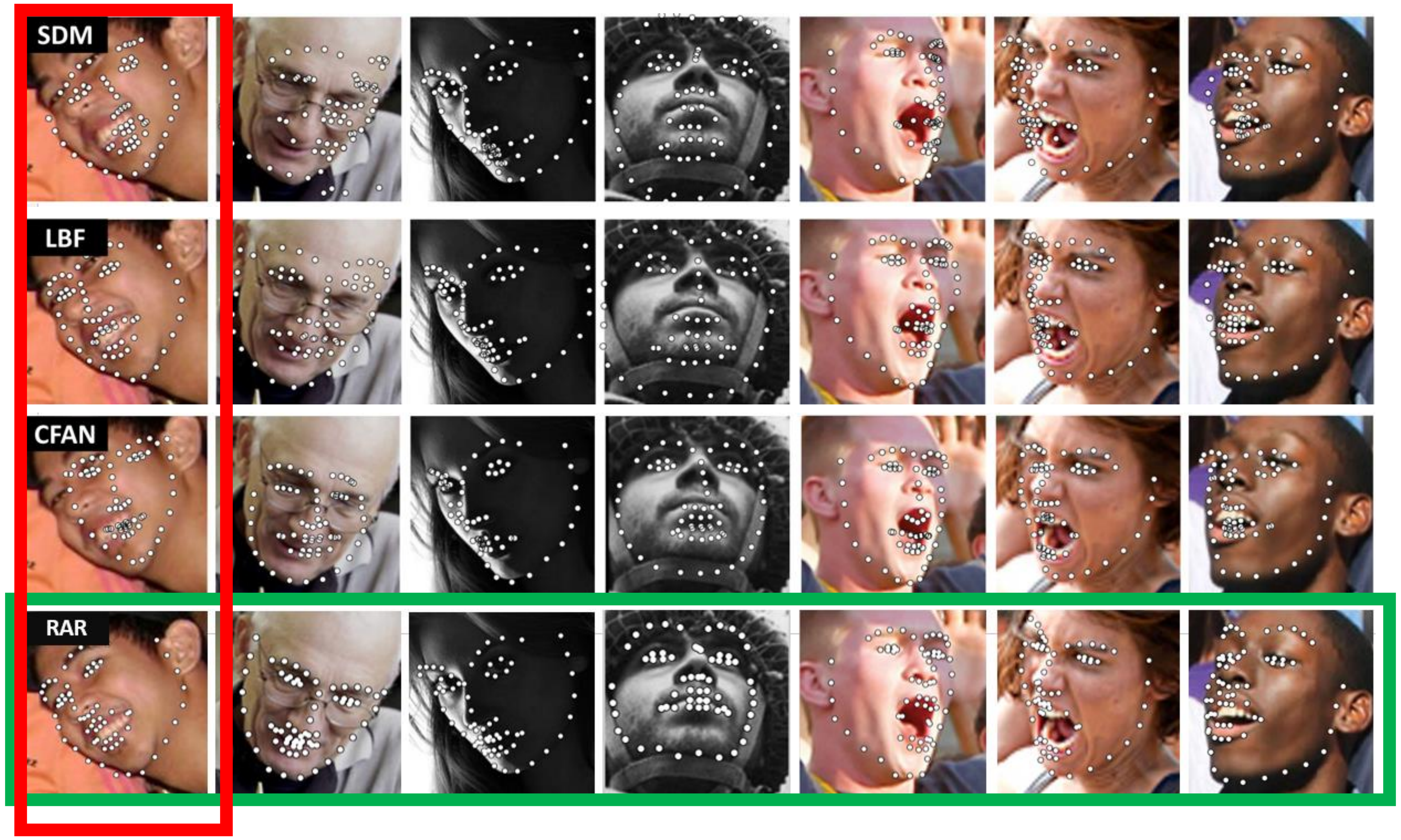
Results on 300W



Methods	300-W Dataset		
	Common	Challenging	Full
Zhu et.al [2012]	8.22	18.33	12.0
RCPR [Burgos,2013]	6.18	17.26	8.35
SDM [Xiong,2013]	5.57	15.40	7.50
LBF [Ren,2014]	4.95	11.98	6.32
LBF Fast [Ren,2014]	5.38	15.50	7.37
CFAN[Zhang, 2014]	5.50	-	-
CFSS [Zhu, 2015]	4.73	9.98	5.76
Ours (RAR)	4.12	8.35	4.94

Zhu: Face detection, pose estimation, and landmark localization in the wild. CVPR 2012
RCPR: Robust face landmark estimation under occlusion. ICCV 2013
SDM: Supervised descent method and its applications to face alignment. CVPR 2013
LBF: Face alignment at 3000 fps via regressing local binary features. ECCV 2014
CFAN: Coarse-to-ne auto-encoder networks(cfan) for real-time face alignment. ECCV2014
CFSS: Face alignment by coarse-to-fine shape searching. CVPR 2015

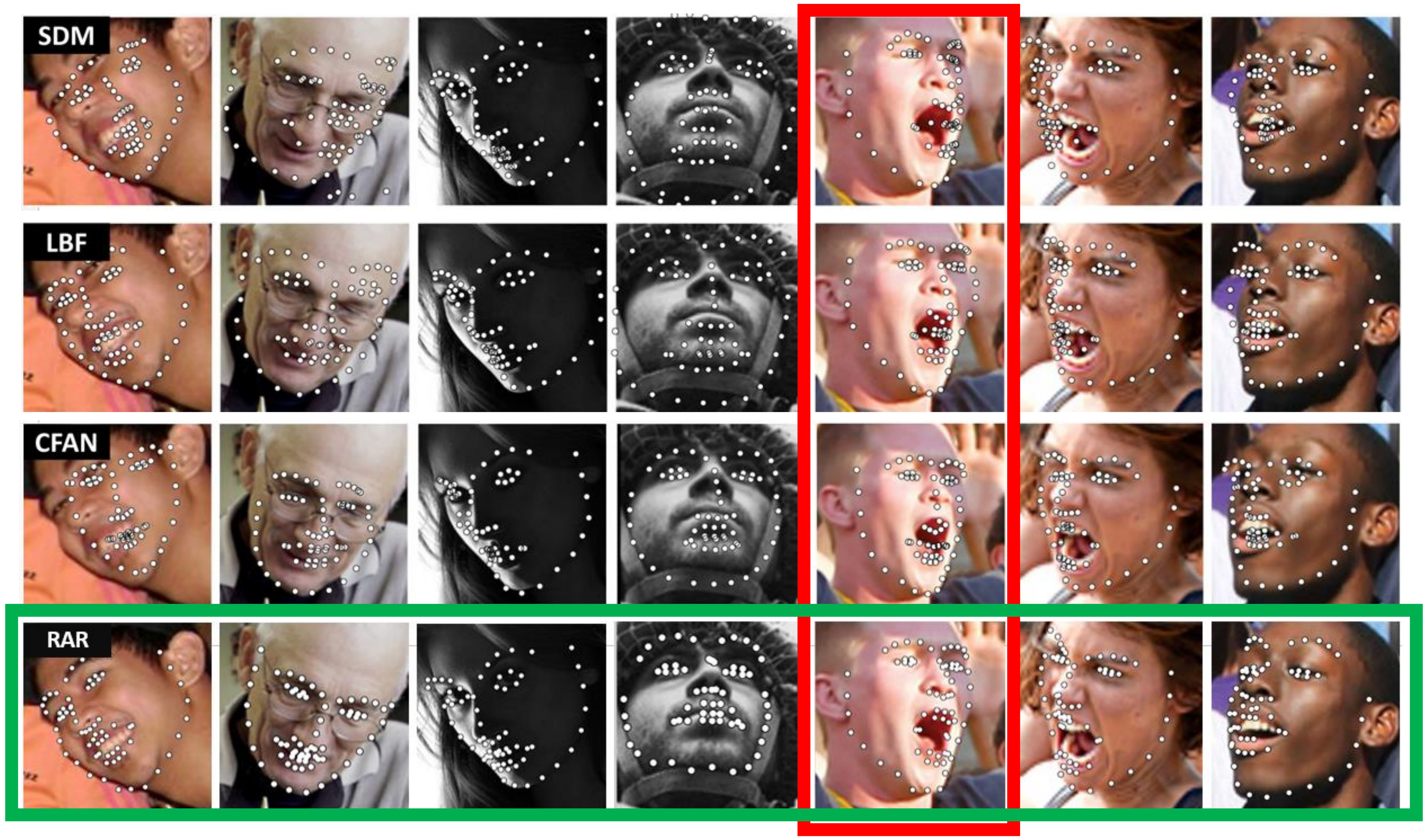
Results on 300W



Methods	300-W Dataset		
	Common	Challenging	Full
Zhu et.al [2012]	8.22	18.33	12.0
RCPR [Burgos,2013]	6.18	17.26	8.35
SDM [Xiong,2013]	5.57	15.40	7.50
LBF [Ren,2014]	4.95	11.98	6.32
LBF Fast [Ren,2014]	5.38	15.50	7.37
CFAN[Zhang, 2014]	5.50	-	-
CFSS [Zhu, 2015]	4.73	9.98	5.76
Ours (RAR)	4.12	8.35	4.94

Zhu: Face detection, pose estimation, and landmark localization in the wild. CVPR 2012
RCPR: Robust face landmark estimation under occlusion. ICCV 2013
SDM: Supervised descent method and its applications to face alignment. CVPR 2013
LBF: Face alignment at 3000 fps via regressing local binary features. ECCV 2014
CFAN: Coarse-to-ne auto-encoder networks(cfan) for real-time face alignment. ECCV2014
CFSS: Face alignment by coarse-to-fine shape searching. CVPR 2015

Results on 300W



Methods	300-W Dataset		
	Common	Challenging	Full
Zhu et.al [2012]	8.22	18.33	12.0
RCPR [Burgos,2013]	6.18	17.26	8.35
SDM [Xiong,2013]	5.57	15.40	7.50
LBF [Ren,2014]	4.95	11.98	6.32
LBF Fast [Ren,2014]	5.38	15.50	7.37
CFAN[Zhang, 2014]	5.50	-	-
CFSS [Zhu, 2015]	4.73	9.98	5.76
Ours (RAR)	4.12	8.35	4.94

Zhu: Face detection, pose estimation, and landmark localization in the wild. CVPR 2012
RCPR: Robust face landmark estimation under occlusion. ICCV 2013
SDM: Supervised descent method and its applications to face alignment. CVPR 2013
LBF: Face alignment at 3000 fps via regressing local binary features. ECCV 2014
CFAN: Coarse-to-ne auto-encoder networks(cfan) for real-time face alignment. ECCV2014
CFSS: Face alignment by coarse-to-fine shape searching. CVPR 2015

Results on COFW and AFLW

COFW Dataset

Methods	Normalized ME	Failure Rate
RCPR	8.50	20.00%
HPM	7.46	13.24%
RPP	7.52	16.20%
TCDCN	8.05	-
RAR	**6.03**	**4.14%**

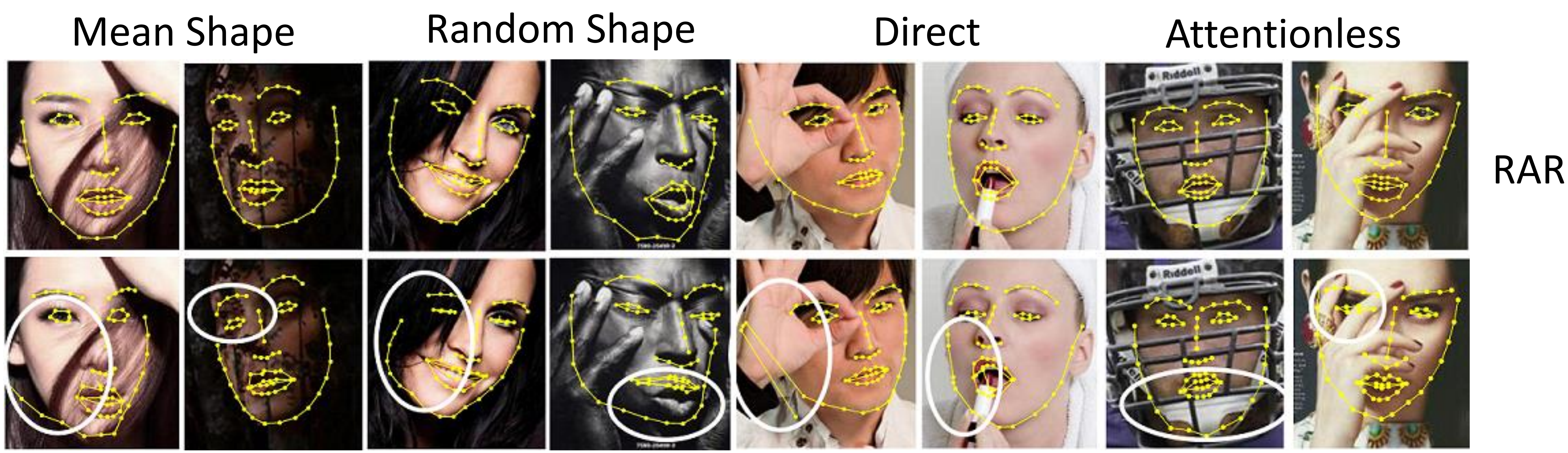
ALFW Dataset

Methods	Normalized ME
RCPR	11.6
SDM	8.50
CFAN	10.95
TCDCN	7.60
RAR	**7.23**

RCPR: Robust face landmark estimation under occlusion. ICCV 2013
HPM: Hierarchical part model. CVPR 2014.
RPP: Regional Predictive Power. TIP 2015.
TCDCN: Task constraint deep convolutional nets. PAMI 2015.
SDM: Supervised descent method and its applications to face alignment. CVPR 2013
CFAN: Coarse-to-ne auto-encoder networks(cfan) for real-time face alignment. ECCV2014
CFSS: Face alignment by coarse-to-fine shape searching. CVPR 2015

Comparison Studies

Dataset	Conv8	Mean Shape	Random Shape	Direct	Robust Initialization
300-W	6.24	5.26	5.22	6.66	**4.94**
COFW	30.14	6.24	6.12	11.52	**6.03**
AFLW	8.14	7.36	7.42	8.15	**7.23**

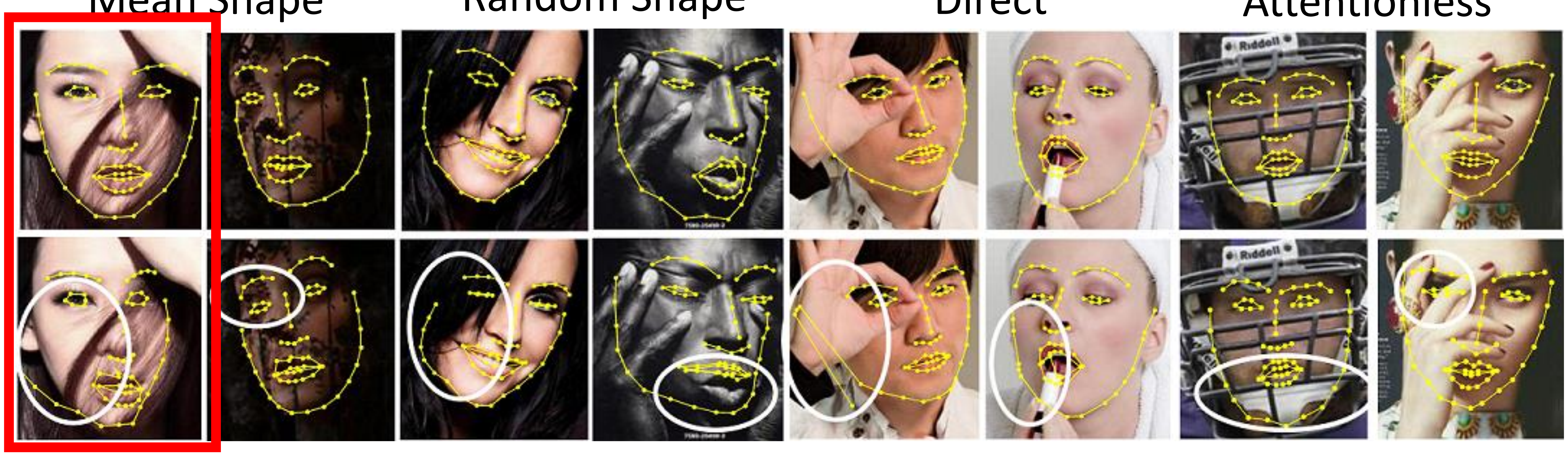


Conv8: Prediction from conv8
Direct: RAR trained with initial shape as conv8
Mean Shape: RAR trained with mean shape as initial shape
Random Shape: RAR trained with random shape as initial shape
Robust: RAR trained with the proposed robust initialization

Comparison Studies

Dataset	Conv8	Mean Shape	Random Shape	Direct	Robust Initialization
300-W	6.24	5.26	5.22	6.66	**4.94**
COFW	30.14	6.24	6.12	11.52	**6.03**
AFLW	8.14	7.36	7.42	8.15	**7.23**

Mean Shape Random Shape Direct Attentionless

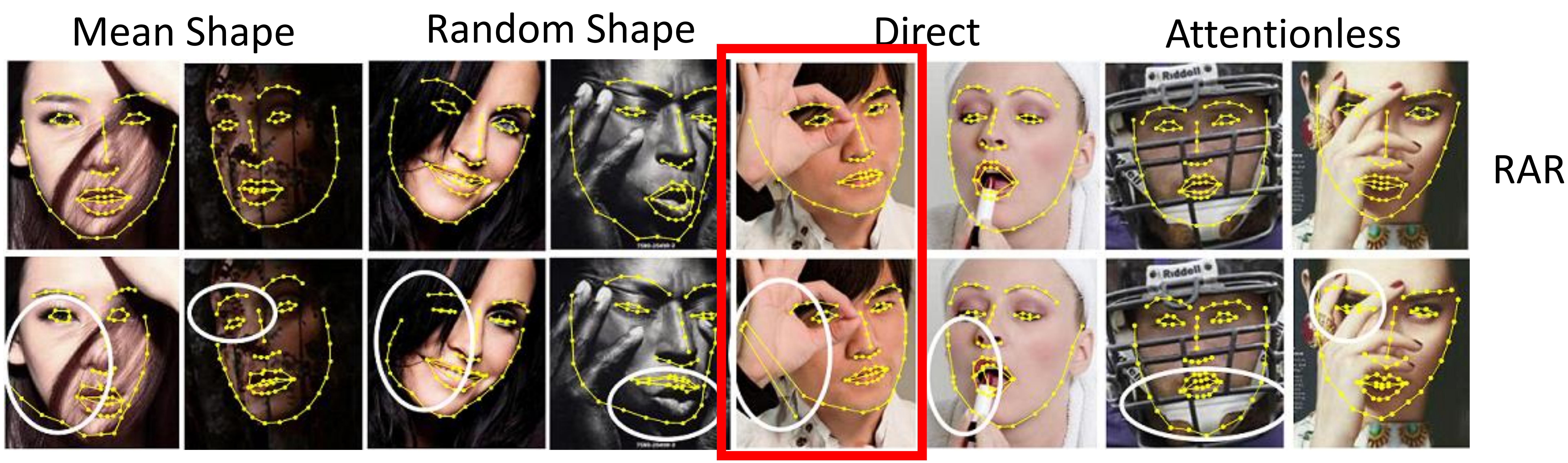


RAR

Conv8: Prediction from conv8
Direct: RAR trained with initial shape as conv8
Mean Shape: RAR trained with mean shape as initial shape
Random Shape: RAR trained with random shape as initial shape
Robust: RAR trained with the proposed robust initialization

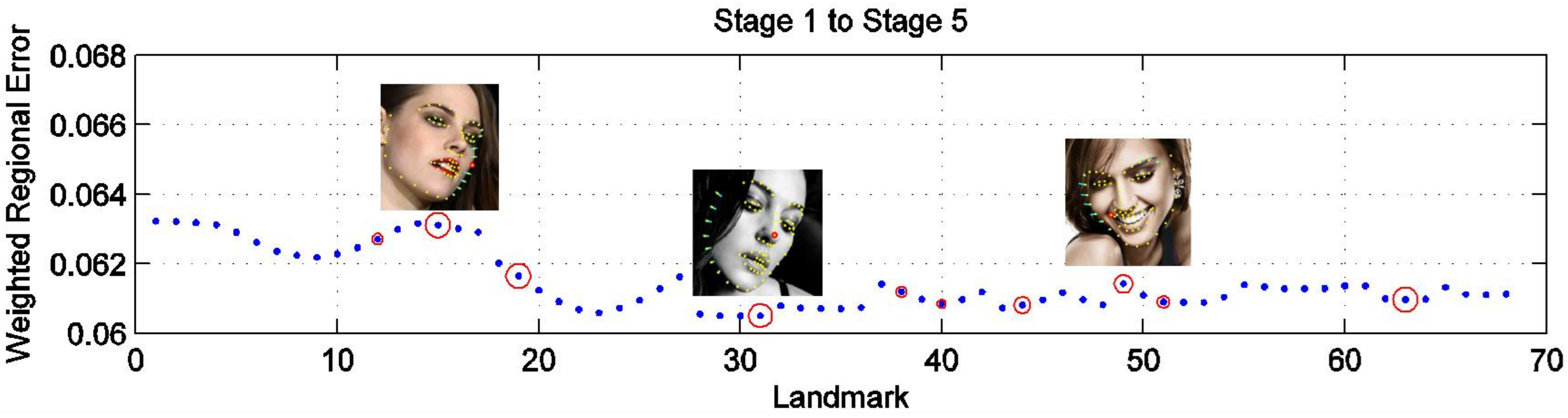
Comparison Studies

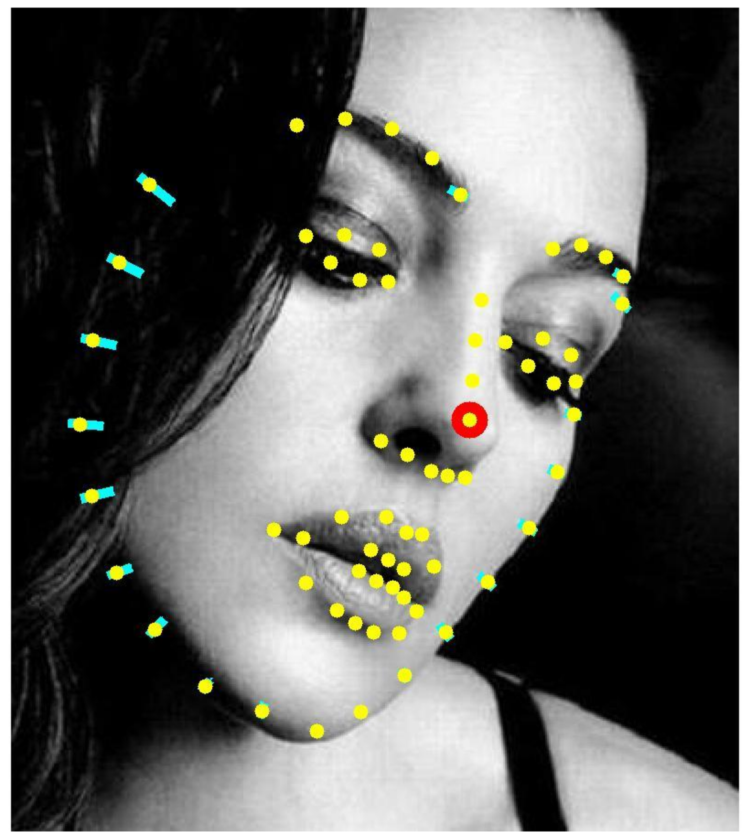
Dataset	Conv8	Mean Shape	Random Shape	Direct	Robust Initialization
300-W	6.24	5.26	5.22	6.66	**4.94**
COFW	30.14	6.24	6.12	11.52	**6.03**
AFLW	8.14	7.36	7.42	8.15	**7.23**



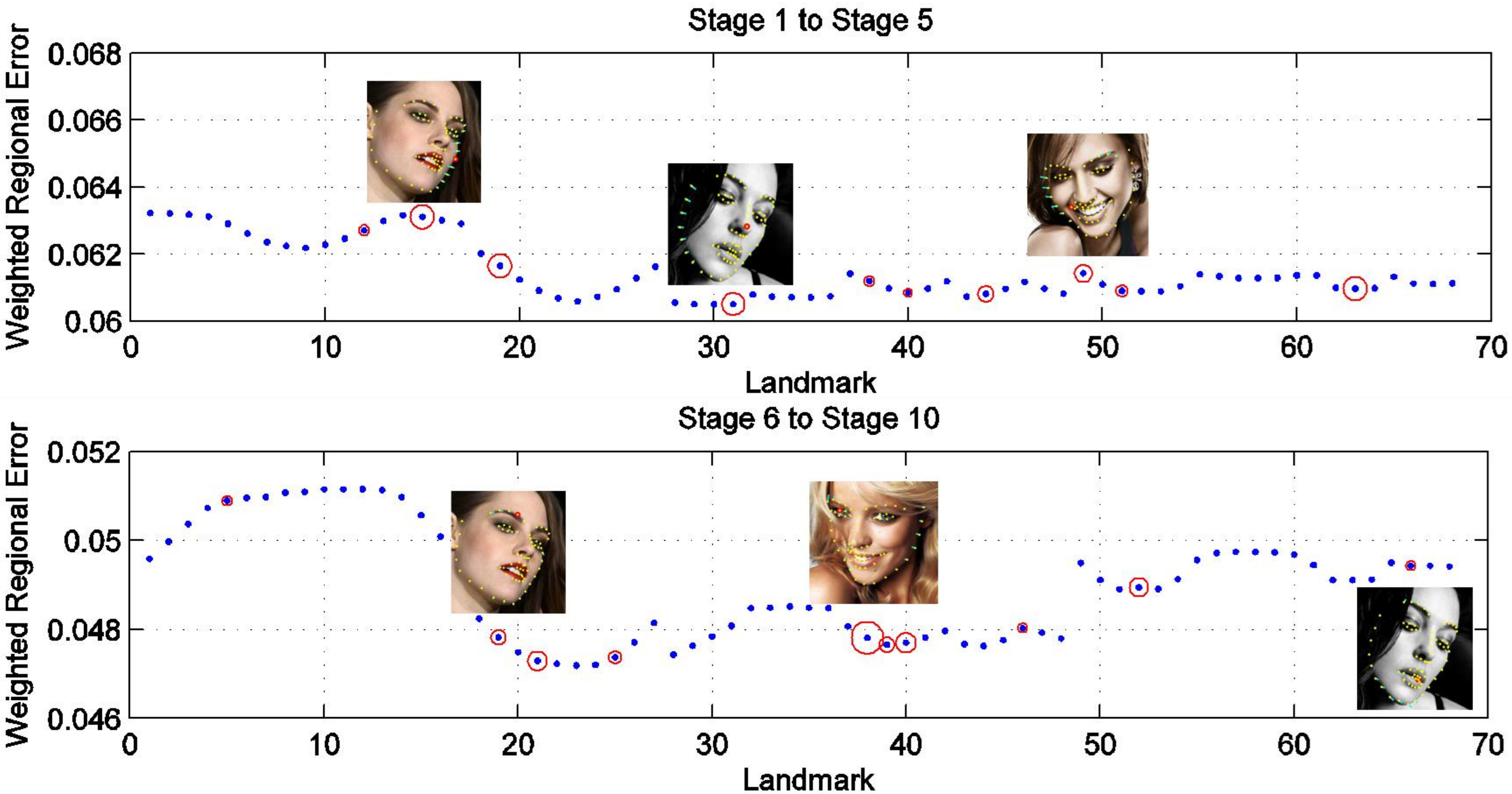
Conv8: Prediction from conv8
Direct: RAR trained with initial shape as conv8
Mean Shape: RAR trained with mean shape as initial shape
Random Shape: RAR trained with random shape as initial shape
Robust: RAR trained with the proposed robust initialization

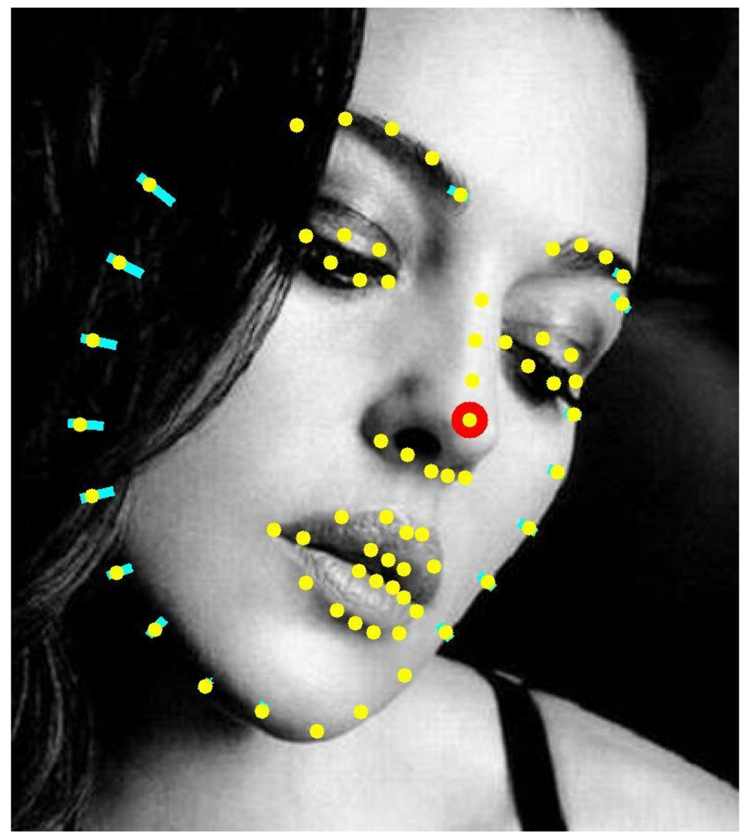
Attention Center Selection Frequencies

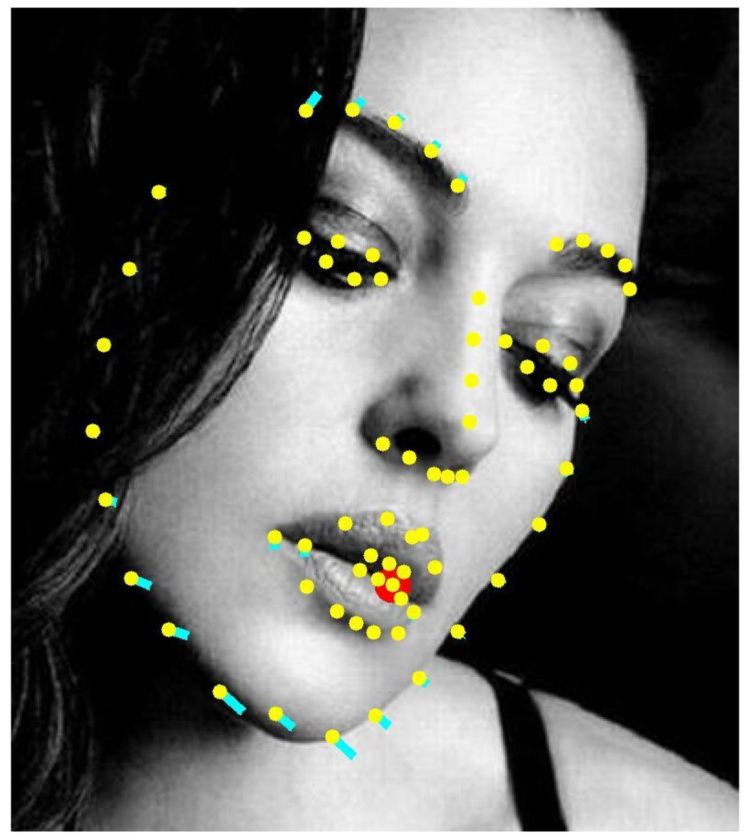




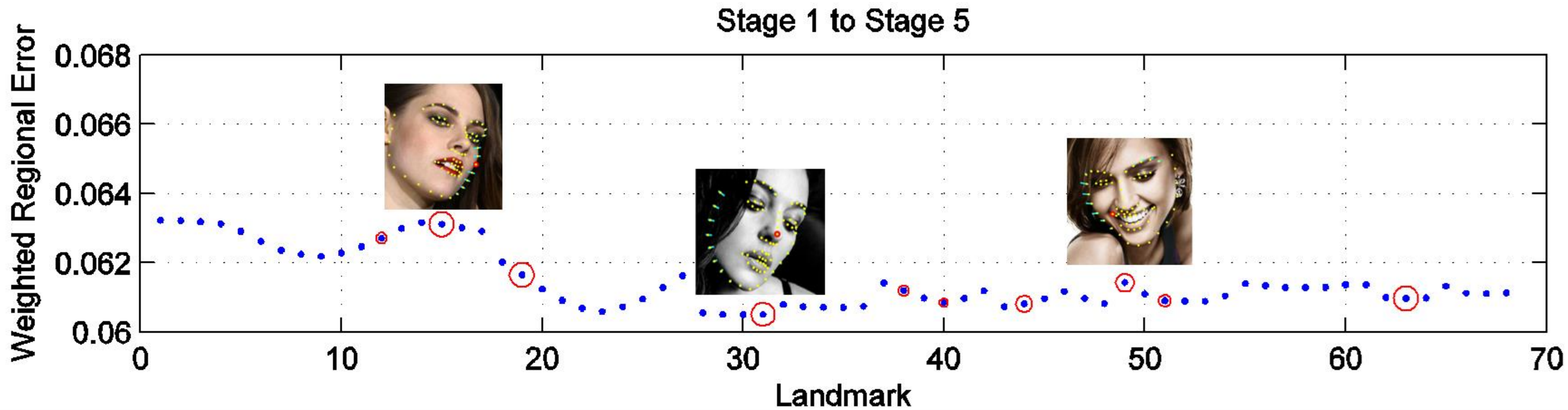
Attention Center Selection Frequencies

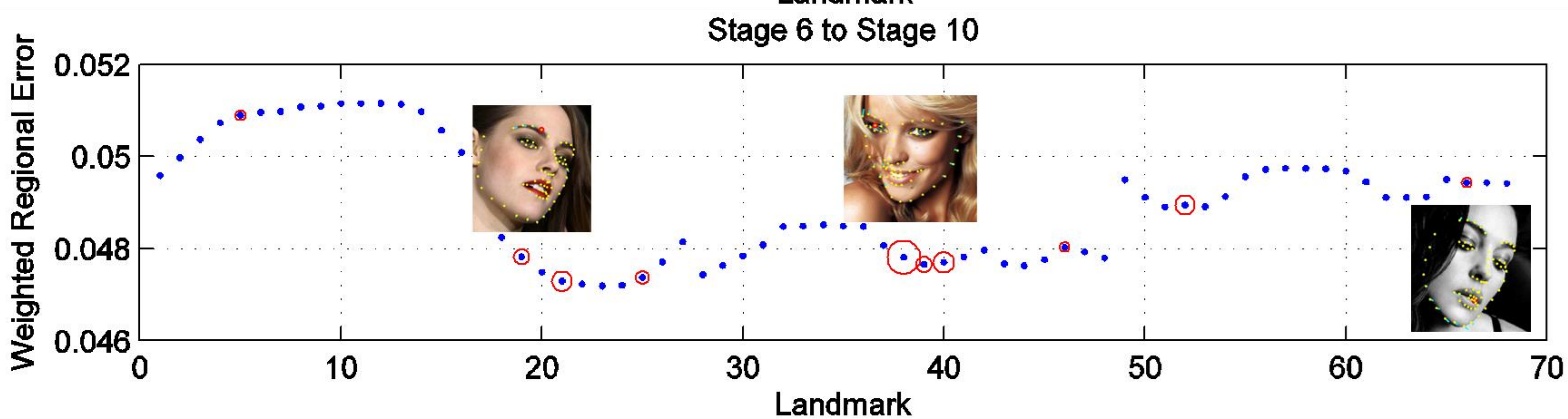


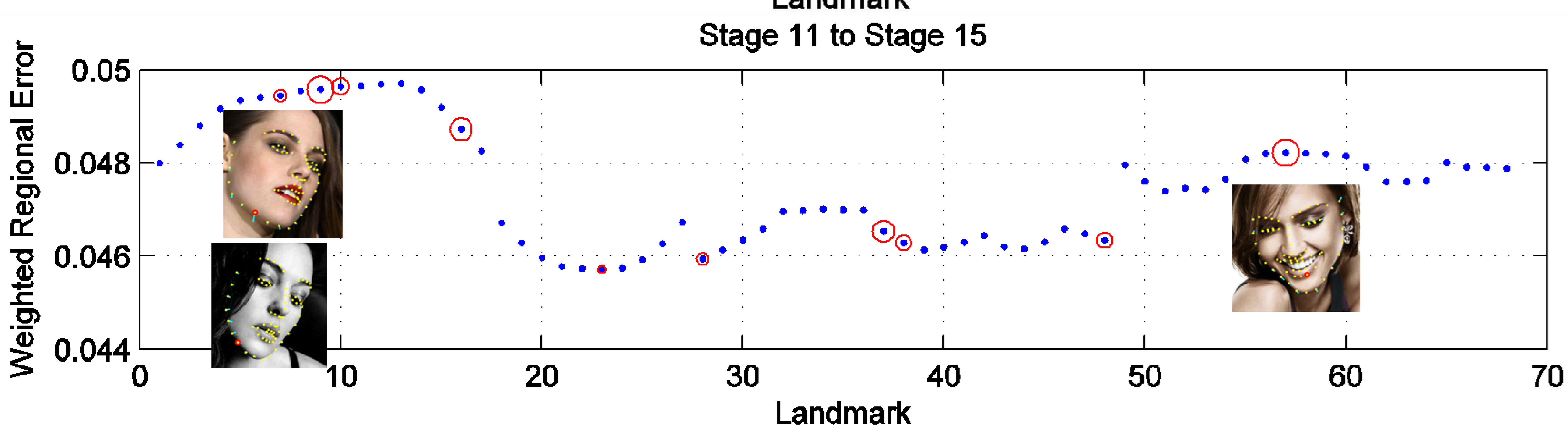


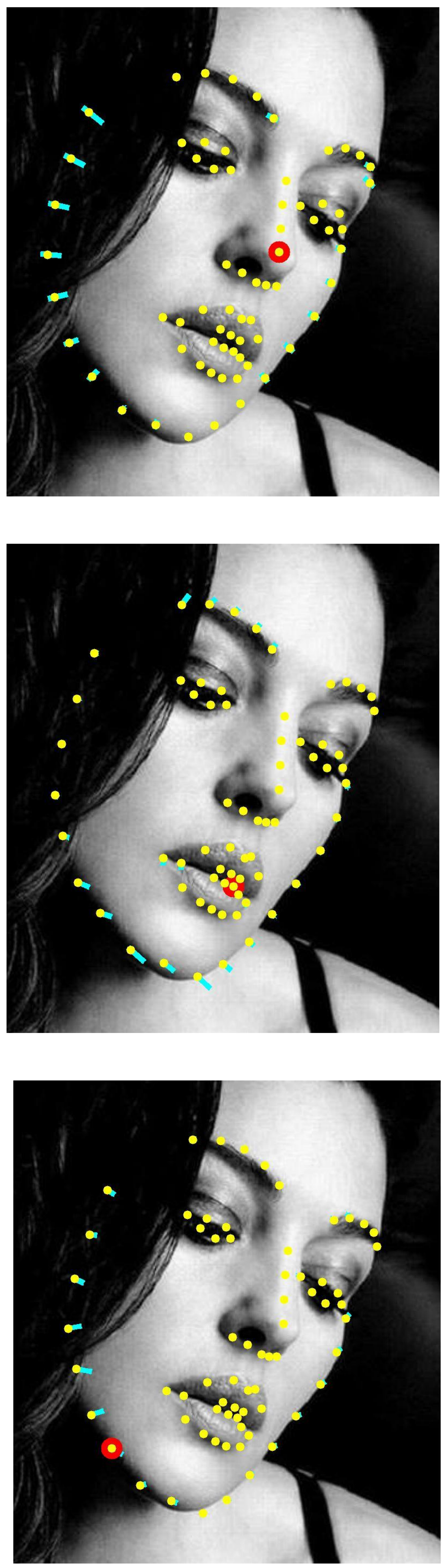


Attention Center Selection Frequencies





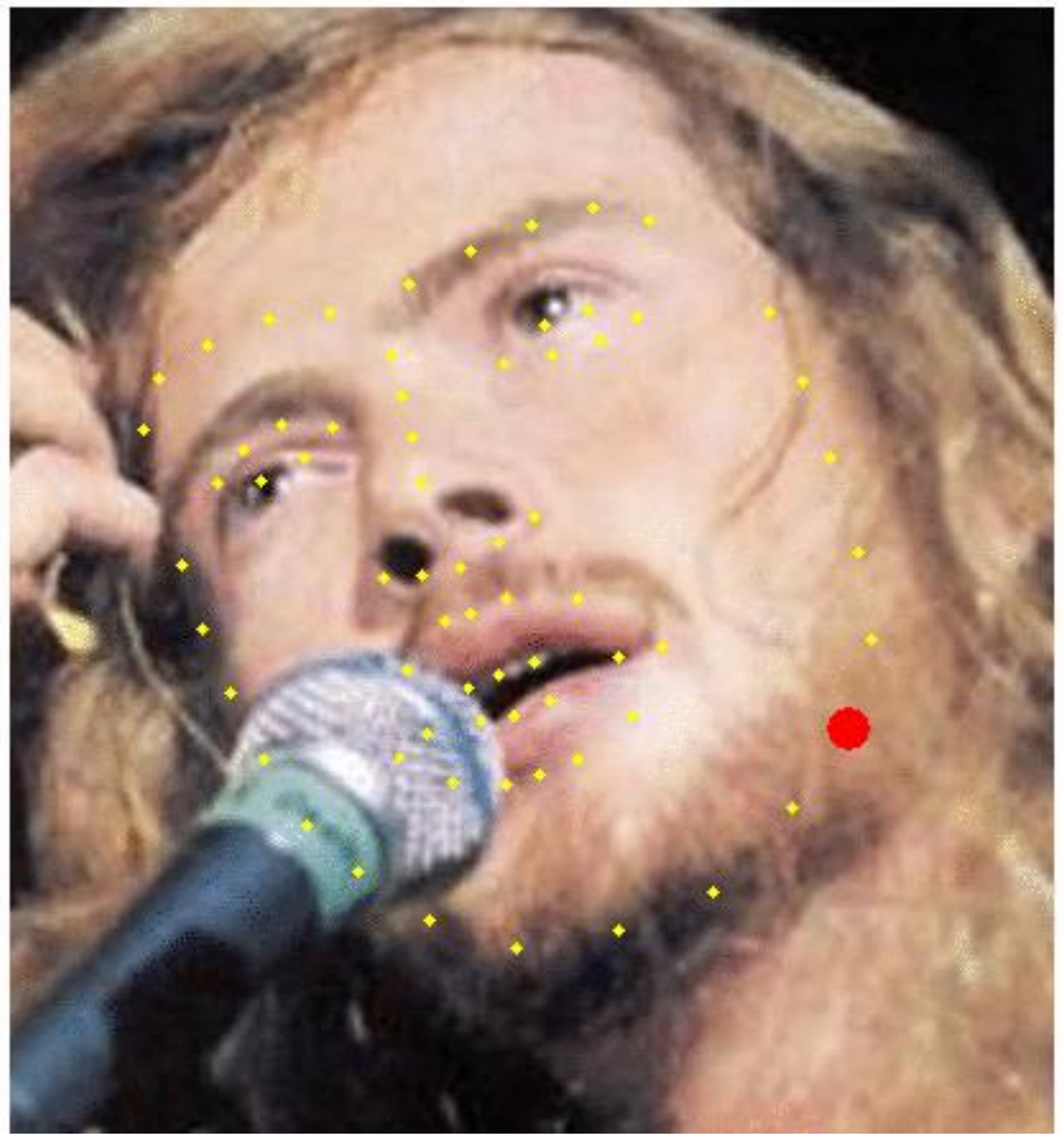


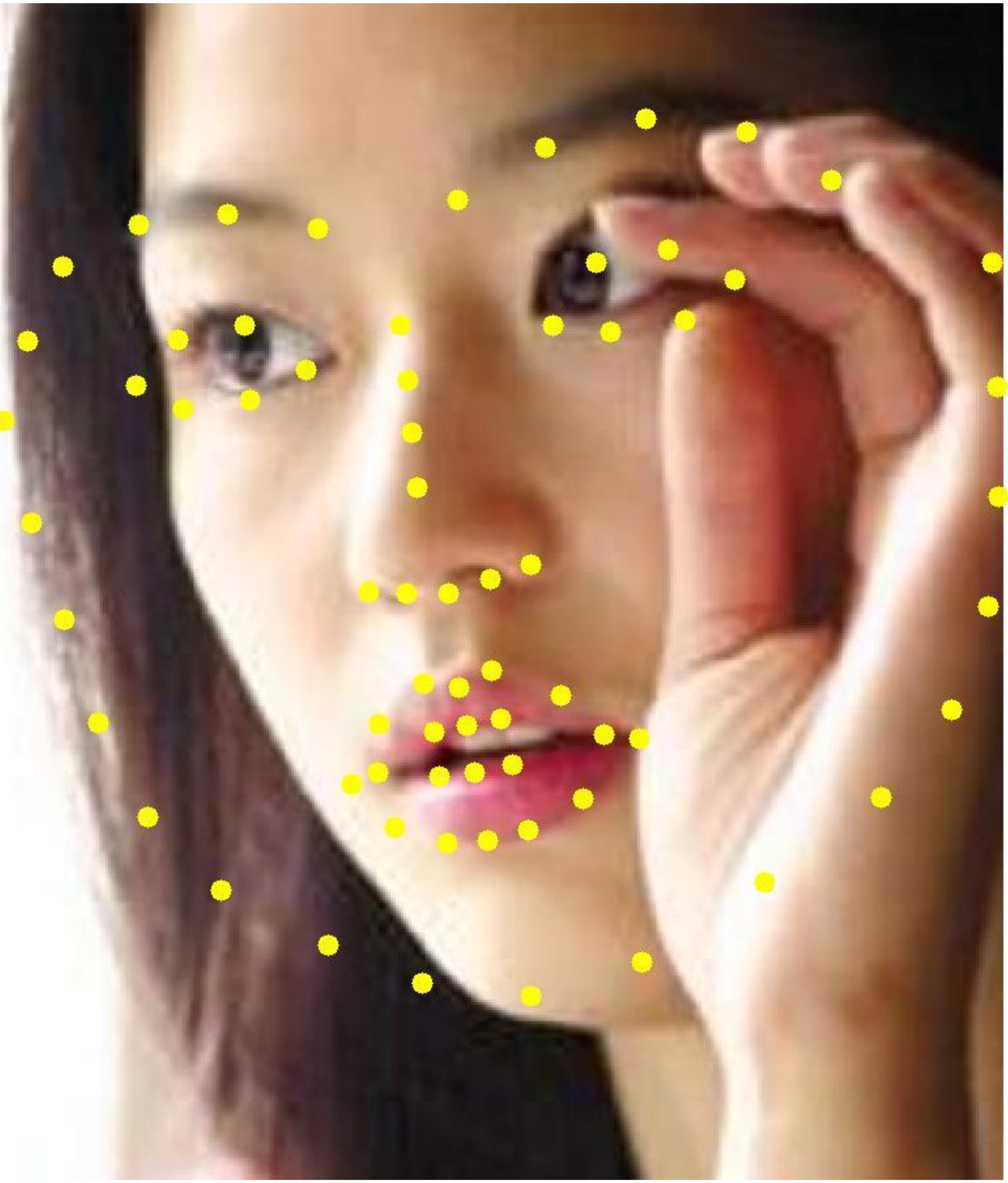


Sample Attentive Refinement

Iter 01

Iter 01

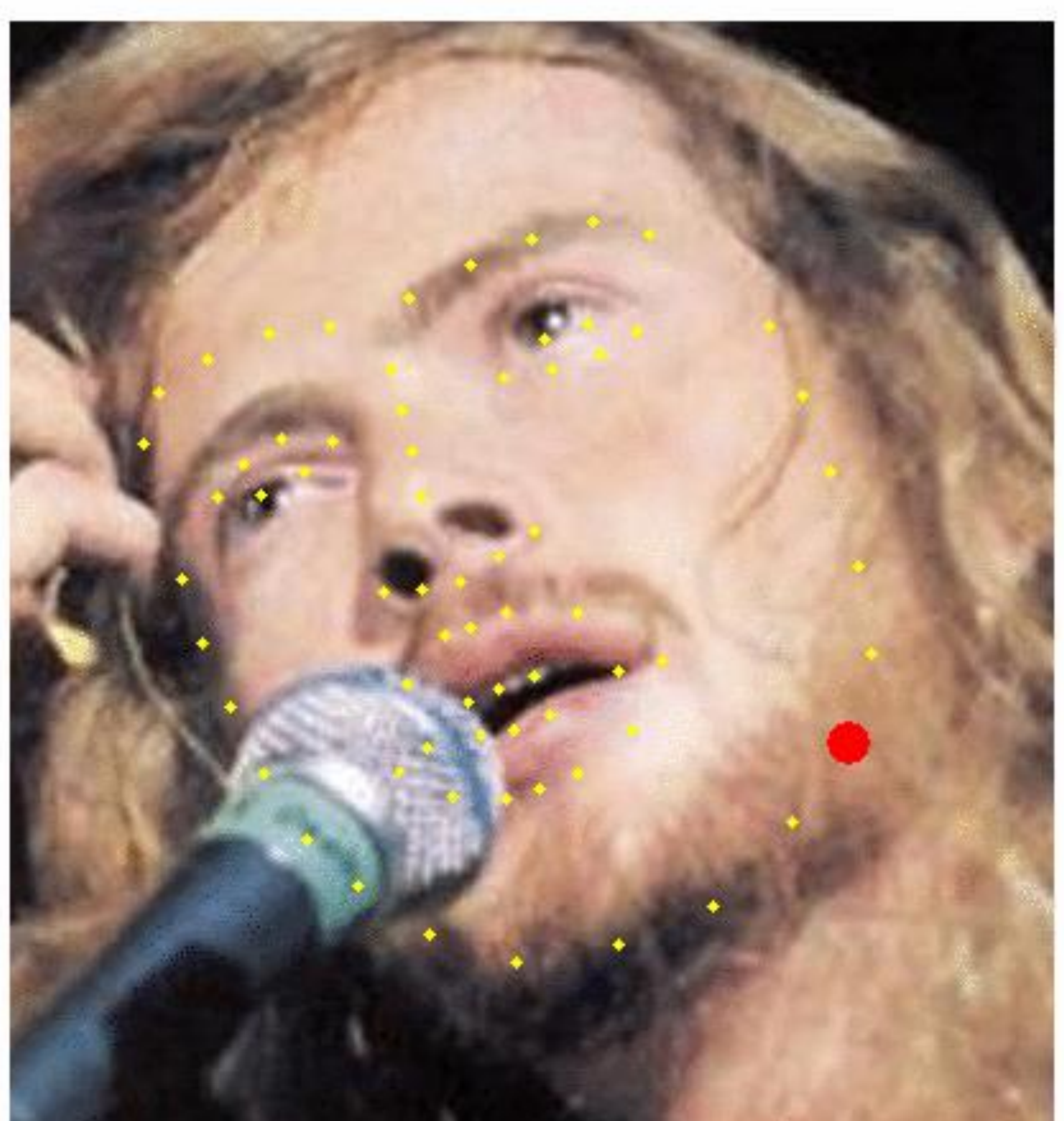


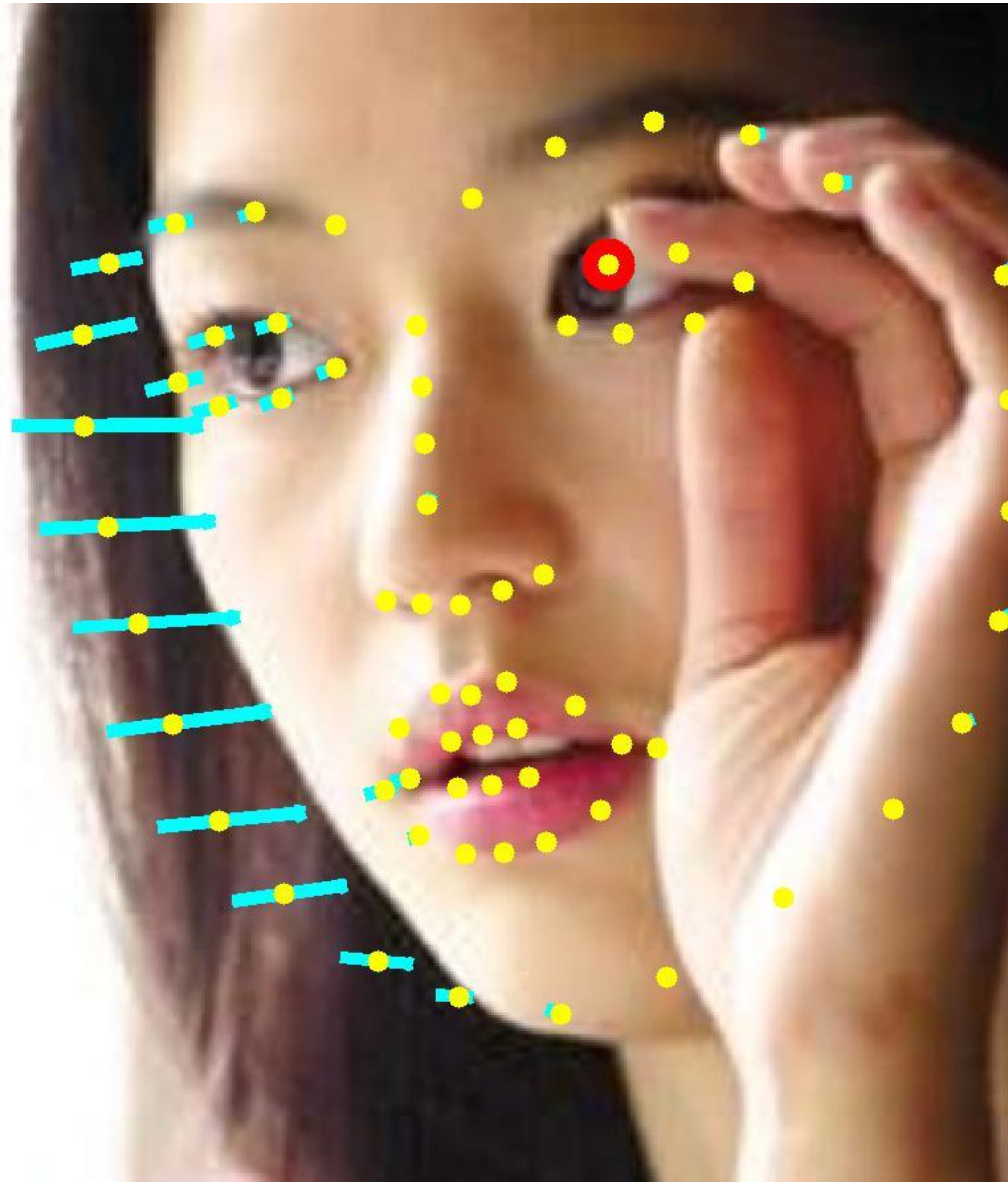


Sample Attentive Refinement

Iter 01

Iter 01



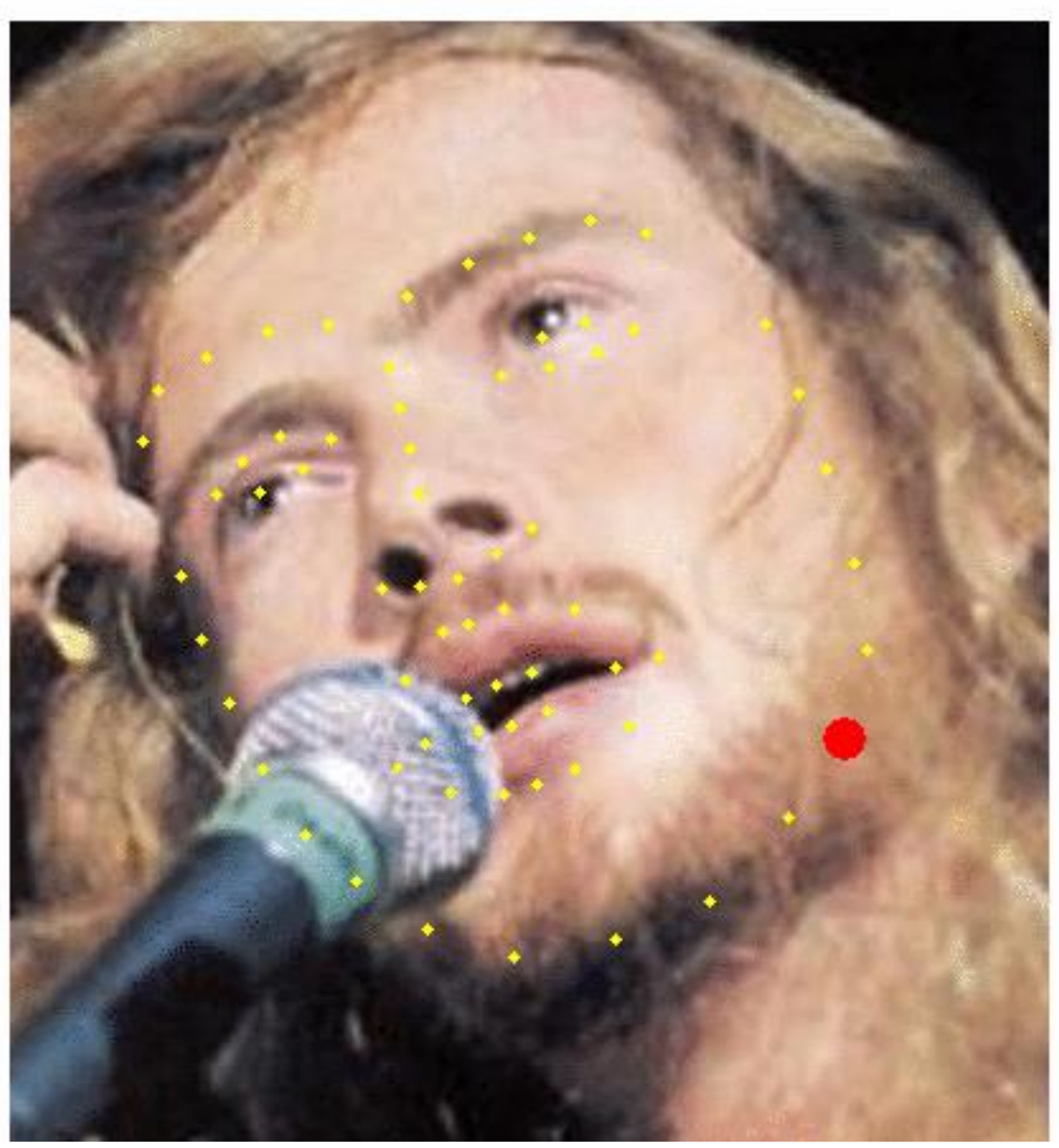


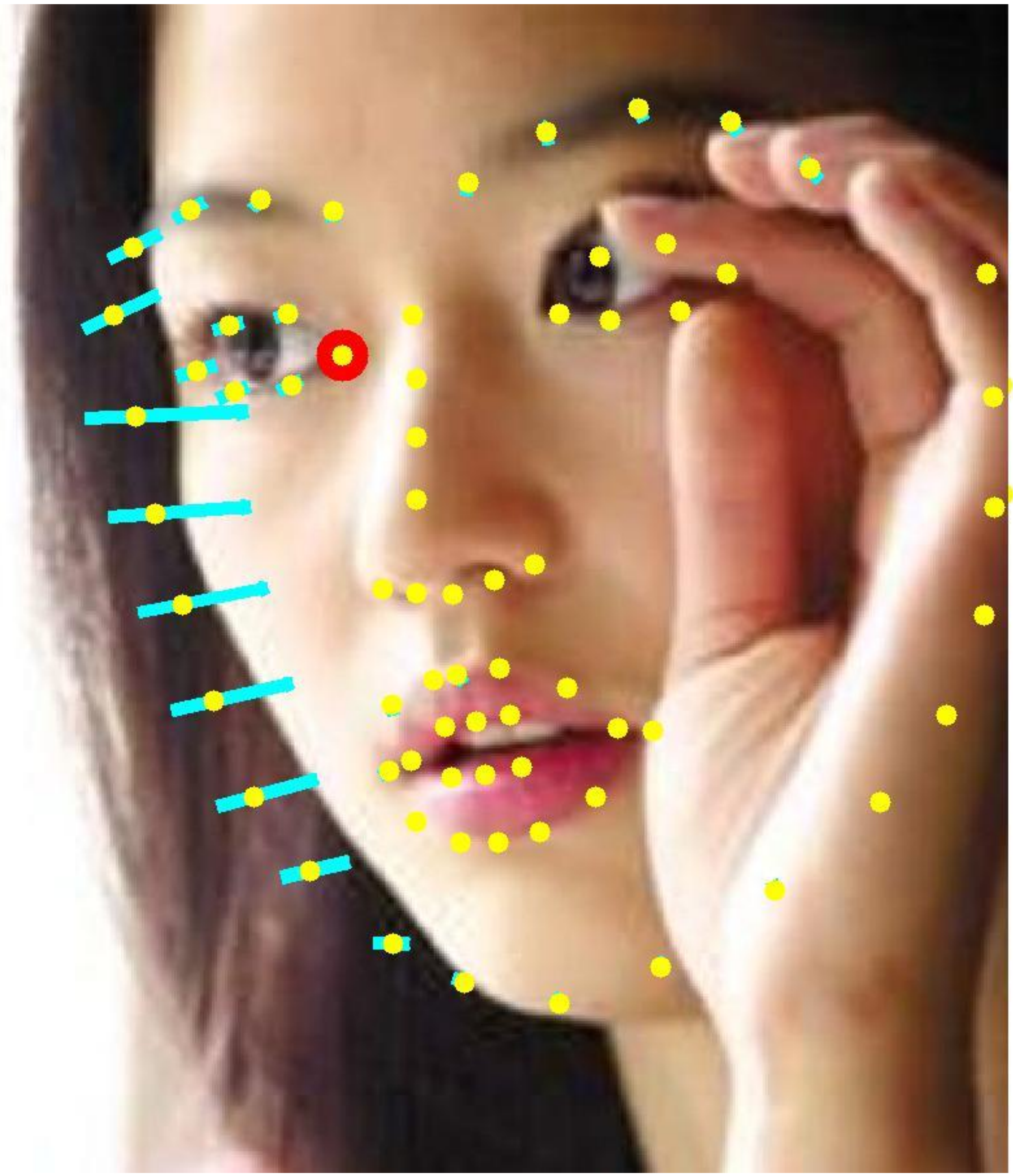
1

Sample Attentive Refinement

Iter 01

Iter 01

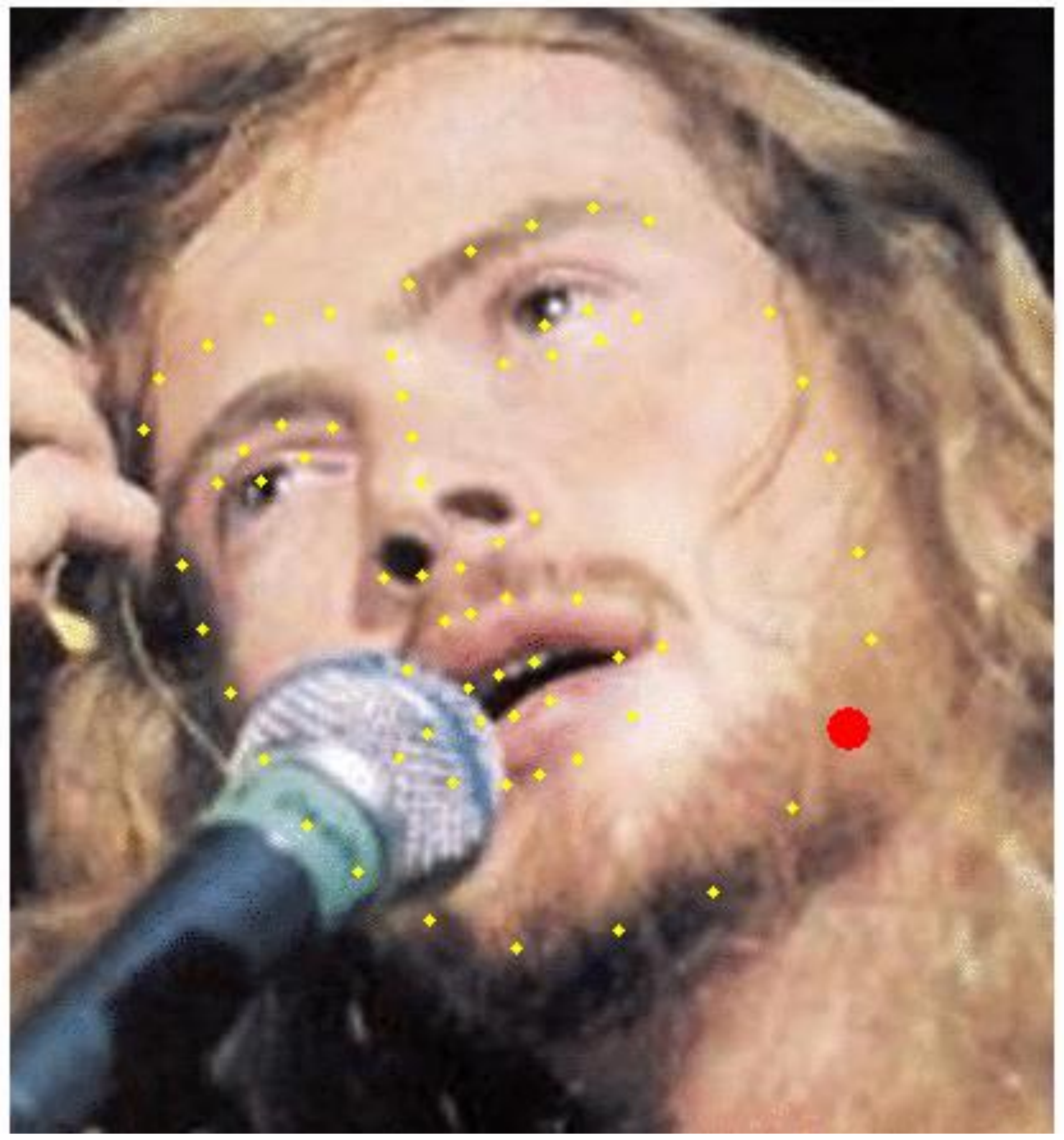


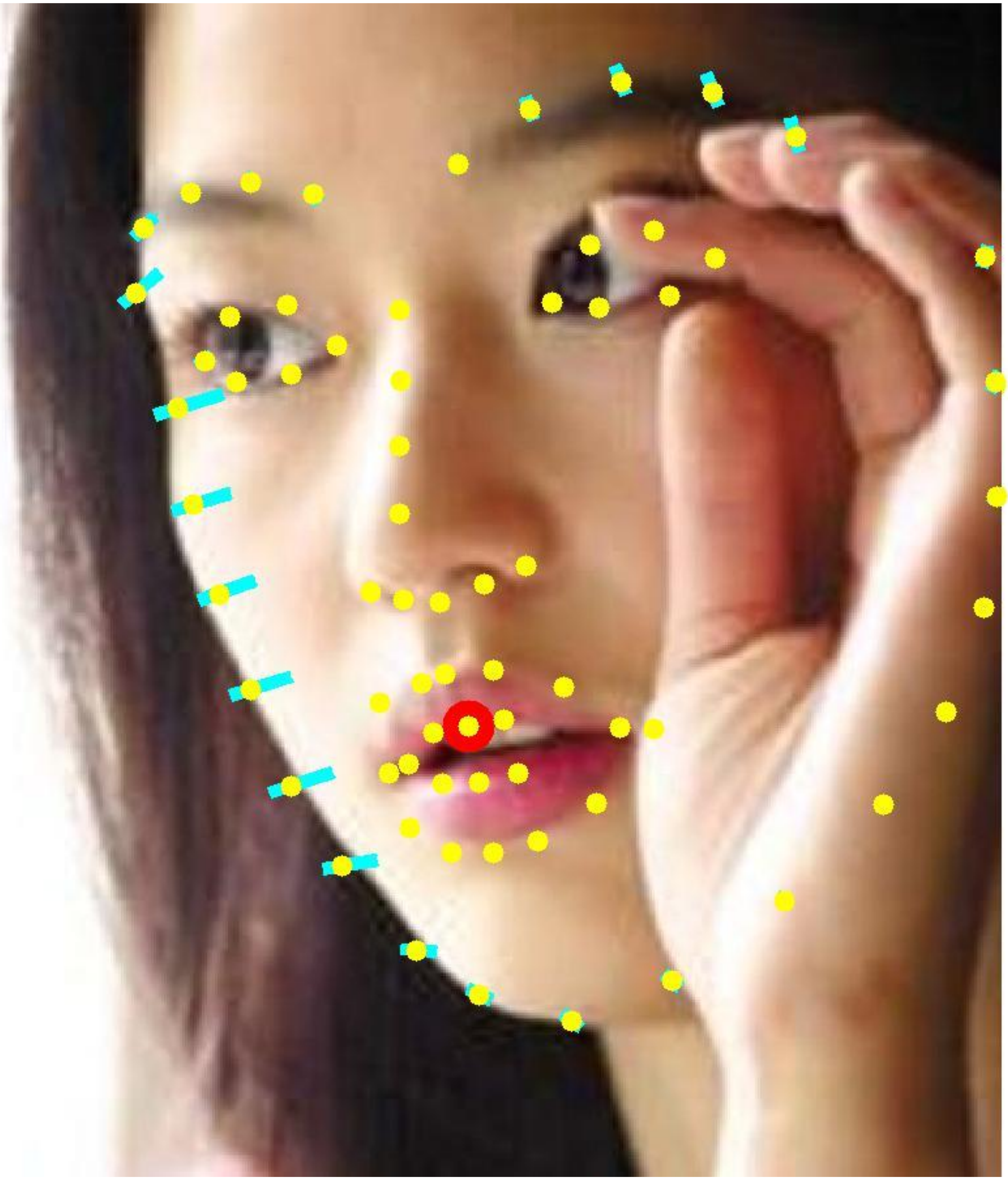


Sample Attentive Refinement

Iter 01

Iter 01



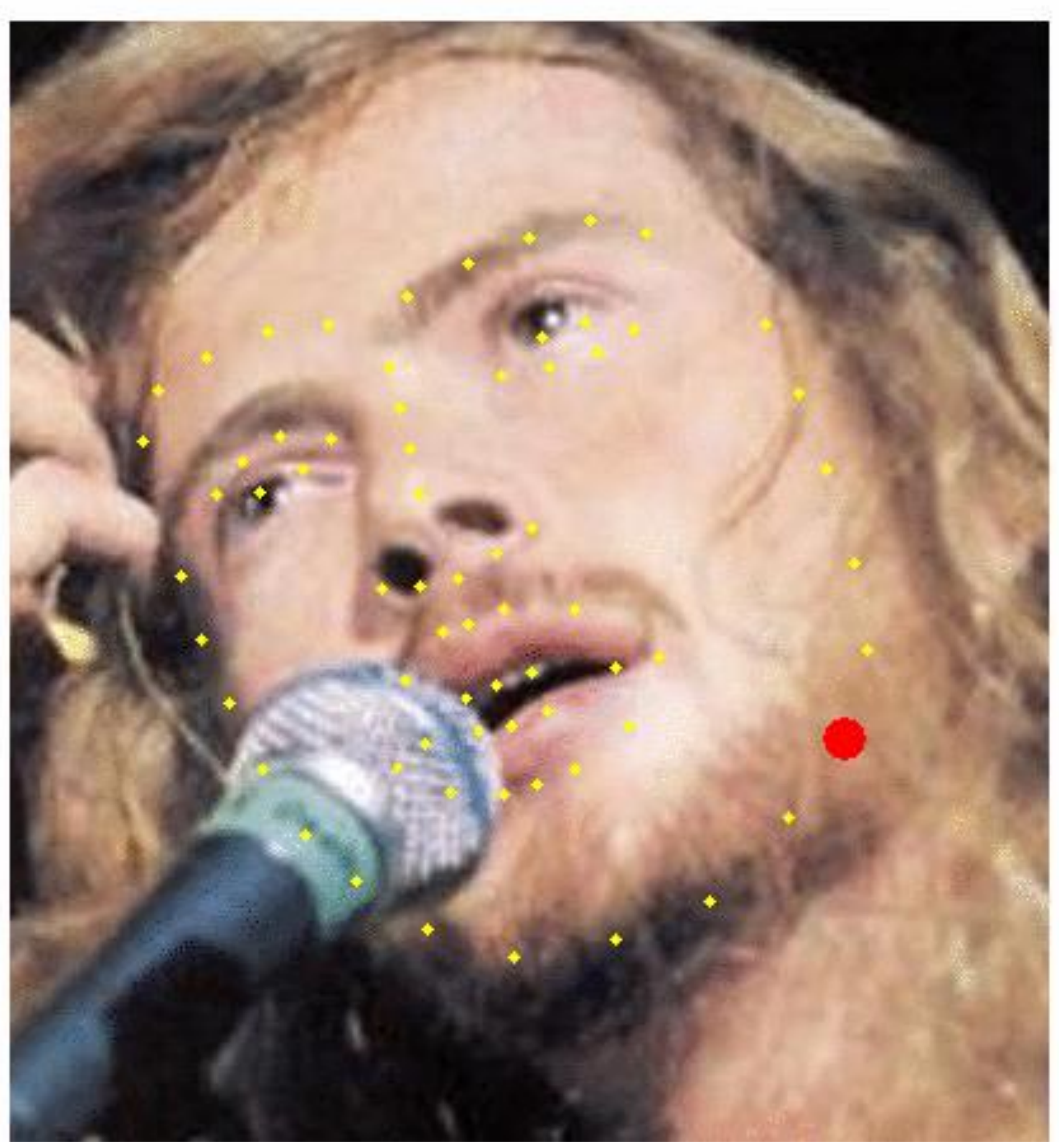


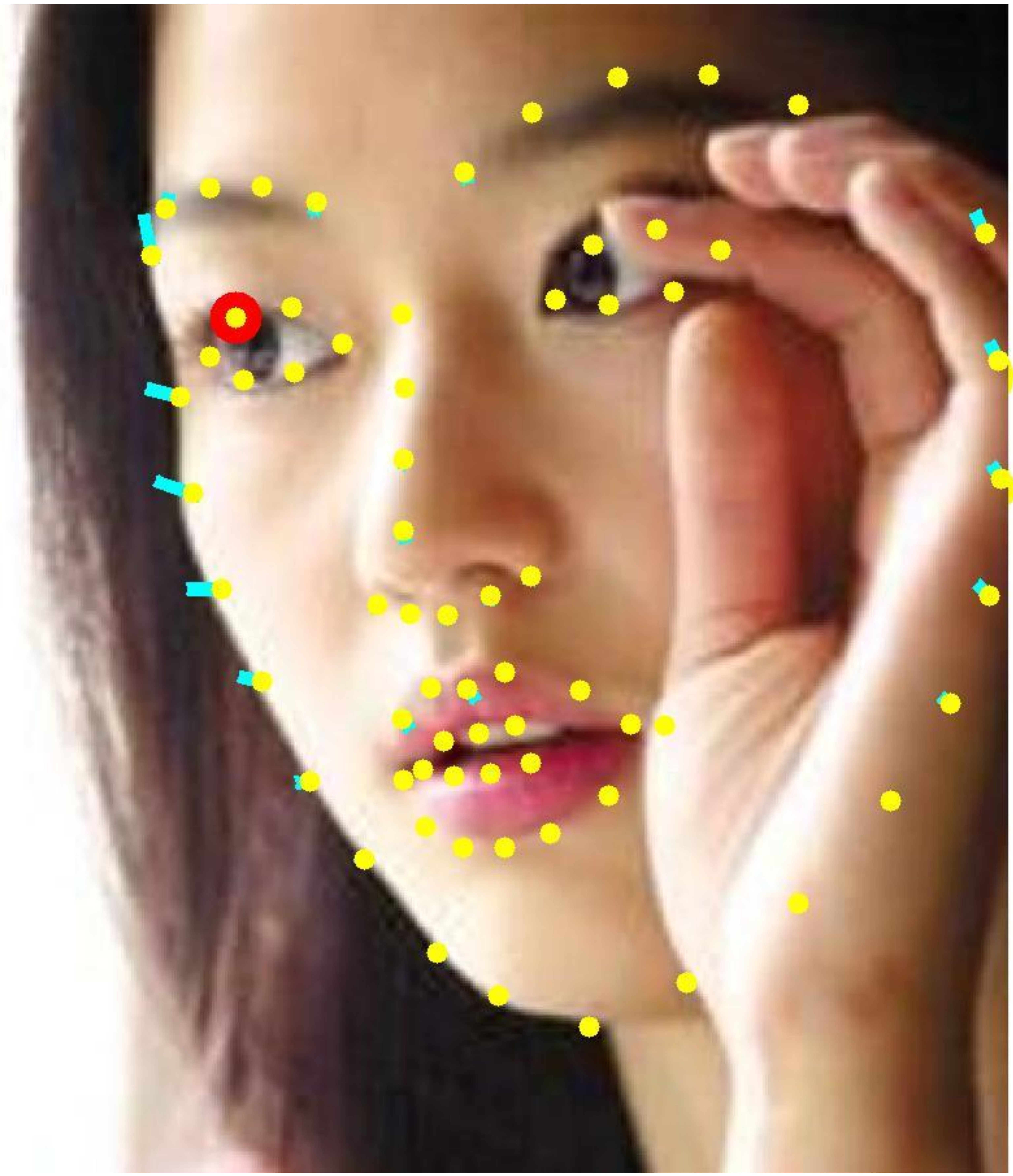
1 2 3

Sample Attentive Refinement

Iter 01

Iter 01



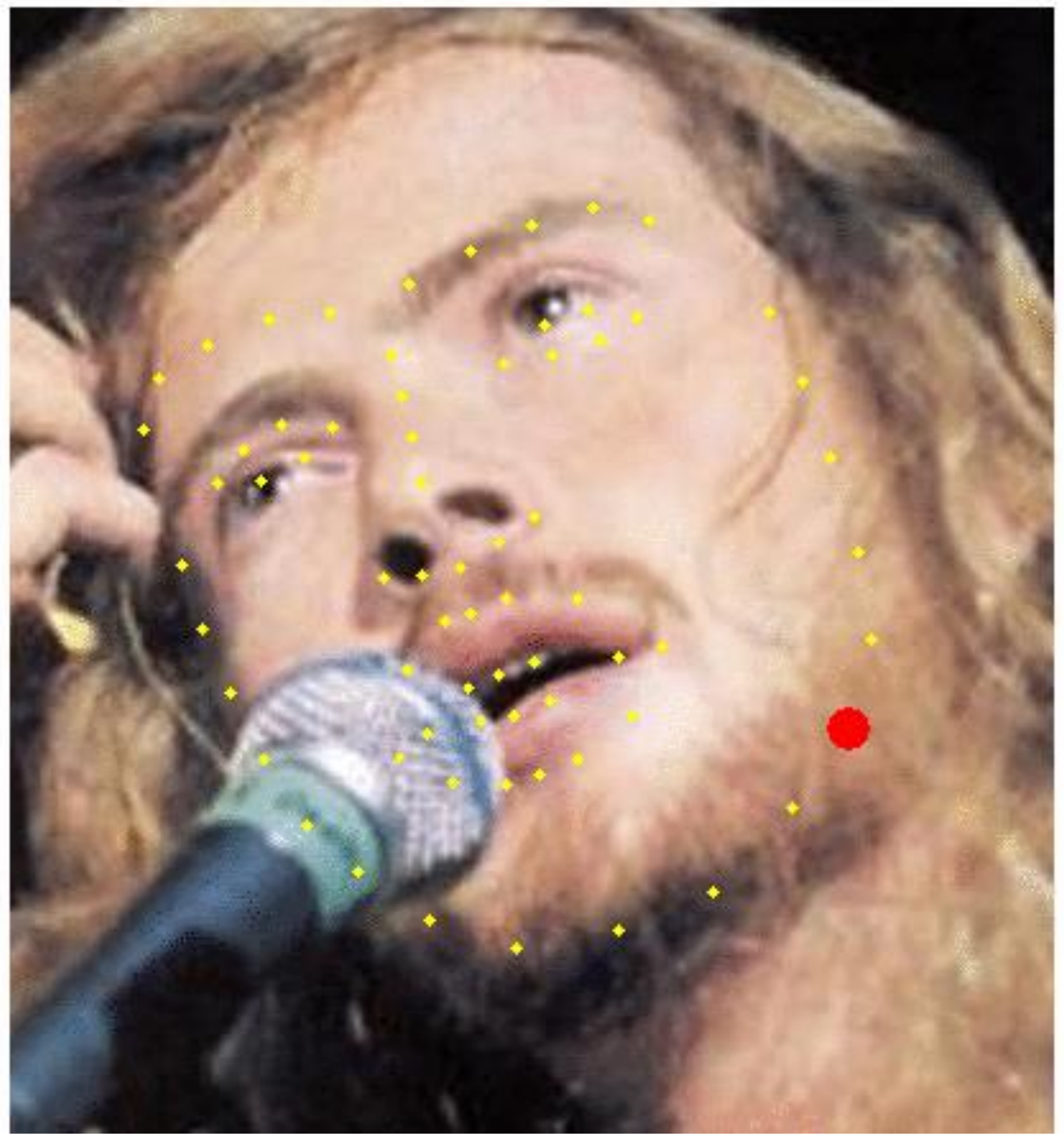


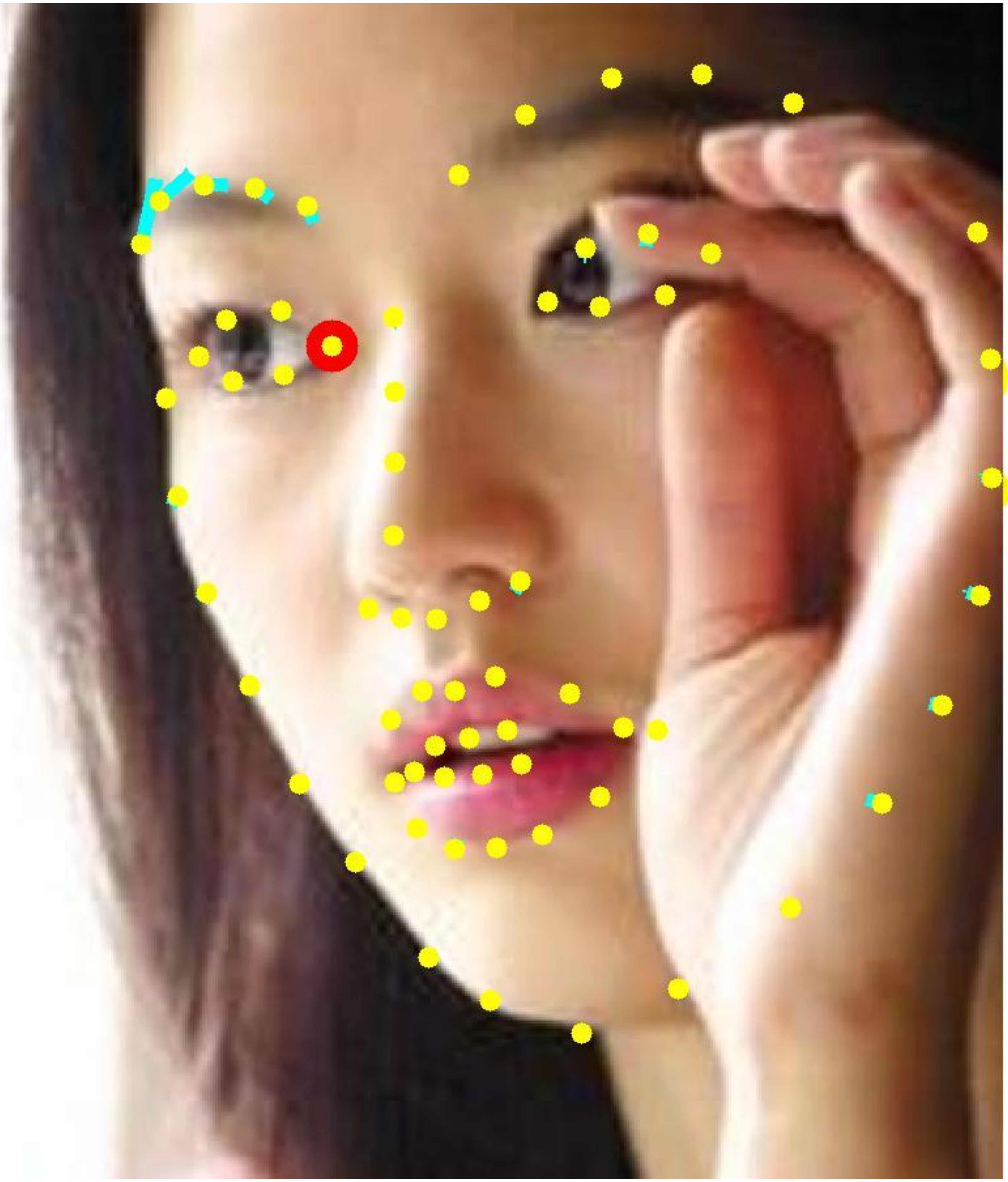
1 2 3 4

Sample Attentive Refinement

Iter 01

Iter 01



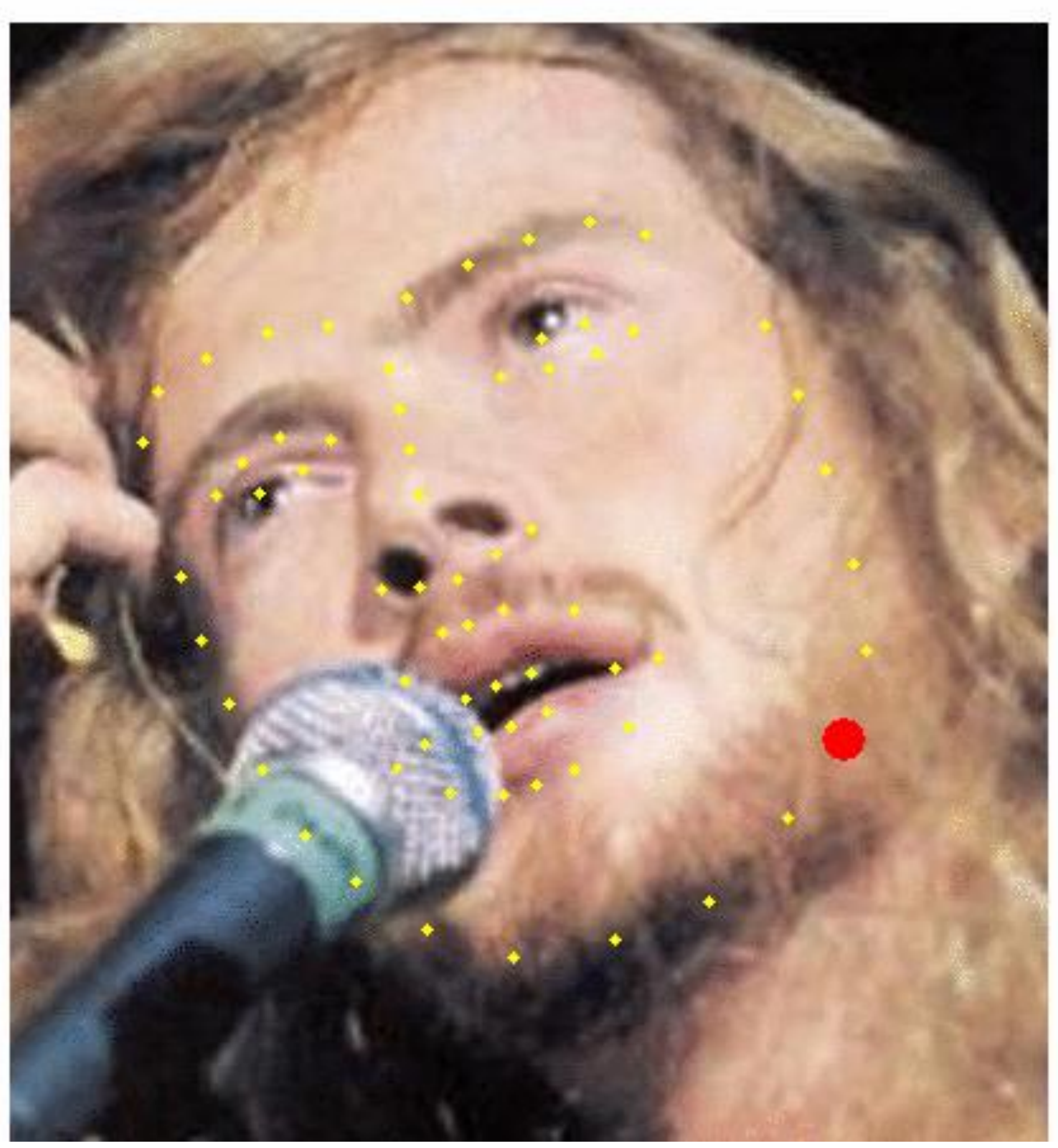


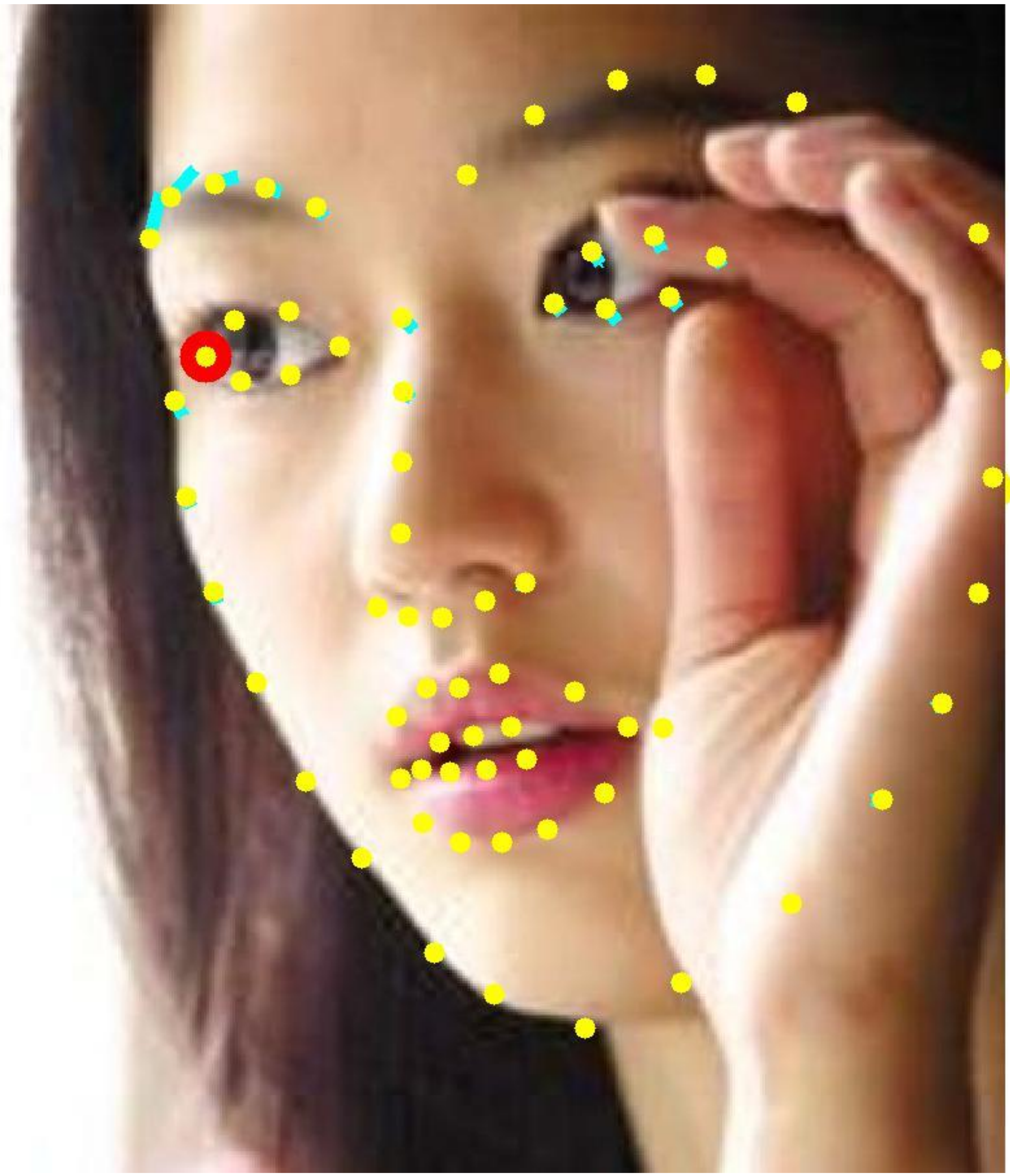
1 2 3 4 5

Sample Attentive Refinement

Iter 01

Iter 01



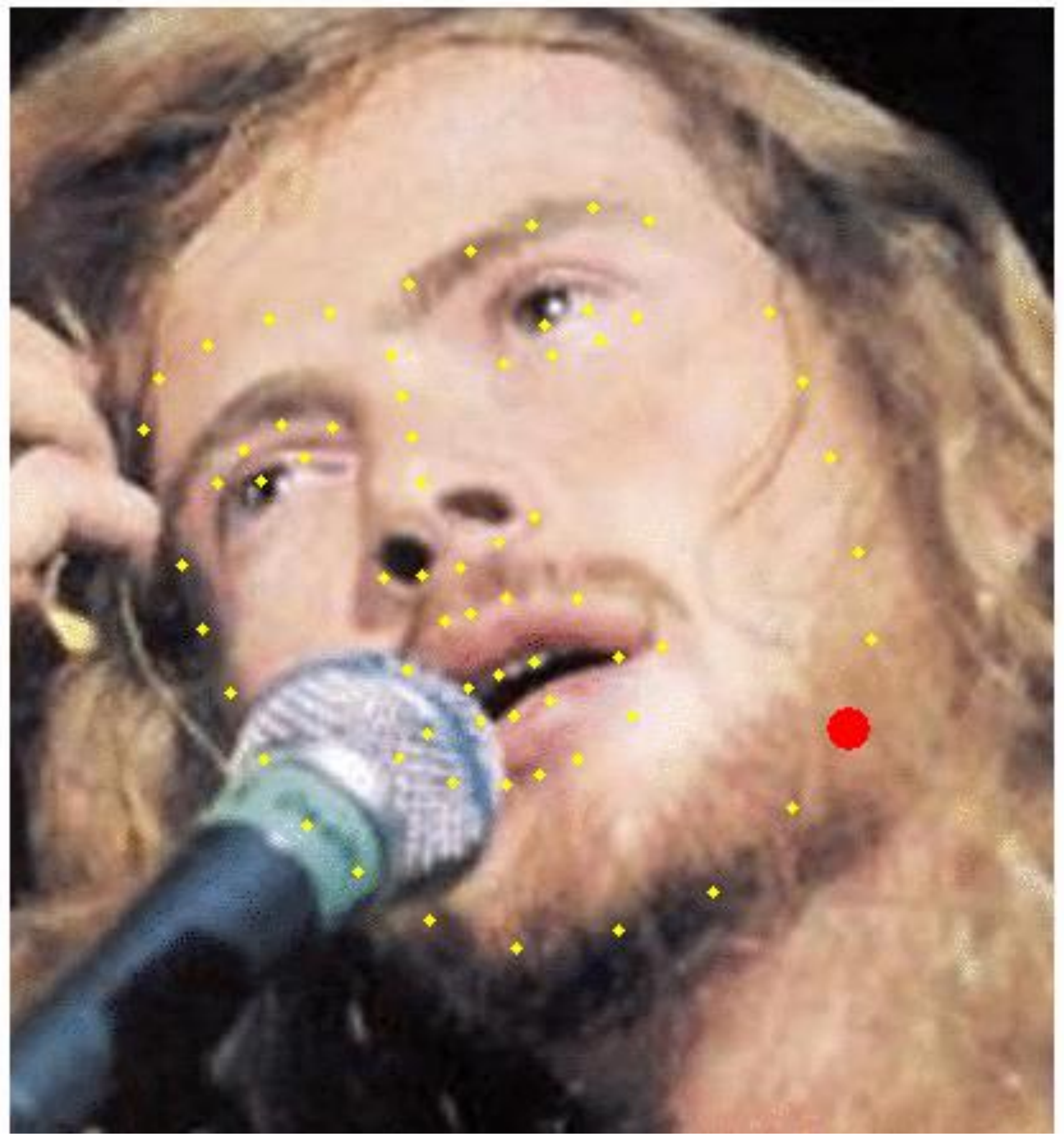


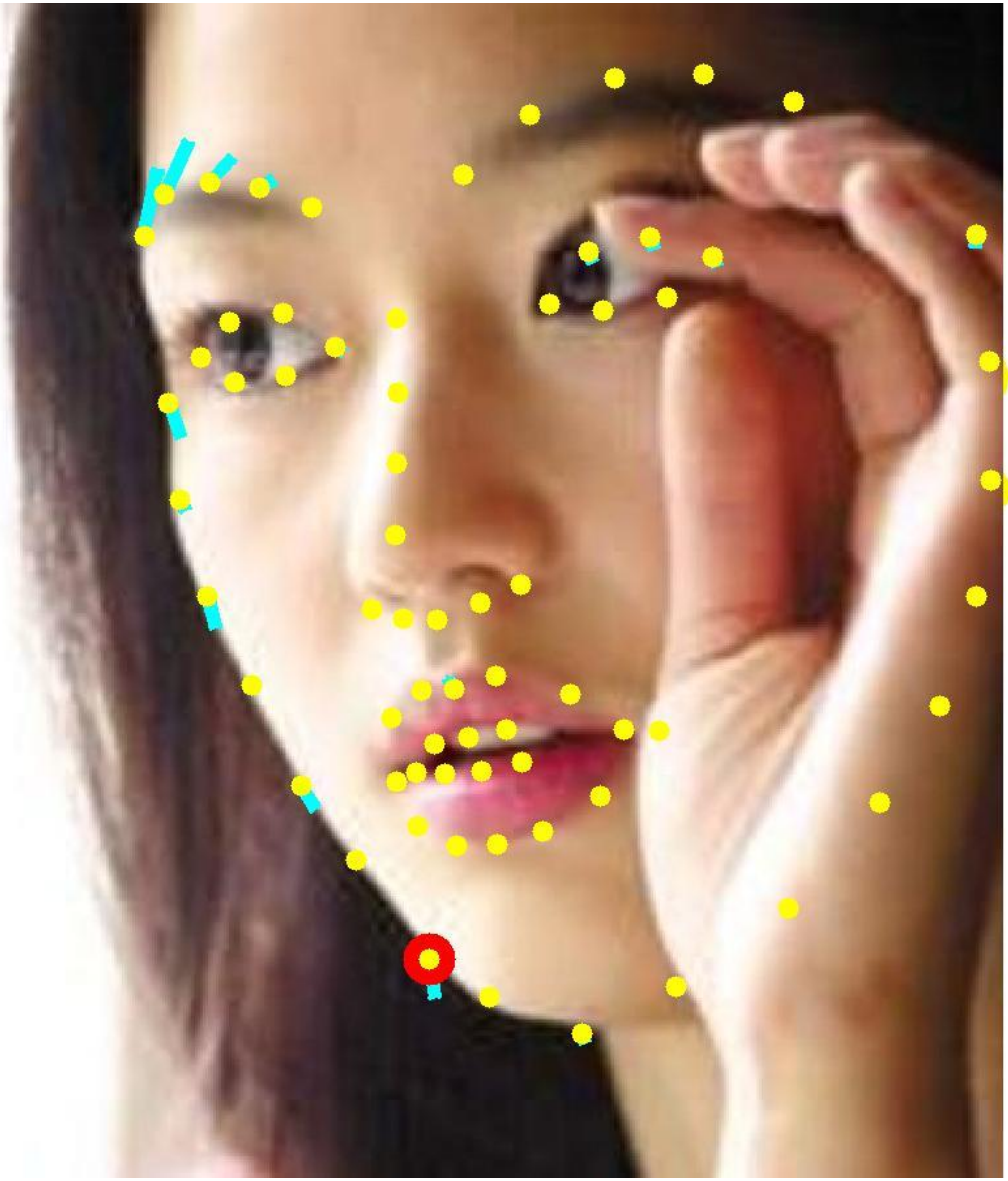
1 2 3 4 5 6

Sample Attentive Refinement

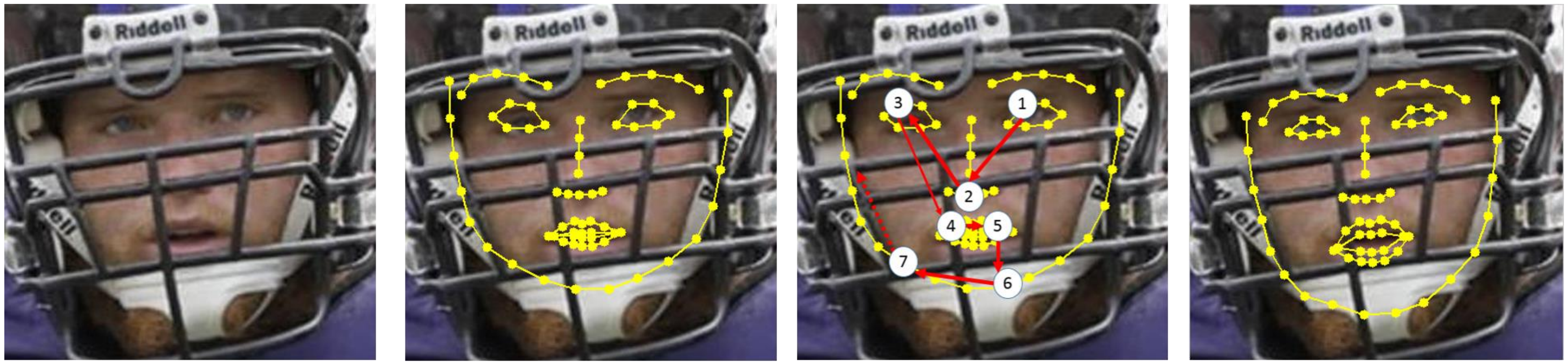
Iter 01

Iter 01





1 2 3 4 5 6 7 8 9 10 11 12 13 14 15



Q&A

Poster Session: O-1A-04