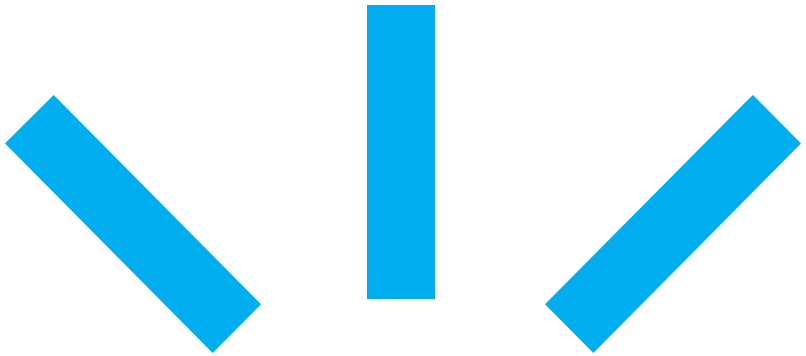


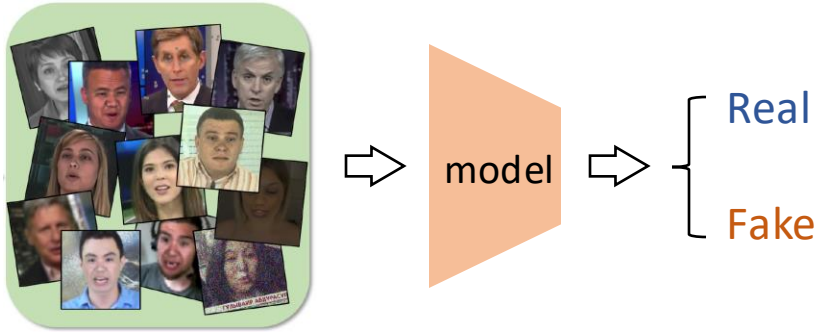
An End-to-End Framework for Joint Deepfake Detection and Fine-Grained Localization

Haotian Liu, Chenhui Pan, Ying Chen, Changfa Mo, Guoying Zhao and Xiaobai Li

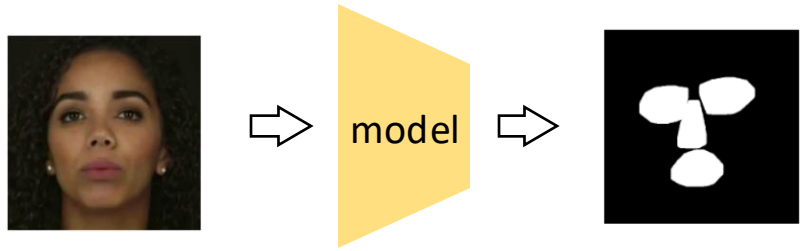
University of Oulu, Zhejiang University

1 Introduction

- **Deepfake Detection:** identify real or fake
- **Deepfake Localization:** locate manipulated region



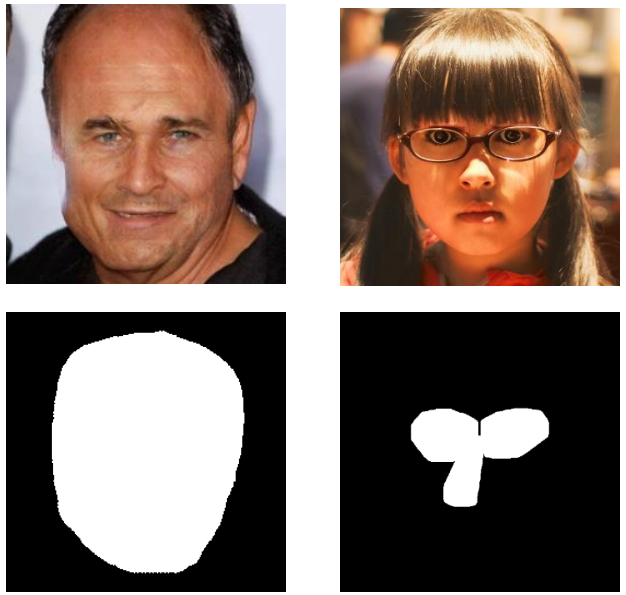
Deepfake Detection



Deepfake Localization

1 Introduction

- Full-face, facial component manipulations
- multiple-face scenarios

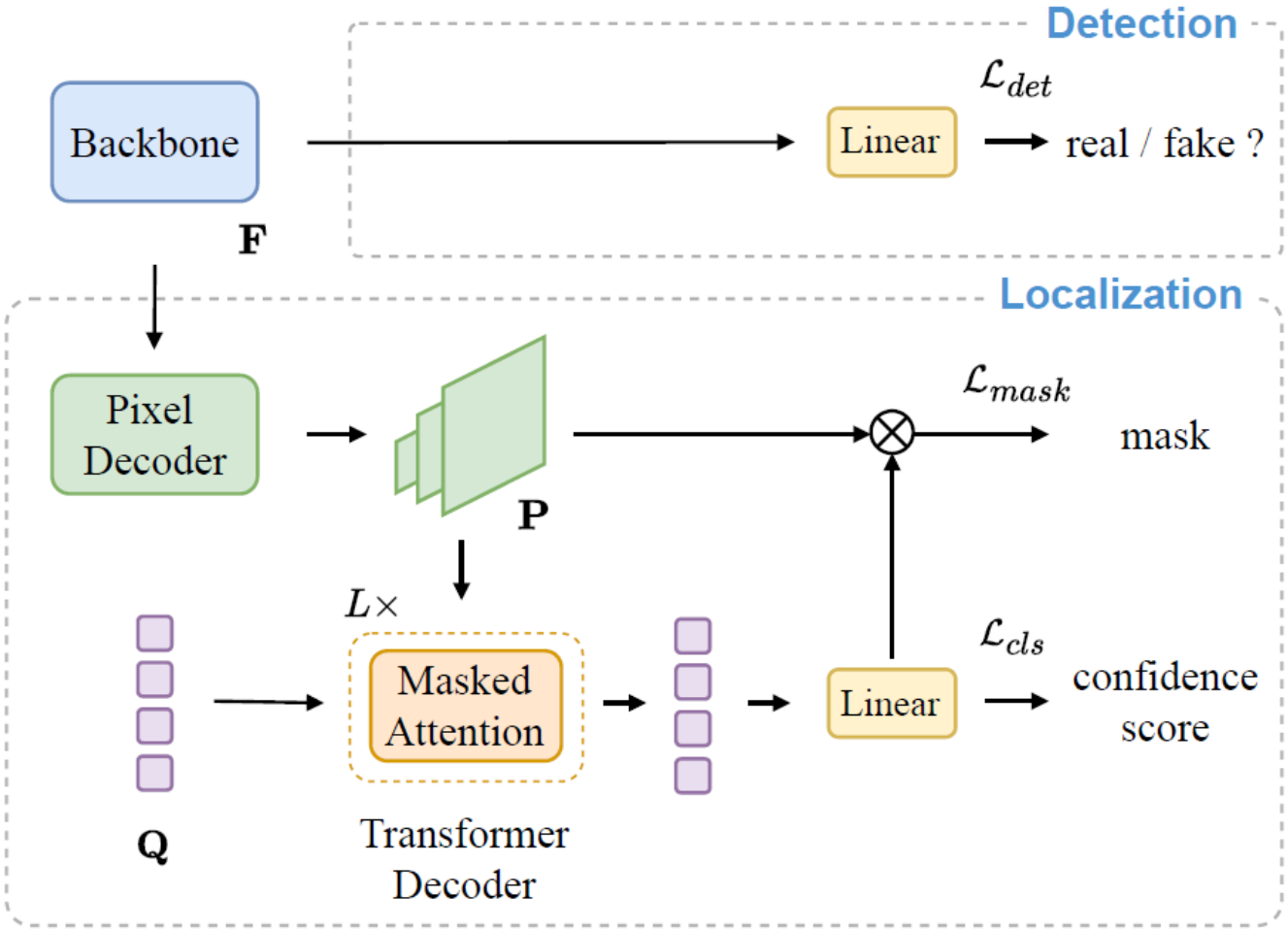


Full-face manipulations,
facial component manipulations

Multiple-face Scenario

2 Methodology

- Our framework jointly performs Deepfake detection and Deepfake localization.

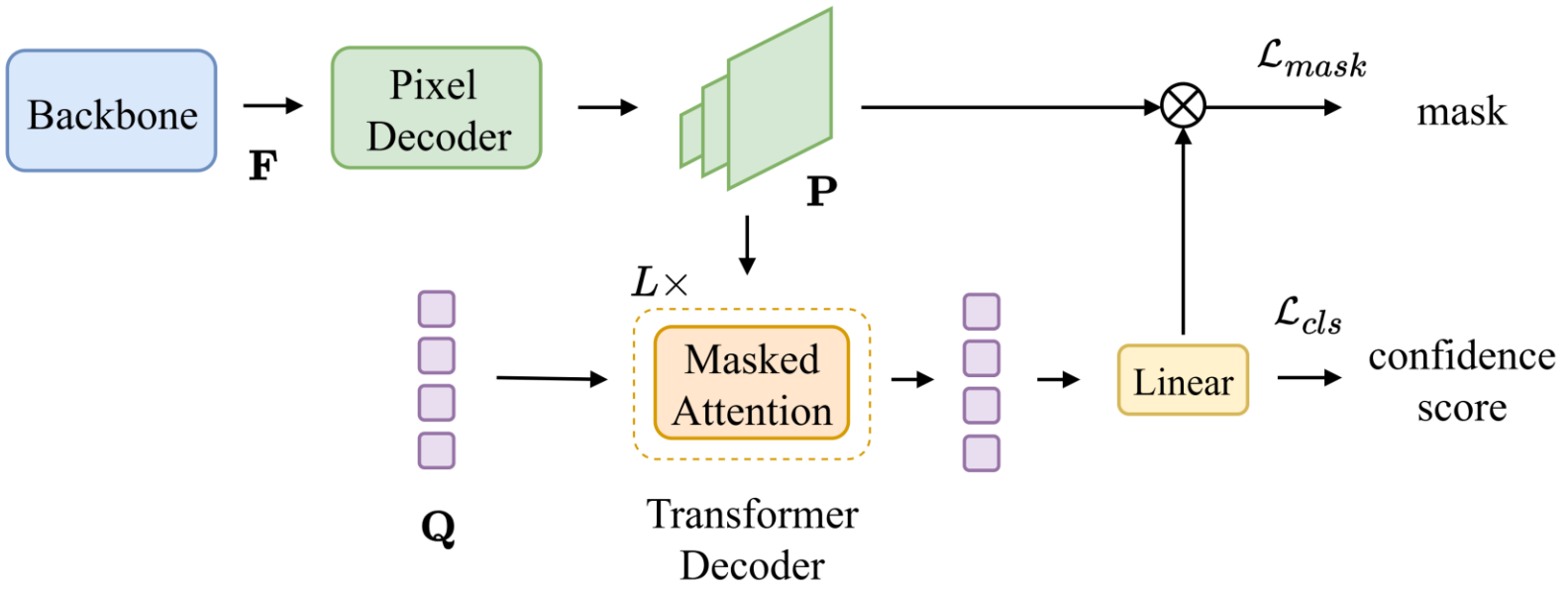


2.1 Masked Deepfake Localization

- Masked-Attention Transformer Decoder

$$\tilde{\mathbf{Q}} = \text{softmax}\left(\mathbf{Q}_l\mathbf{K}_l^T + \mathcal{M}\right)\mathbf{V}_l + \mathbf{Q}.$$

$$\mathcal{M}(i,j) = \begin{cases} 0, & \text{if } \mathbf{M}_{l-1}(i,j) > 0.5, \\ -\infty, & \text{otherwise.} \end{cases}$$



2.2 End-to-End Optimization

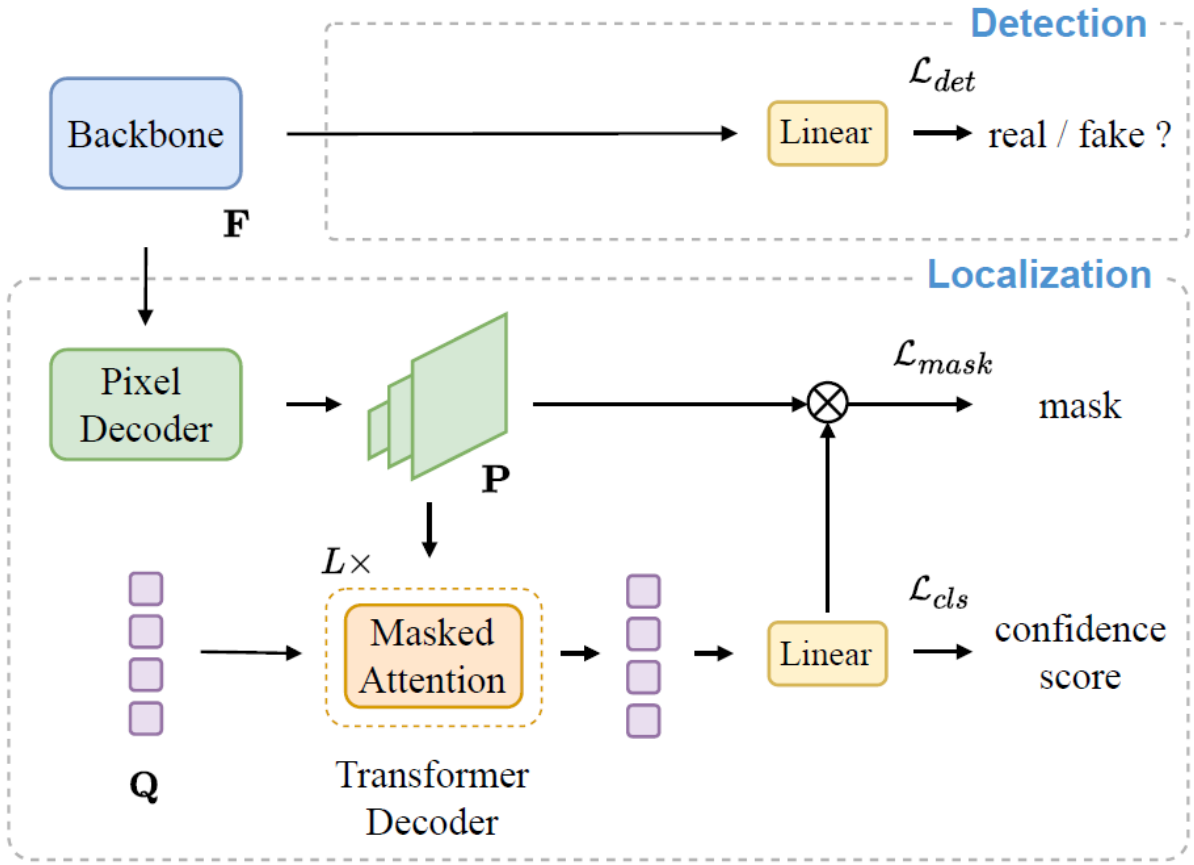
- Deepfake Detection

$$\mathcal{L}_{\text{det}} = \mathcal{L}_{\text{bce}}(y, \hat{y})$$

- Deepfake Localization

$$\mathcal{L}_{\text{loc}} = \mathcal{L}_{\text{mask}} + \mathcal{L}_{\text{cls}}$$

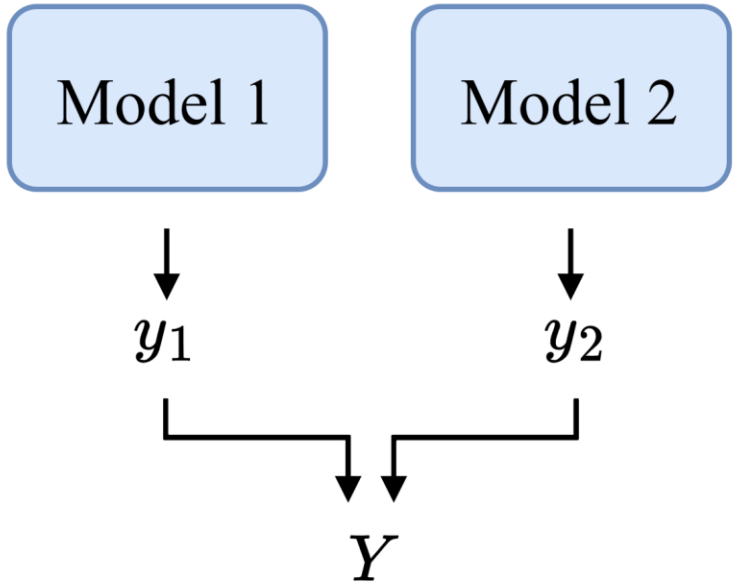
$$\mathcal{L} = \lambda_{det} \cdot \mathcal{L}_{\text{det}} + \lambda_{loc} \cdot \mathcal{L}_{\text{loc}}$$



2.3 Ensemble

- Ensemble the outputs of different models

$$Y = \frac{w_1 \cdot y_1 + w_2 \cdot y_2}{w_1 + w_2}$$



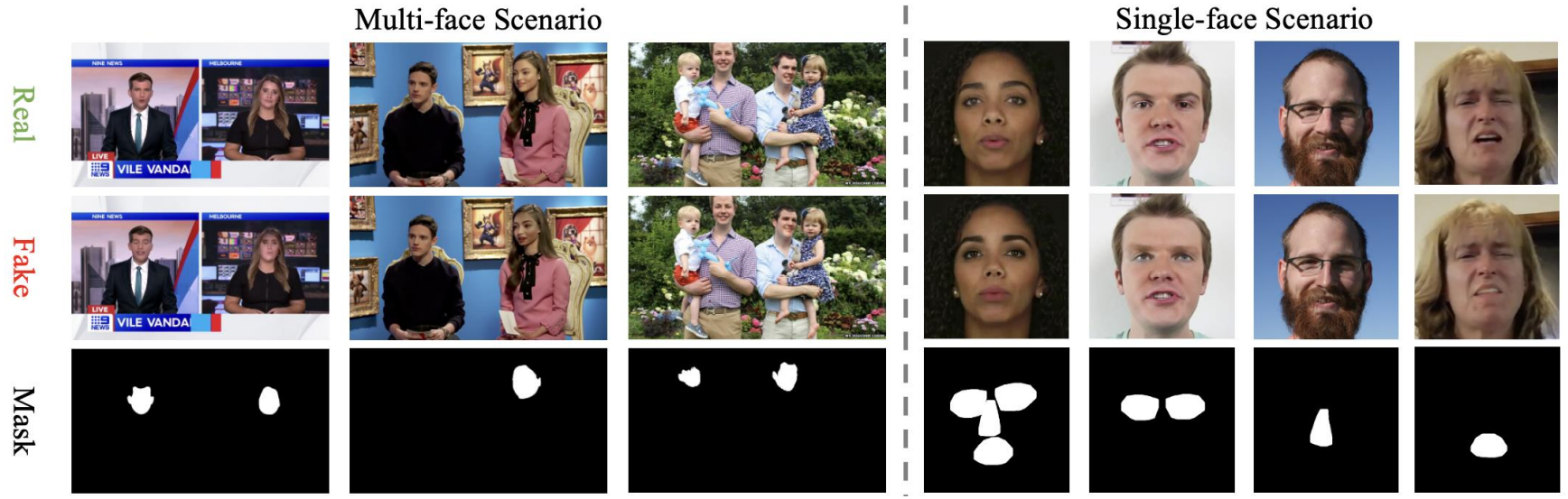
3 Experiments

DDL-I dataset (Deepfake Detection and Localization)

- 1.2 million images with pixel-level annotations
- 61 representative Deepfake methods

Metrics

- Detection: AUC
- Localization: pixel-level F1, IoU



Ddl: A dataset for interpretable deepfake detection and localization in real-world scenarios

3 Implementation Details

- Backbone: Swin-Tiny
- Input shape: 2048×512
- Augmentation
  - random resizing, cropping, and flipping
  - photometric distortions (brightness, contrast, …)

- End-to-End training with both detection and localization annotations

3.1 Comparison Results

Table 1: Comparison results of Deepfake detection and localization on the DDL-I validation set. Bold indicates the best results.

Method	Detection	Localization	
	AUC	F1	IoU
HRNet-w18 [Wang *et al.*, 2020]	99.47	95.17	90.77
UperNet-Swin-T [Xiao *et al.*, 2018]	99.85	96.53	93.28
SegFormer-B5 [Xie *et al.*, 2021]	**99.92**	96.78	93.75
SAM [Kirillov *et al.*, 2023]	99.51	93.70	88.14
IML-ViT [Ma *et al.*, 2023]	-	94.10	90.20
Mesorch [Zhu *et al.*, 2025]	-	94.41	90.40
Ours	98.38	**97.90**	**95.88**

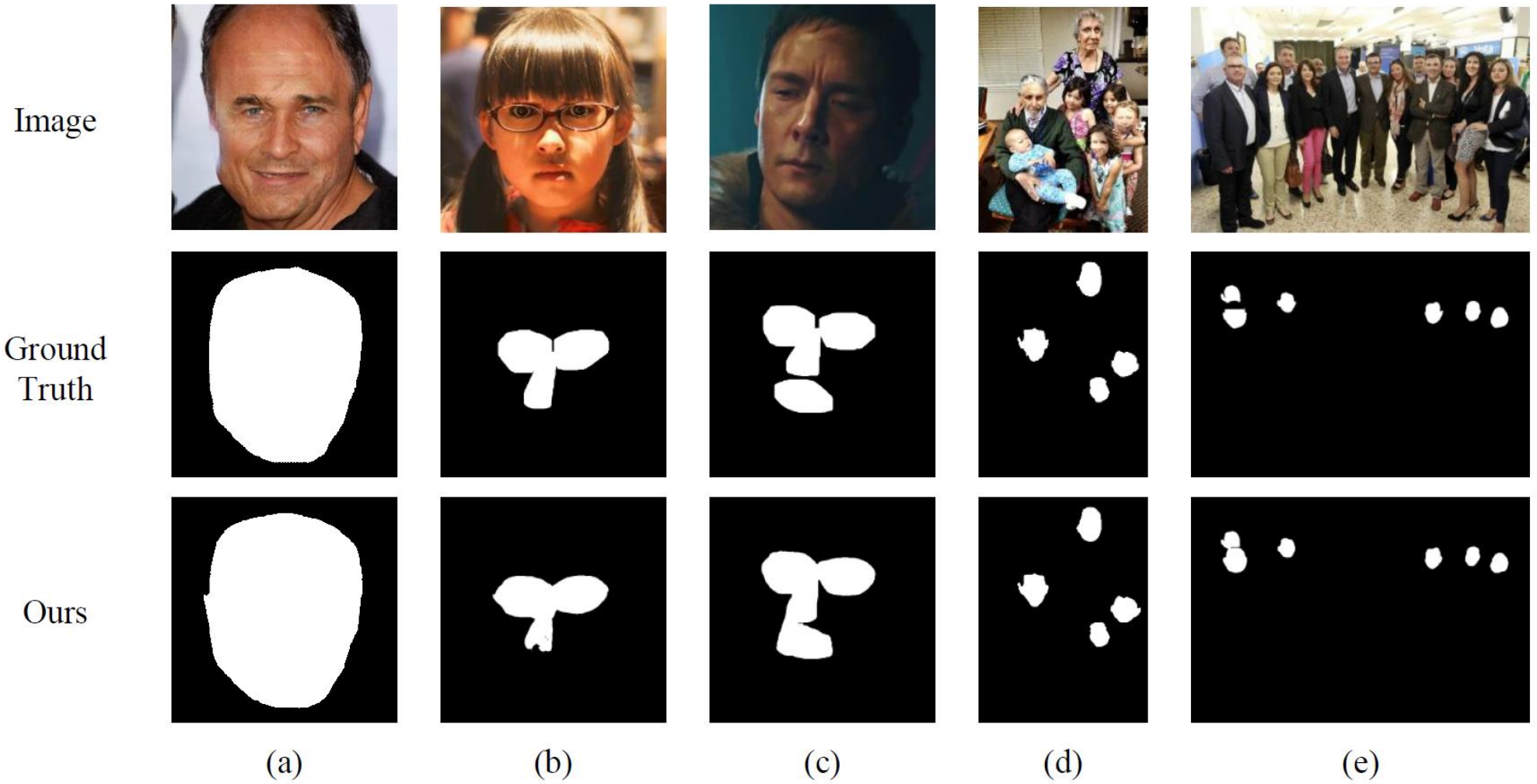
3.2 Ablation Study

Table 2: Ablation results of backbone selection and ensemble strategies on the DDL-I testing set.

Method	Detection	Localization	
	AUC	F1	IoU
Single backbone			
ResNet-50	88.32	72.76	66.58
Swin-T	93.42	77.78	71.79
Swin-S	94.24	76.18	70.79
Ensemble (Swin-T + Swin-S)			
Average	94.01	77.07	71.44
Re-weighted	94.95	-	-
Ours	94.95	77.78	71.79

Overall score: 81.50%

3.3 Visualization



full-face forgeries (a), manipulations of specific facial components (b–c), and complex scenes containing both real and fake faces (d–e).

4 Conclusion

Summary

- We propose an end-to-end framework that jointly performs Deepfake detection and fine-grained localization.
- We utilize learnable queries and masked attention mechanisms to suppress background noise and locate small and sparse forgery regions.
- Our final solution achieving a leading position in the IJCAI Deepfake Challenge.

Future work

- Explore the underlying relationship between Deepfake detection and localization.

Thank you!