FaceCrafter: Identity-Conditional Diffusion with Disentangled Control over Facial Pose, Expression, and Emotion

Kazuaki Mishima [1] Antoni Bigata Casademunt [2] Stavros Petridis [2] Maja Pantic [2] Kenji Suzuki [1]

[1] Institute of Science Tokyo, Tokyo, Japan

[2] Imperial College London, London, United Kingdom

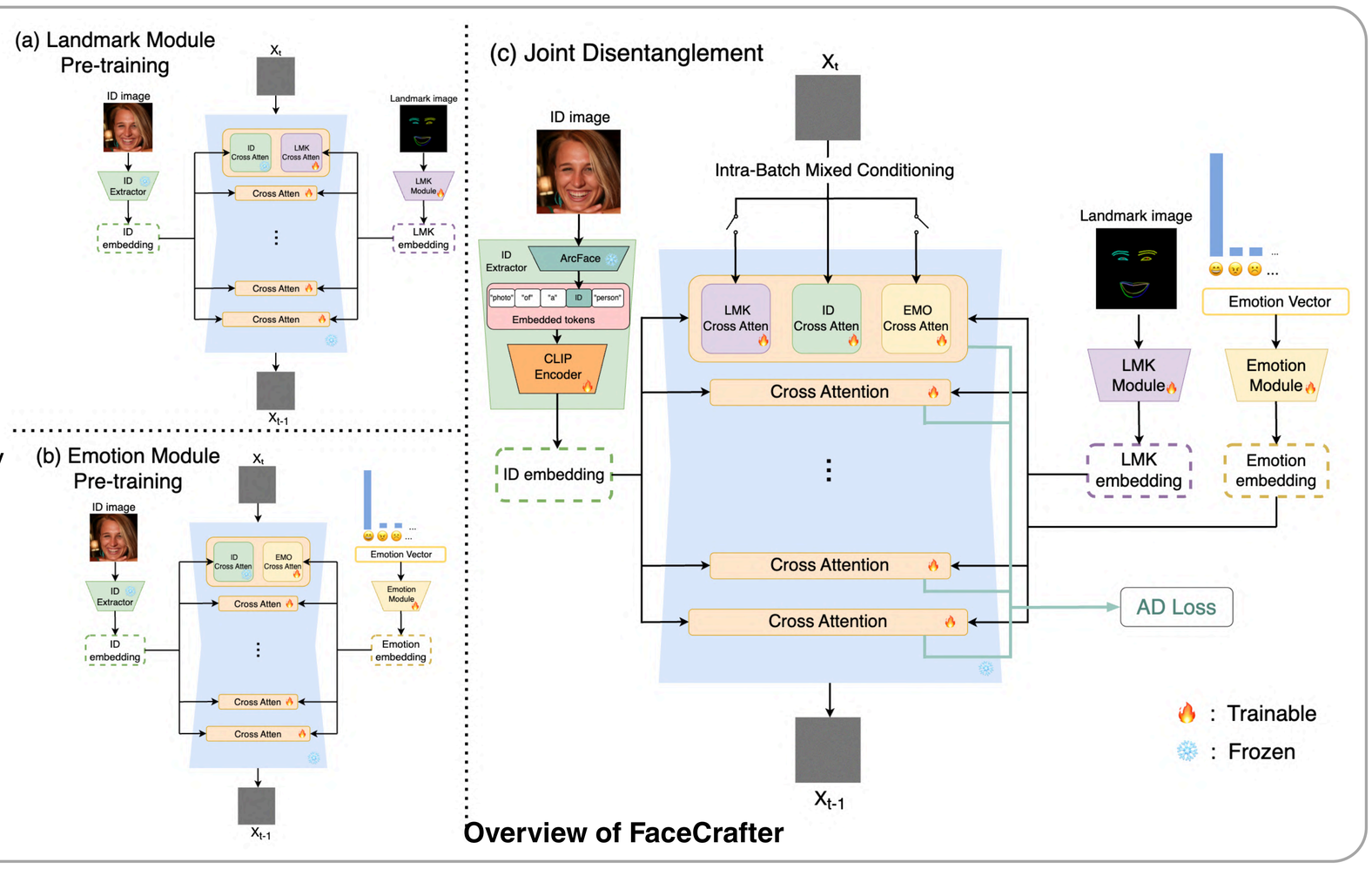
BMVC 2025

Introduction

- Facial images contain both **identity attributes (structure, age, gender)** and **mutable factors (pose, expression, emotion)**.
- Existing datasets lack sufficient diversity in mutable factors, limiting robust identity-aware generation.
- Diffusion models improve realism but still struggle to control pose/expression/emotion while preserving identity.
- FaceCrafter addresses this via **lightweight modules + disentanglement**, enabling independent control and strong ID-only performance, positioning it as a **more robust identity-conditional foundation model**.

Method

- **Base Model:** Arc2Face, an ID-conditional diffusion model built on Stable Diffusion, using ArcFace features as identity embeddings.
- **Control Modules:**
  - Landmark → pose/expression (MediaPipe)
  - Emotion → valence-arousal + 8-class vectors (EmoNet)
  - Lightweight modules added to cross-attention using IP-Adapter style
- **Pre-training Stage (a, b):** Pre-train modules separately
- **Joint Disentanglement Training (c):**
  - Integrates pre-trained Landmark and Emotion control modules with the ID module.
  - Uses Intra-Batch Mixed Conditioning: mixes combinations of ID, pose, and emotion within each batch to encourage disentanglement.
  - Applies Attention Disentanglement (AD) Loss to explicitly separate features of different control conditions.



Overview of FaceCrafter

Results

- **Quantitative Results:**
  - Outperforms existing methods in pose, expression, and emotion control using only 5% additional parameters.
  - Emotion accuracy improved by 25% over baselines; +9% gain with Emotion Control Module.
  - Joint Disentanglement Stage improves overall control compared to Pre-training Stage.
  - Strong generative ability and diversity even under ID-only conditioning.
- **Ablation Study of AD Loss:**
  - AD Loss significantly boosts diversity, especially for expressions.
  - Slight FID drop without AD Loss indicates minor trade-off with image quality.
- **Qualitative Results:**
  - More accurate control of pose and expression than baselines.
  - LMK + EMO control improves emotion generation.
  - Under ID-only conditioning, generates more diverse poses, expressions, and emotions than Arc2Face.

Method	Exp-RMSE ↓	Pose-RMSE ↓	V-RMSE ↓	A-RMSE ↓	Dist-RMSE ↓	Class-ACC(%) ↑	ID-Sim ↑	FID ↓	#Extra-Params ↓
Arc2Face+ControlNet	8.12	0.130	0.360	0.247	0.86	35.1	**0.76**	44.64	361M
CapHuman	7.82	0.166	0.428	0.224	0.87	34.6	0.71	88.80	361M
FaceCrafter-Pre-trained(LMK)	7.11	0.124	0.267	0.221	0.71	46.5	0.75	46.85	**11.7M**
FaceCrafter-Pre-trained(LMK+EMO)	7.12	0.128	0.218	0.205	0.59	56.5	0.72	53.25	18M
FaceCrafter(LMK)	**6.78**	**0.119**	0.250	0.203	0.63	52.4	0.71	**38.80**	**11.7M**
FaceCrafter(LMK+EMO)	6.81	0.126	**0.217**	**0.184**	**0.52**	**61.2**	0.69	41.54	18M

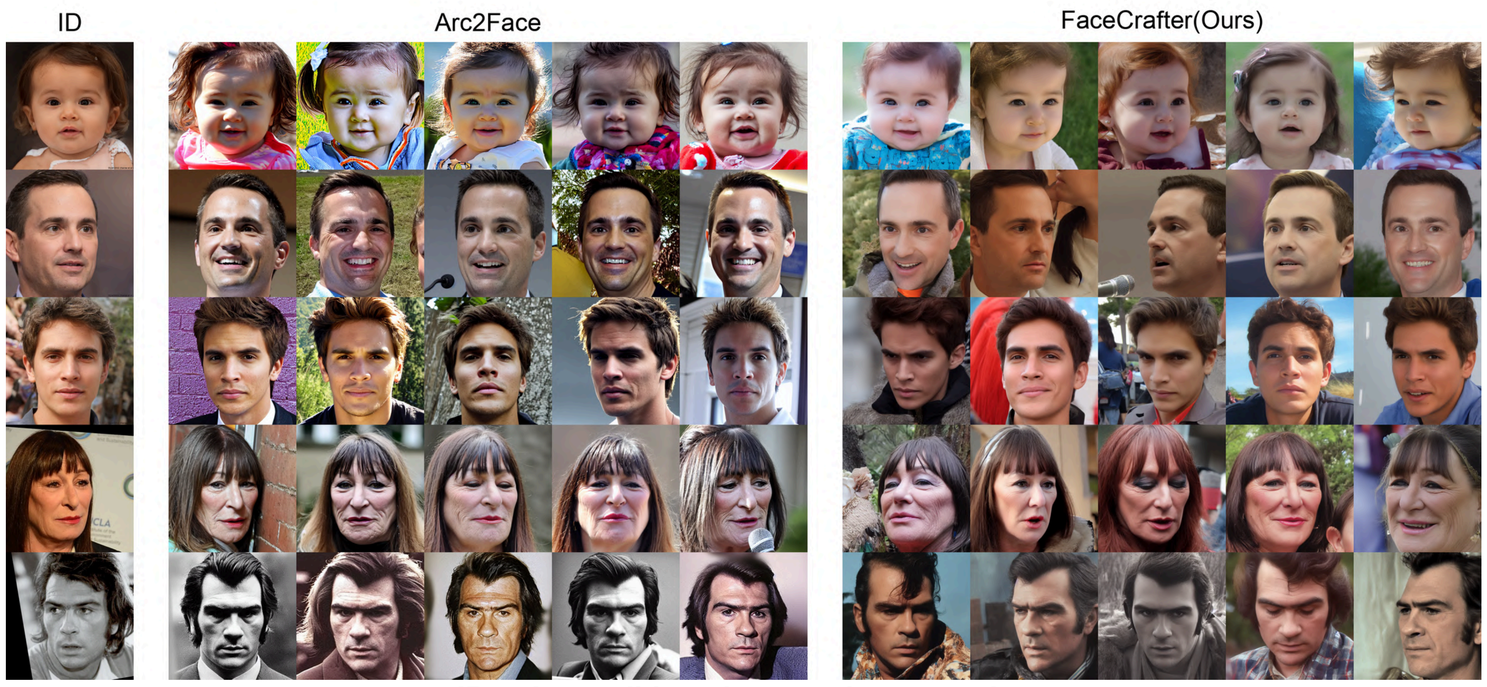
Control performance comparison

Method	LPIPS ↑		Exp. (ℓ_2) ↑		Pose (ℓ_2) ↑		FID ↓	
	Synth	AgeDB	Synth	AgeDB	Synth	AgeDB	Synth	AgeDB
FastComposer	0.389	0.487	3.597	4.678	0.163	0.225	13.517	31.736
PhotoMaker	0.410	0.424	3.920	4.283	0.167	0.165	13.295	8.410
InstantID	0.386	0.437	3.733	4.569	0.059	0.082	22.859	18.598
IPA-FaceID (SDXL)	0.402	0.462	4.648	5.812	0.181	0.197	7.104	24.105
IPA-FaceID-Plus	0.320	0.384	2.706	3.518	0.150	0.194	14.880	11.817
IPA-FaceID-Plusv2	0.356	0.429	3.147	4.092	0.185	0.236	9.752	10.798
Arc2Face	0.506	0.508	6.375	5.966	0.317	0.273	5.673	6.628
FaceCrafter	**0.552**	**0.589**	**6.971**	**6.930**	**0.429**	**0.424**	**2.112**	**6.501**

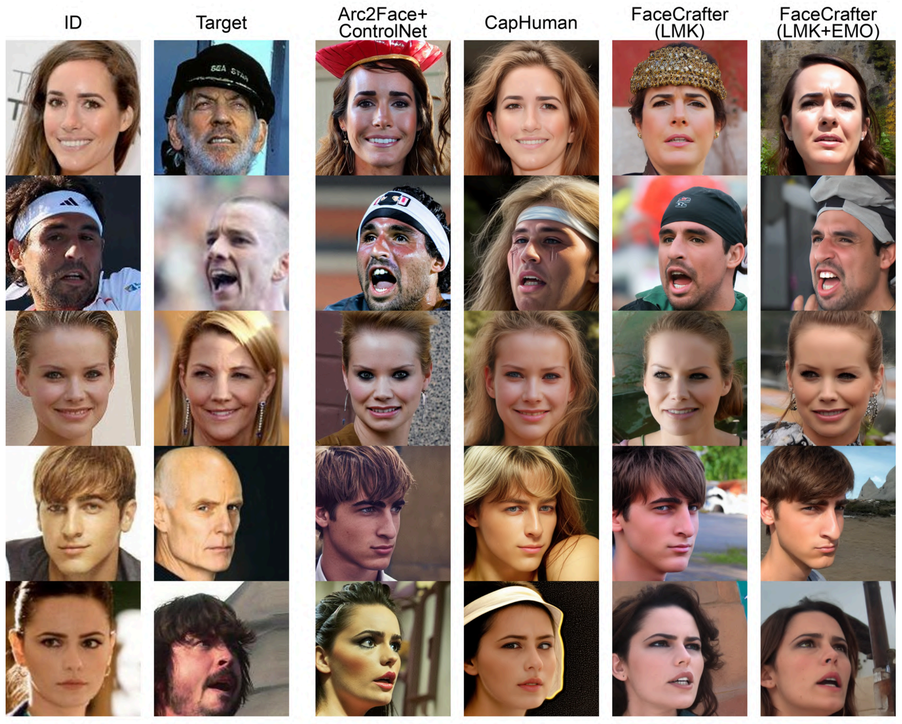
Diversity performance comparison

Method	LPIPS ↑		Exp. (ℓ_2) ↑		Pose (ℓ_2) ↑		FID ↓	
	Synth	AgeDB	Synth	AgeDB	Synth	AgeDB	Synth	AgeDB
w/o AD Loss	0.538	0.564	6.540	6.430	0.411	0.397	2.140	**5.381**
Ours	**0.552**	**0.589**	**6.971**	**6.930**	**0.429**	**0.424**	**2.112**	6.501

Ablation study of AD Loss



ID-only conditioned generation examples



Qualitative comparison under pose & expression control