FaceXFormer: A Unified Transformer for Facial Analysis

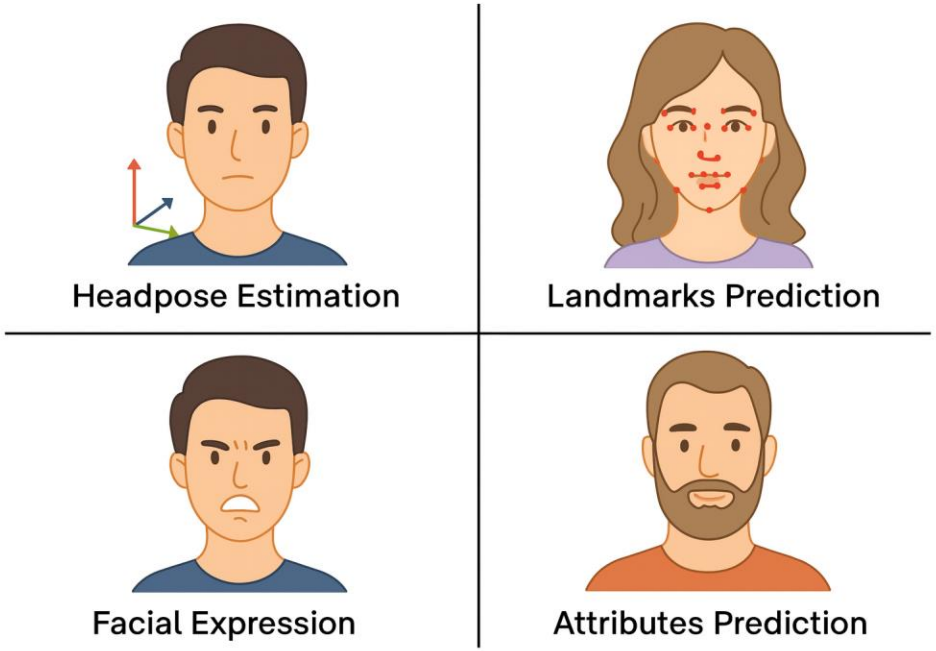
Kartik Narayan*, Vibashan VS*, Rama Chellappa, Vishal M. Patel

Project Page: https://kartik-3004.github.io/facexformer/
Paper: https://arxiv.org/pdf/2403.12960v2
GitHub: https://github.com/Kartik-3004/facexformer

* Equal Contribution

What is Facial Analysis ?

- Facial analysis refers to interpreting human faces in images or videos.
- Facial analysis includes tasks such as:
  - Headpose Estimation
  - Landmarks Prediction
  - Facial Expression Recognition
  - Attributes prediction (Beard, Moustache,)
    … **(more tasks)**
- Applications in:
  - Online Examination Monitoring
  - Surveillance
  - Retail and Marketing
  - Healthcare/Patient Monitoring



Facial Analysis Tasks

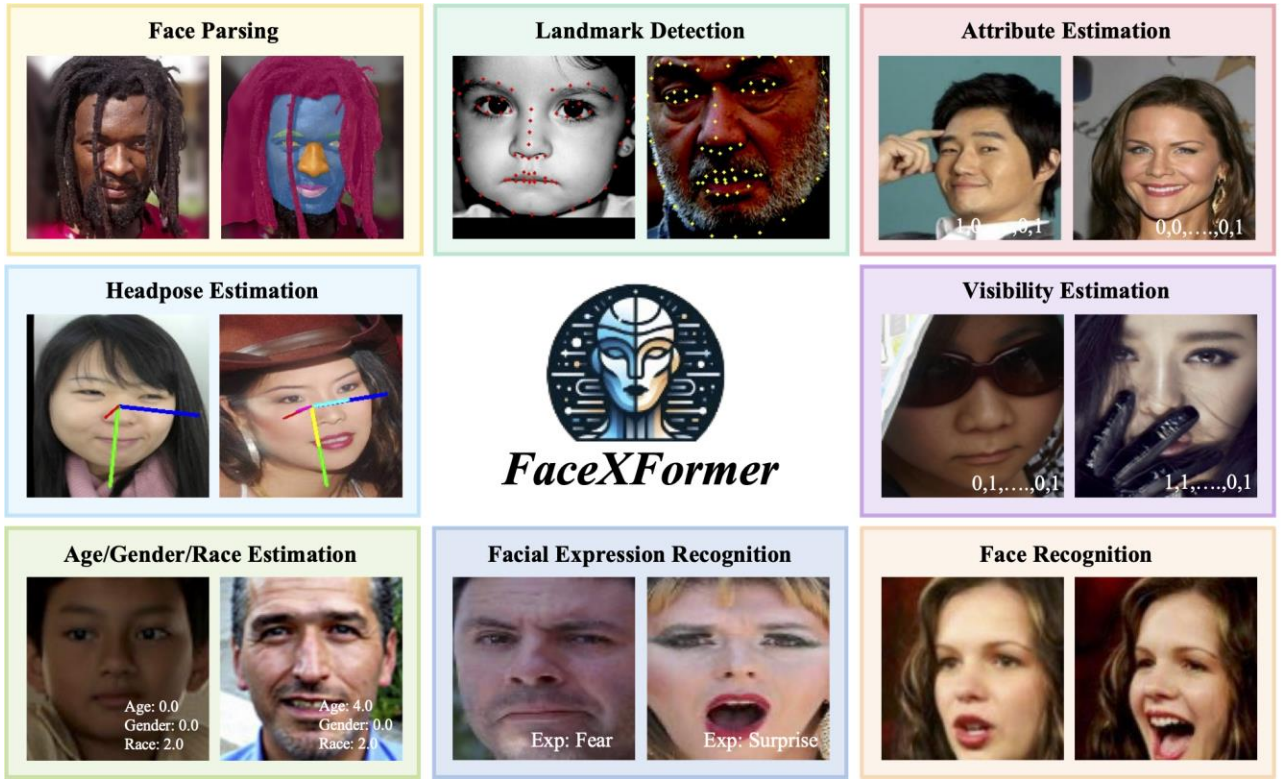
1. Face Parsing
2. Landmarks Detection
3. Headpose Estimation
4. Attributes Prediction
5. Age Estimation
6. Gender Prediction
7. Race Estimation
8. Visibility Prediction
9. Facial Expression Recognition
10. Face Recognition

Drawback of specialized models

- Tasks-specific Architectures
  - Developing and running multiple models
  - Computationally expensive
  - Increases system complexity
  - Inefficient for real-time applications

Unification !!!

- A single unified model can learn shared, task-invariant facial representations, which improves robustness, especially in challenging “in-the-wild” conditions.
- A unified architecture enables faster inference.



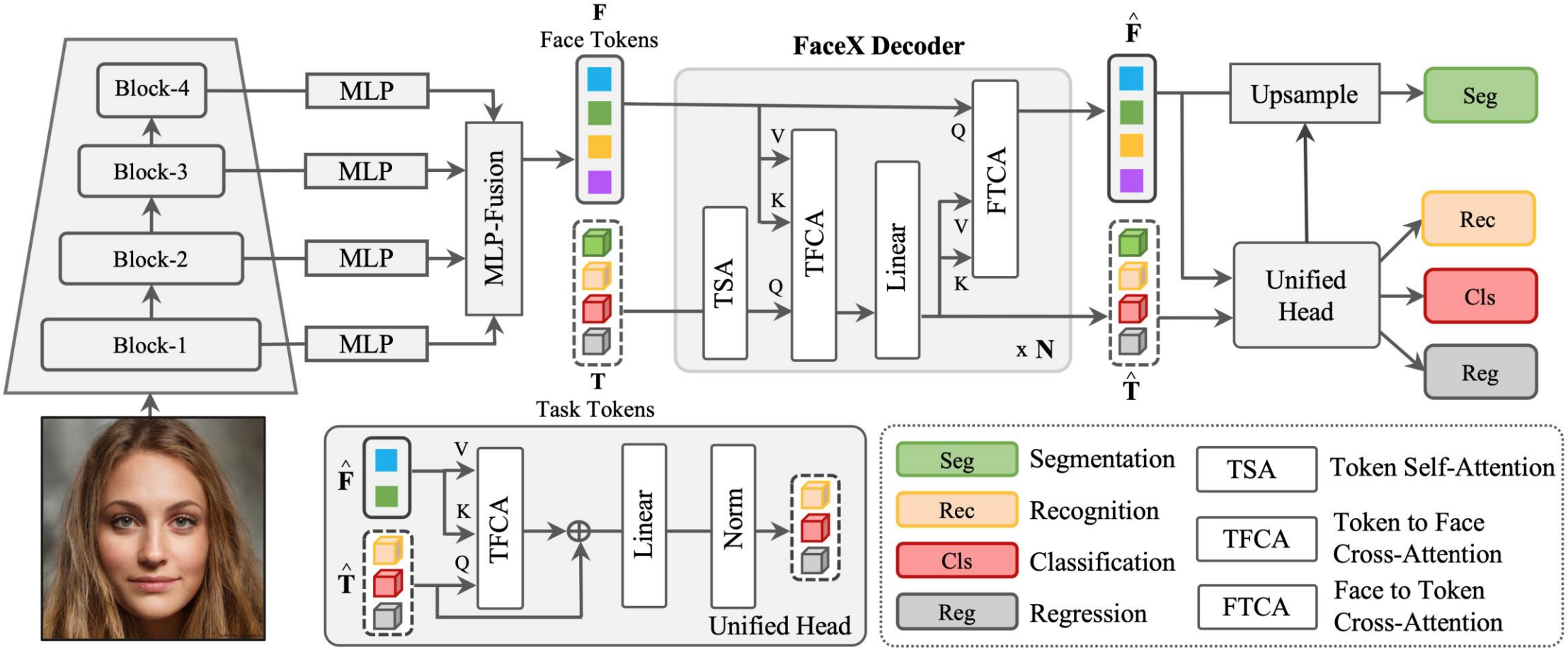
Previous multi-task models

- All-in-One and HyperFace are old works which use convnets.
- QFace and Faceptor are recent methods but cater to less tasks and are architecturally heavy, which increases inference time.

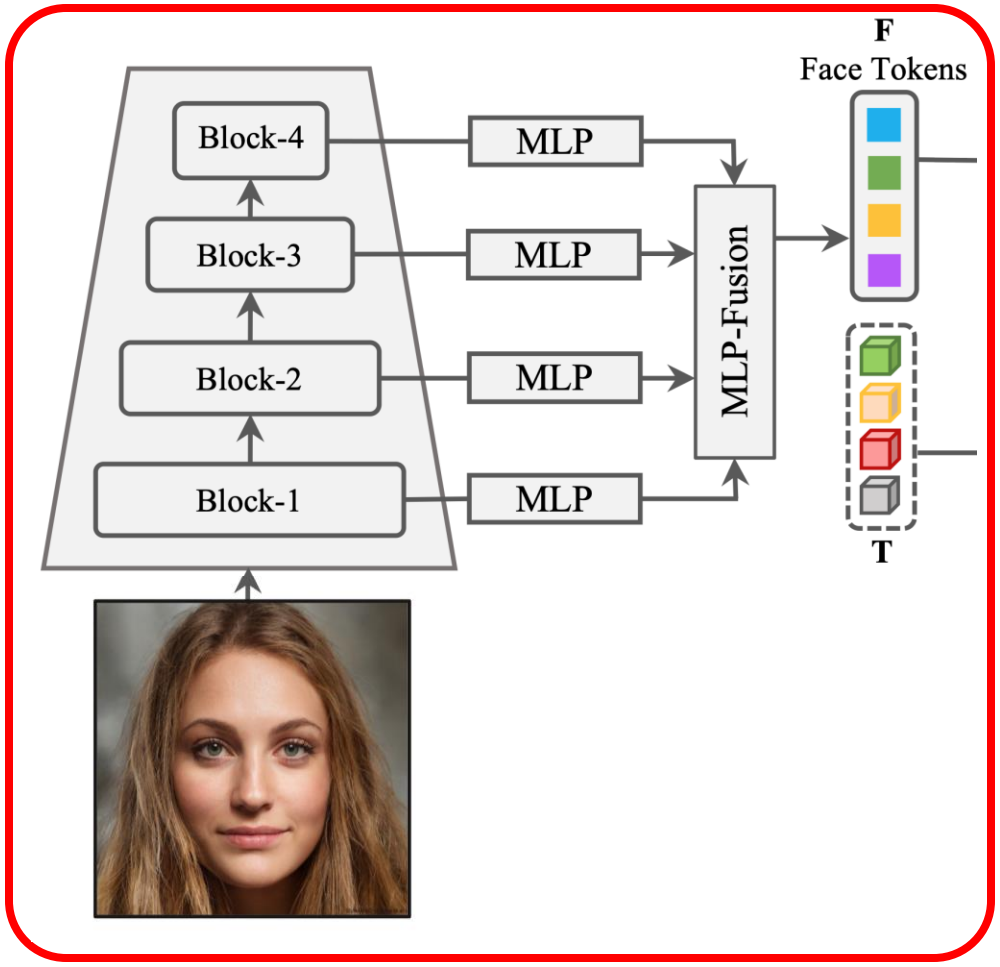
	QFace [91]	**Faceptor** [74]	***FaceXFormer***
Tasks	4	7	10
Model Size	-	178.9 M	109.29 M
Face Pretrained	Yes	Yes	No
FPS (fp32)	-	14.30	33.21
Decoder	Query2Label [54]	Pixel Decoder & Transformer Decoder	FaceX
Decoder Layers	9	9	2

Methods	FP	LD	HPE	Attr	Age	Gen	Race	Vis	Exp	FR
Single-Task Models										
DML-CSR [130]	✓									
FP-LIIF [83]	✓									
SegFace [69]	✓									
Wing [23]		✓								
HRNet [101]		✓								
WHENet [134]			✓							
TriNet [10]			✓							
img2pose [3]			✓							
TokenHPE [124]			✓							
SSPL [88]				✓						
VOLO-D1 [42]					✓					
DLDL-v2 [24]					✓					
3DDE [97]								✓		
MNN [98]								✓		
KTN [46]									✓	
DMUE [85]									✓	
CosFace [100]										✓
ArcFace [16]										✓
AdaFace [39]										✓
Multi-Task Models										
SSP+SSG [35]	✓			✓						
Hetero-FAE [28]				✓	✓	✓	✓		✓	
FairFace [36]					✓	✓	✓			
MiVOLO [42]					✓	✓				
MTL-CNN [141]		✓		✓					✓	
ProS [18]	✓	✓		✓						
FaRL [133]	✓	✓		✓	✓	✓				
HyperFace [77]		✓	✓			✓		✓		✓
AllinOne [78]		✓	✓		✓	✓			✓	✓
Swinface [73]				✓	✓				✓	✓
QFace [91]			✓	✓	✓				✓	
Faceptor [74]	✓	✓		✓	✓	✓			✓	✓
FaceXFormer	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

FaceXFormer Architecture



FaceXFormer Architecture

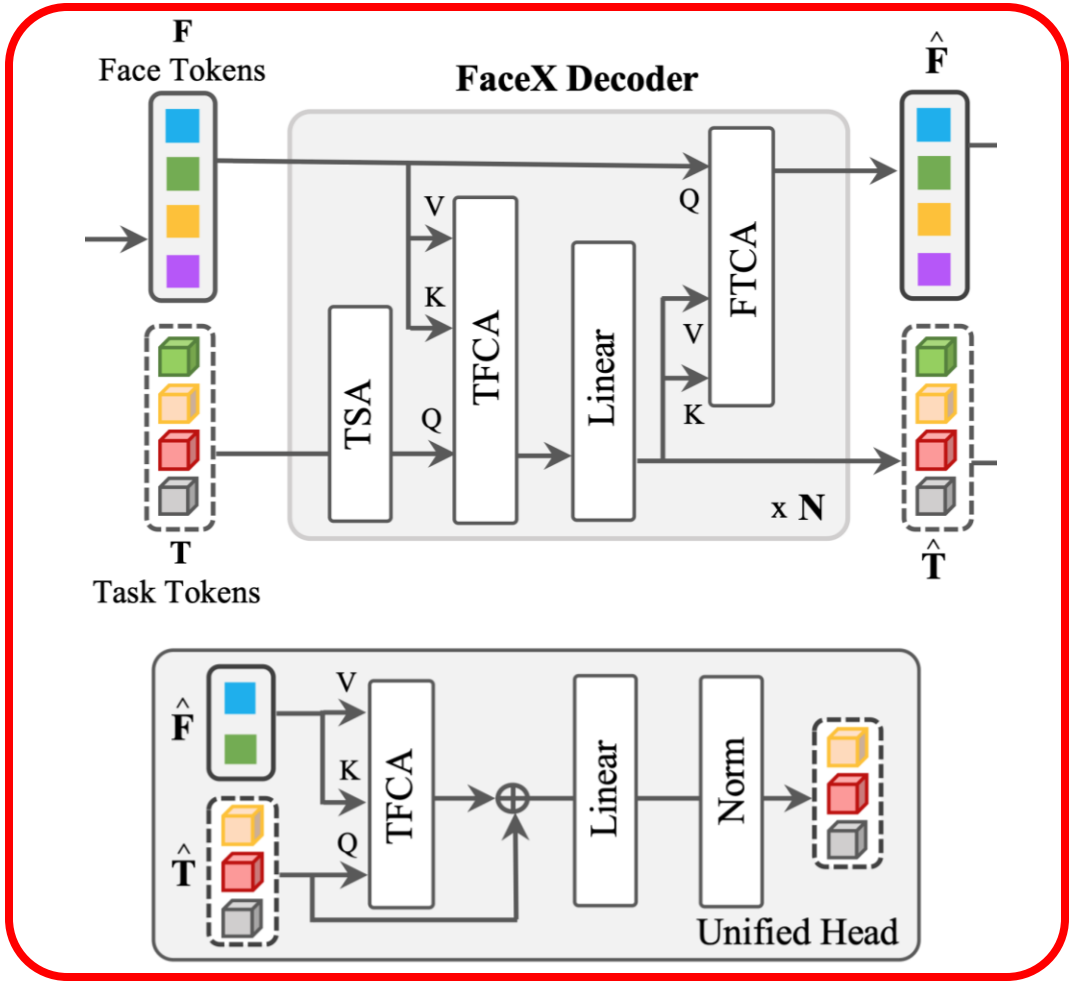


- We use Swin-B transformer as our backbone.
- We extract multi-scale features from intermediate layers.
- We pass the extracted features to an MLP layer to modify the feature dimension to make enable fusion.
- We pass it through an MLP fusion to obtain face tokens.

- Block 1: [16, 128, 128, 128] → [16, 512, 128, 128]
- Block 2: [16, 256, 64, 64] → [16, 512, 128, 128]
- Block 3: [16, 512, 32, 32] → [16, 512, 128, 128]
- Block 4: [16, 1024, 16, 16] → [16, 512, 128, 128]

After 1x1 Conv:
[16, 512, 128, 128]

FaceXFormer Architecture

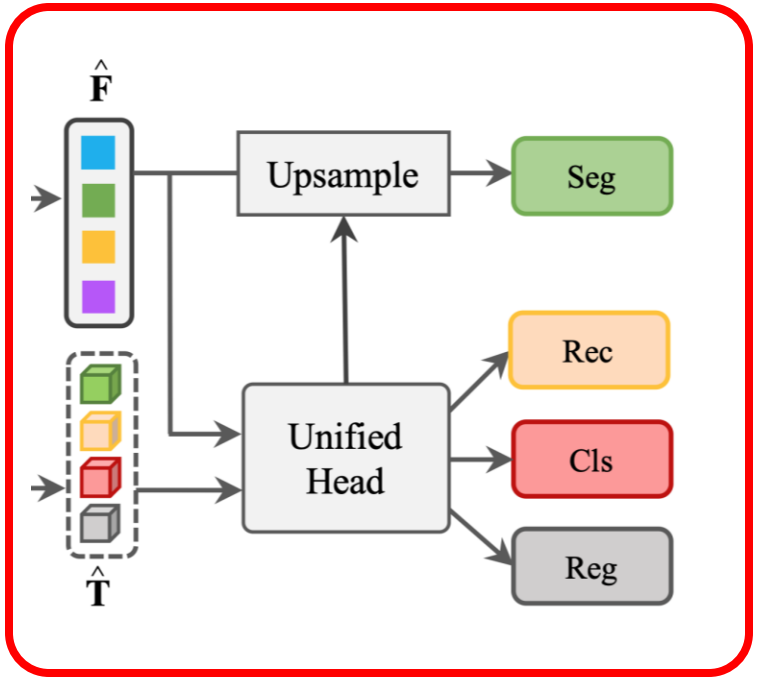


(1) TSA (Task Self-attention):
Each task token is updated by interacting with other set of task tokens and captures task-specific interactions.

(2) TFCA (Task to Face Cross Attention):
Task tokens interact with fused face features which facilitates task-focused feature extraction.

(3) FTCA (Face to Task Cross Attention):
We use the updated task tokens and perform another cross-attention, which enables face features to be modified with task-specific details.

FaceXFormer Architecture



- Segmentation Head:
  - Transpose Convolution
- Recognition Head:
  - Partial FC
- Landmark Prediction Head:
  - StackedHGNet
- Other Heads:
  - MLP

Results

Method	Input Res.	Skin	Hair	Nose	L-Eye	R-Eye	L-Brow	R-Brow	L-Lip	I-Mouth	U-Lip	Mean F1 ↑
Wei et al. [105]	512	96.4	91.1	91.9	87.1	85.0	80.8	82.5	91.0	90.6	87.9	88.43
EHANet [55]	512	96.0	93.9	93.7	86.2	86.5	83.2	83.1	90.3	93.8	88.6	89.53
EAGRNet [91]	473	96.2	94.9	**94.0**	88.6	89.0	85.7	85.2	**91.2**	**95.0**	88.9	90.87
AGRNet [92]	473	96.5	87.6	93.9	88.7	89.1	85.5	85.6	91.1	92.0	89.1	89.91
$FaRL_{scratch}$ [130]	512	96.2	94.9	93.8	89.0	89.0	85.3	85.4	90.0	91.7	88.1	90.34
DML-CSR [127]	473	95.7	94.5	93.9	89.4	89.6	85.5	85.7	91.0	91.8	89.1	90.62
FP-LIIF [80]	256	96.4	95.1	93.7	88.5	88.5	84.5	84.3	90.3	92.1	87.5	90.09
SwinFace [71]	×	×	×	×	×	×	×	×	×	×	×	×
QFace [88]	×	×	×	×	×	×	×	×	×	×	×	×
Faceptor [72]	512	**96.6**	**96.2**	93.9	89.4	89.1	**86.2**	**86.3**	90.6	91.6	89.0	90.89
FaceXFormer	224	96.4	95.7	93.8	**90.1**	**90.3**	86.0	85.9	90.6	92.1	**89.2**	**91.01**

Table 3. Performance comparison for face parsing on the CelebAMask-HQ dataset [42]. The symbol × indicates that the model does not perform the corresponding task. **Red** = First Best, Blue = Second Best. × indicates a model that doesn't perform the task.

Results

Methods	Expression (RAF-DB) ↑	Methods	Visibility (COFW) ↑	Methods	Age (MAE) UTKFace ↓
DLP-CNN [45]	80.89	RCPR [7]	40	OR-CNN [68]	5.74
gACNN [48]	85.07	Wu et al. [111]	44.43	Axel Berg et al. [5]	4.55
IPA2LT [120]	86.77	Wu et al. [110]	49.11	CORAL [8]	5.47
RAN [101]	86.90	ECT [122]	63.4	Gustafsson et al. [26]	4.65
CovPool [2]	87.00	3DDE [94]	63.89	R50-SORD [70]	4.36
SCN [100]	87.03	MNN [95]	72.12	VOLO-D1 [40]	4.23
DACL [21]	87.78			DLDL-v2 [23]	4.42
KTN [44]	88.07			MWR [84]	4.37
DMUE [82]	88.76				
SwinFace [71]	86.54	SwinFace [71]	×	SwinFace [71]	×
QFace [88]	**92.86**	QFace [88]	×	QFace [88]	×
Faceptor [72]	87.58	Faceptor [72]	×	Faceptor [72]	**4.10**
FaceXFormer	**88.24**	***FaceXFormer***	**72.56**	***FaceXFormer***	4.17

Results

Methods	Headpose (BIWI) Yaw	Pitch	Roll	MAE ↓	Methods	Landmarks (300W) Full ↓	Com ↓	Chal ↓	Methods	CelebA Acc. ↑
HopeNet [78]	4.81	6.61	3.27	4.89	LAB [109]	3.49	2.98	5.19	PANDA-1 [123]	85.43
QuatNet [30]	5.49	4.01	2.94	4.15	Wing [22]	4.04	3.27	7.18	LNets+ANet [53]	87.33
FSA-Net [116]	4.27	5.49	2.93	4.14	DeCaFa [14]	3.39	2.93	5.26	SSP+SSG [34]	88.24
EVA-GCN [114]	6.01	4.78	2.98	3.98	HRNet [98]	3.32	2.87	5.15	MOON [77]	90.94
TriNet [9]	4.11	4.75	3.04	3.97	PicassoNet [106]	3.58	3.03	5.81	NSA [56]	90.61
img2pose [3]	4.56	**3.54**	3.24	3.78	AVS+SAN [18]	3.86	3.21	6.46	MCNN-AUX [28]	91.29
MNN [95]	3.98	4.61	**2.39**	3.66	LUVLi [39]	3.23	2.76	5.16	MCFA [137]	91.23
MFDNet [51]	**3.40**	4.68	2.77	3.62	HIH [41]	3.09	**2.65**	4.89	DMM-CNN [57]	91.70
TokenHPE [121]	3.95	4.51	2.71	3.72	PIPNet [33]	3.19	2.78	4.89	SSPL [85]	91.77
WHENet [131]	3.99	4.39	3.06	3.81	SLPT [112]	3.17	2.75	4.90	FaRL [130]	91.39
SwinFace [71]	×	×	×	×	SwinFace [71]	×	×	×	SwinFace [71]	91.38
QFace [88]	–	–	–	–	QFace [88]	×	×	×	QFace [88]	91.56
Faceptor [72]	×	×	×	×	Faceptor [72]	3.16	2.75	4.84	Faceptor [72]	91.39
FaceXFormer	3.91	3.97	2.67	**3.52**	***FaceXFormer***	**3.05**	2.66	**4.67**	***FaceXFormer***	**91.83**

Results

Method	LFW	CFP-FP	AgeDB	CALFW	CPLFW	Mean ↑
CosFace [97]	99.81	98.12	98.11	95.76	92.28	96.81
ArcFace [15]	99.83	98.27	98.28	95.45	92.08	96.78
VPL-ArcFace [16]	99.83	99.11	**98.60**	**96.12**	93.45	**97.42**
AdaFace [37]	99.83	**99.11**	98.17	96.02	**93.93**	97.41
SwinFace [71]	**99.87**	98.60	98.15	96.10	93.42	97.22
QFace [88]	×	×	×	×	×	×
Faceptor [72]	99.40	96.34	93.65	94.75	92.27	95.28
FaceXFormer	99.68	96.75	96.35	95.50	91.41	95.94

Ablation

- (Row 1) Without MLP fusion we see a small drop in performance
- (Row 2) Without the proposed Bi-directional CA, we see a huge drop in performance.
- (Row 3) With just CA we see a considerable drop in performance.

The above highlights the impact of the components proposed

MLP Fusion	Self Attn.	Cross Attn.	Bi-dir. CA	HPE (MAE)	Lnd (NME)	Attr. (Acc.)	Age (MAE)
	✓		✓	3.70	4.70	91.21	4.21
✓	✓			17.16	31.49	79.90	16.15
✓	✓	✓		4.63	5.30	89.98	4.84
✓	✓		✓	**3.52**	**4.67**	**91.83**	**4.17**

THANK YOU