FaceCPT: Toward Cross-Modal Facial Representation Learning with Face-Caption Pre-Training

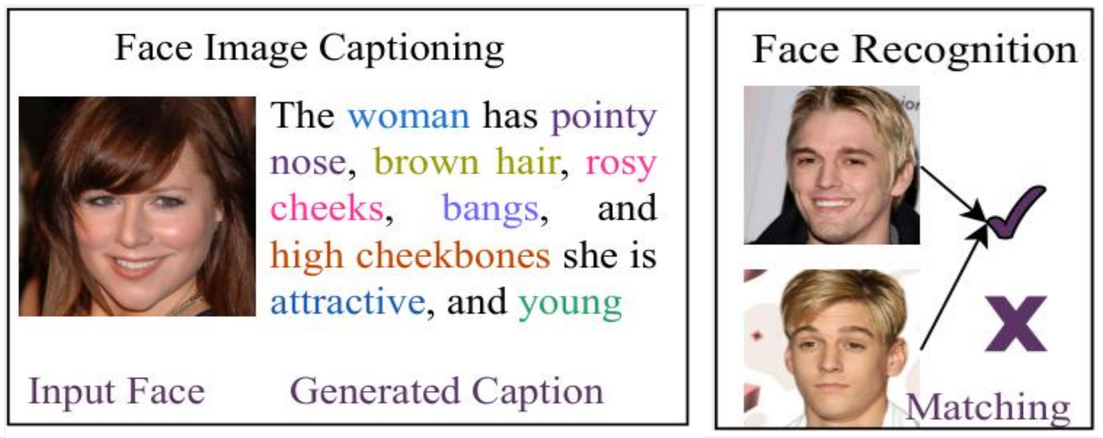
Md Mahedi Hasan, Shoaib Meraj Sami, Nasser Nasrabadi, Jeremy Dawson

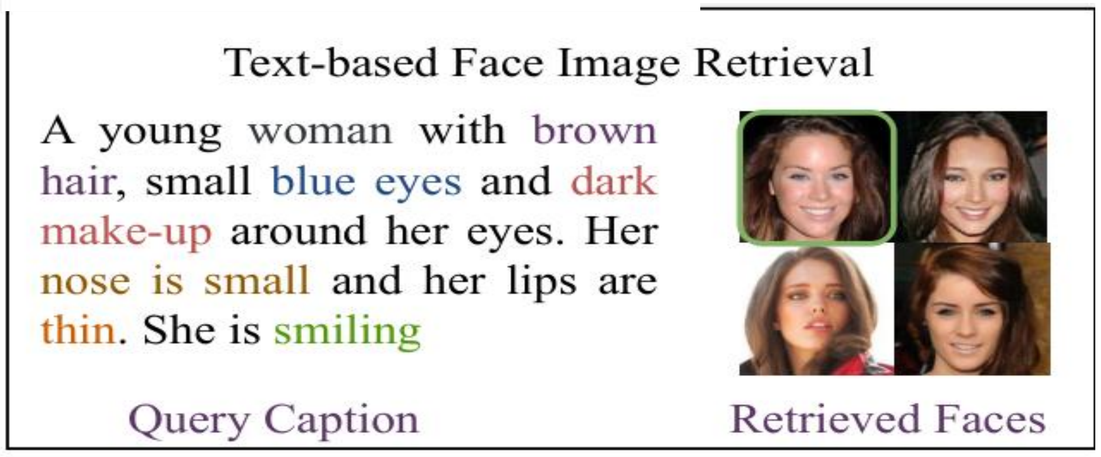
West Virginia University
Morgantown, West Virginia, USA

BMVC 2025

Introduction & Motivation

- Current face recognition (FR) models experience significant performance drops under challenging conditions such as noise, low resolution, and poor image quality.
- Can we learn a **universal facial representation** by pre-training on web-sourced face-caption pairs for low-resolution, surveillance-related tasks?
- These downstream tasks play a critical role in intelligent video surveillance systems, border security, and criminal investigation.
- We introduce a unified model that learns general facial representation through contrastive pre-training on web-sourced data.





Challenges and Limitations

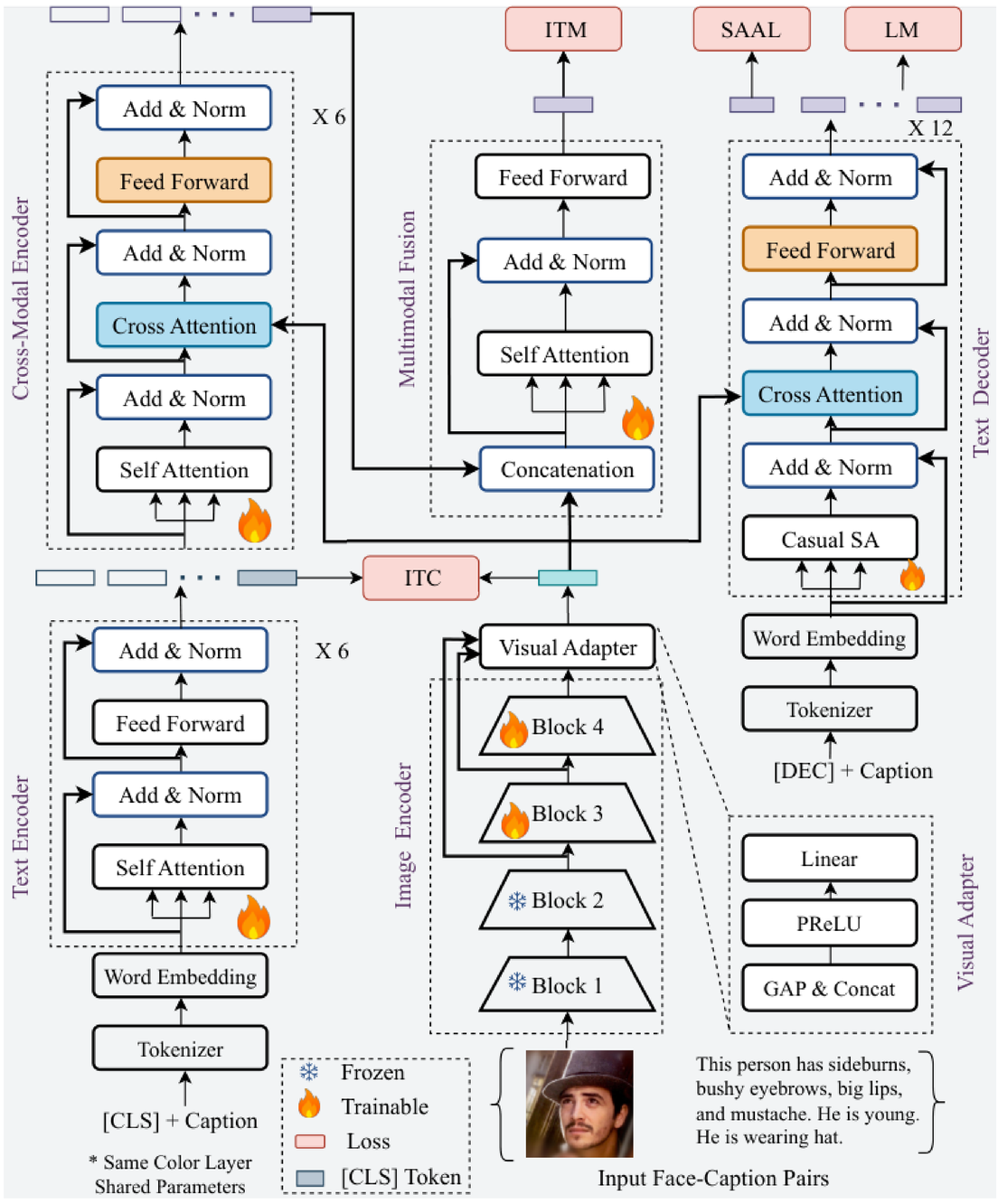
Traditional approaches have

- Relies on supervised training for each specific task
- Requires large, manually annotated datasets
- Demands substantial computational resources
- Performs poorly on low-resolution FR tasks
- Demonstrates limited zero-shot learning capabilities

Issues in Face-Caption datasets

- Subjects often share the same facial attributes
- Captions describe only a few or a single attribute
- Captions may lack crucial details
- Incorrect attributes may be present in the caption
- Captions can vary drastically

Overview of the FaceCPT Framework



Evaluation on Face Analysis

Methods	Res	LFW	CFP-FP	AgeDB-30	CALFW	CPLFW
Cosface [57]	112^2	99.81	**98.12**	**98.11**	**95.76**	92.28
ViT* [13]	224^2	**99.83**	96.19	97.82	95.92	**92.55**
FaRL* [64]	224^2	99.60	96.70	95.55	95.38	92.38
Faceptor* [44]	112^2	99.40	96.34	93.65	94.75	92.27
FaceCPT-R50	112^2	99.68	96.48	96.52	95.48	92.12
FaceCPT-R101	112^2	99.73	96.56	96.64	95.56	92.36

Table: Comparison with FRL methods on benchmark FR evaluation datasets using 1:1 verification.

Methods	RAF-DB [30] (Acc. ↑)	AgeDB [39] (MAE ↓)
FaRL [64]	88.31	5.64
Faceptor [44]	91.26	4.96
SwinFace [43]	90.97	-
QFace [52]	**92.86**	**4.67**
Ours (FaceCPT-R101)	91.42	4.89

Table: Comparison of SOTA FRL methods for age estimation and facial expression recognition

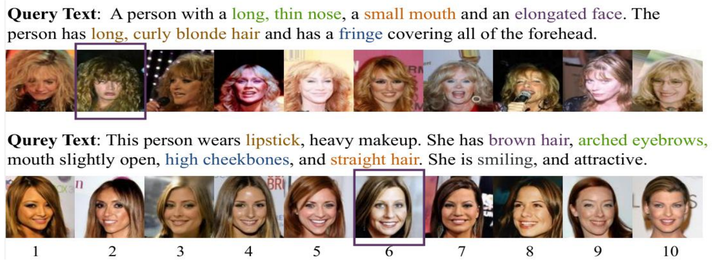
Methods	Training	CelebA [34]	LFWA [34]
DMM† [36]	FSL	91.70	86.56
SlimCNN† [49]		91.24	76.02
AFFAIR† [29]		91.45	86.13
Rot† [15]	SSL	91.69	85.72
FixMatch† [51]		89.78	83.84
VAT† [38]		91.44	84.68
SSPL† [50]		91.77	86.53
MCF* [59]		91.33	-
FRL* [4]		89.65	-
FRA* [14]		**92.20**	-
Faceptor [44]		91.39	-
FaceXFormer [40]		91.79	-
FaRL† [64]		91.88	**86.69**
Ours (FaceCPT-R50)		91.92	86.40

Table: Comparison with various fully supervised (FSL), self-supervised (SSL), and FRL methods for facial attribute recognition using mean accuracy (%)

Evaluation on Face Understanding

Pre-trained Models	Resolution	LFW [20] R@5	LFW [20] R@10	Age-DB [39] R@5	Age-DB [39] R@10
Vision-Language Pre-training					
BLIP-base [27]	384^2	2.79	4.89	2.52	5.56
mPLUG-large [25]	336^2	2.78	7.78	5.69	12.30
BLIP2-OPT$_{2.7B}$	364^2	3.09	7.85	6.76	12.88
Facial Representation Learning					
FaRL [64]	224^2	5.54	9.85	6.97	12.98
FLIP [31]	224^2	5.86	9.90	7.10	13.11
Ours(FaceCPT-R50)	112^2	7.78	10.91	7.43	13.12
Ours(FaceCPT-R101)	112^2	**8.14**	**11.25**	**8.56**	**14.06**

Table: Comparison of zero-shot transferability with baselines



Visualization of FaceCPT's retrieved images for the text-to-face retrieval task.