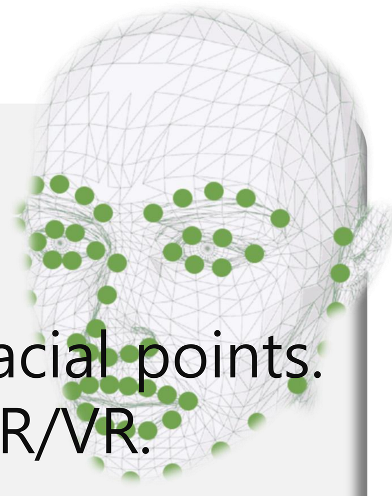
TOWARDS MULTI-DOMAIN FACE LANDMARK DETECTION WITH SYNTHETIC DATA FROM DIFFUSION MODEL

*Yuanming Li, Gwantae Kim, Jeong-gi Kwak, Bon-hwa Ku, Hanseok Ko**

School of Electrical Engineering, Korea University

Brand Engagement Network

Background

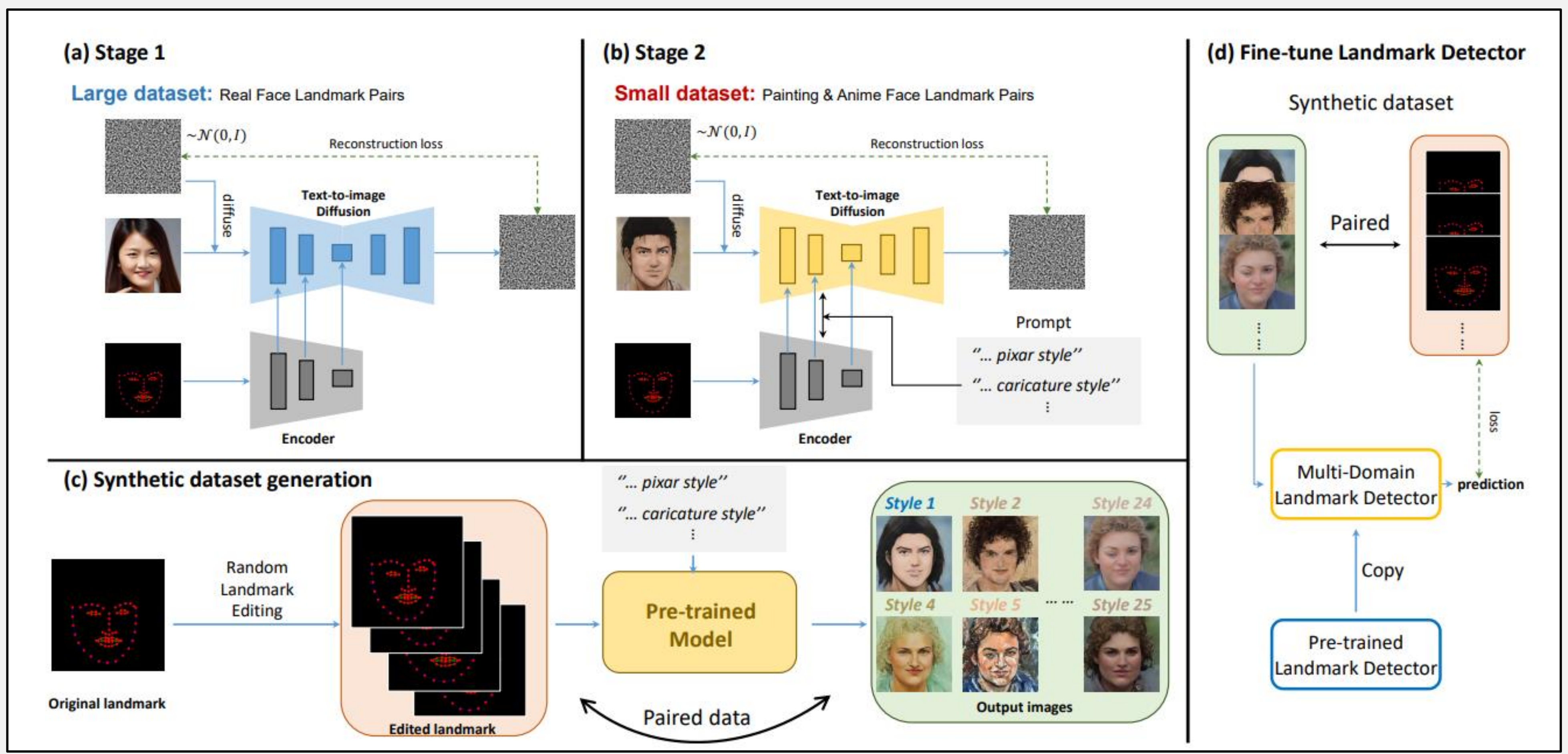


- **Facial Landmark Detection:** Identifies key facial points. Critical for 3D reconstruction, recognition, AR/VR.
- **Challenge:** Extending accuracy to art, cartoons, caricatures. Limited by scarce diverse data.
- **Traditional Methods Limitations:** Rely on warping, style translation. Struggle with significant domain gaps.

Contributions

- We enhance **multi-domain face landmark detection** using synthetic data from a diffusion model.
- A **two-stage training method for synthetic dataset** generation : initially leveraging a large real-face dataset with ControlNet, then fine-tuning on a smaller, diverse domain dataset.
- We generate a **multi-domain face landmark dataset** across 25 styles, comprising 400 annotated images per style.

Method



Stage 1: Initial Training

- Utilized a large dataset of real-face and landmark pairs.
- Trained ControlNet to generate face images conditioned on facial landmarks.

Stage 2: Domain Adaptation

- Fine-tuned ControlNet using a small, diverse domain face dataset.
- Adjusted facial landmarks and styles through text prompts.

Synthetic Dataset Generation

- Edited random landmark attributes to create a variety of styles.
- Generated 400 images for each of 25 styles, resulting in a 10,000-image dataset.

Fine-tuning Landmark Detector

- Employed the synthetic dataset to fine-tune a pre-trained face landmark detector.
- Enhanced model's performance on the ArtFace and Caricature datasets.

Experimental Setup

- Implemented based on the Stable Diffusion model with a 1.4 billion parameter T2I model.
- Training was done on the FFHQ dataset for 200k steps and on a small multi-domain dataset for 100k steps.
- Utilized DDIM sampler with classifier-free guidance for landmark-guided face generation.
- The entire training process was efficient, requiring only a single NVIDIA RTX Titan GPU and was completed within a day, highlighting the model's practical applicability for quick deployment and testing.
- Evaluated using NME for landmark accuracy, FR for error instances, and AUC for overall performance.

Results and Discussion

Synthesis images

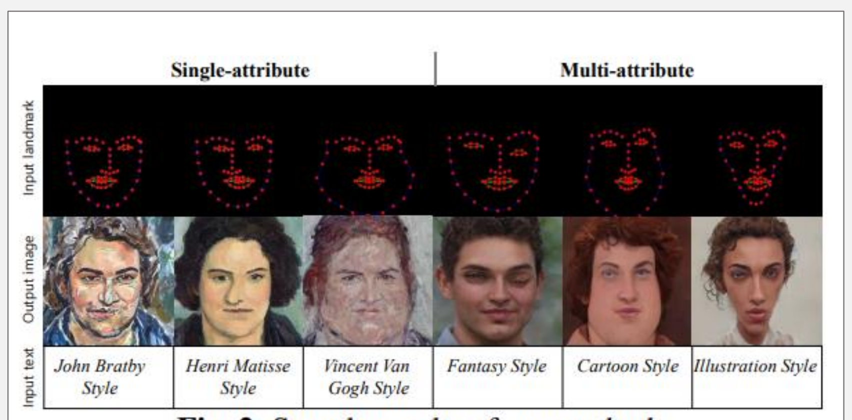


Fig. 2. Sample results of our method.

- Displays various generated images demonstrating the model's capability to accurately align with edited facial landmarks across different styles, from single-attribute changes to complex, multi-attribute transformations.

Quantitative comparison

Table 1. Quantitative comparison with evaluated baselines.

	ArtFace			CariFace		
Metric	NME	$FR_{10\%}$	$AUC_{10\%}$	NME	$FR_{10\%}$	$AUC_{10\%}$
Ours	4.64	2.26	0.5548	5.54	6.29	0.4838
foa	4.69	3.75	0.5388	8.26	22.31	0.2997
ArtFace	6.50	10.62	0.4573	12.04	44.41	0.1476
CariFace	-	-	-	4.54	0.71	0.5477
STAR	6.20	13.21	0.5142	7.16	13.73	0.3875

Qualitative comparision

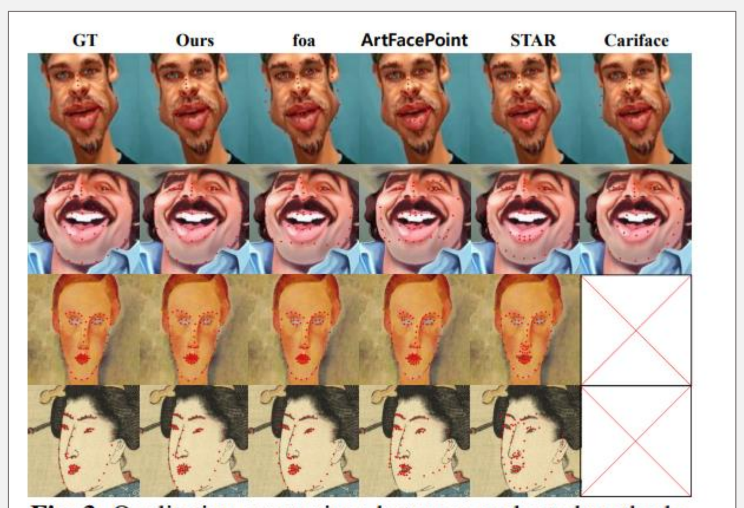


Fig. 3. Qualitative comparison between evaluated methods.

- Offers a qualitative comparison between the proposed method and existing techniques, showing superior alignment and detail capture in generated facial landmarks.

Ablation study for image synthesis

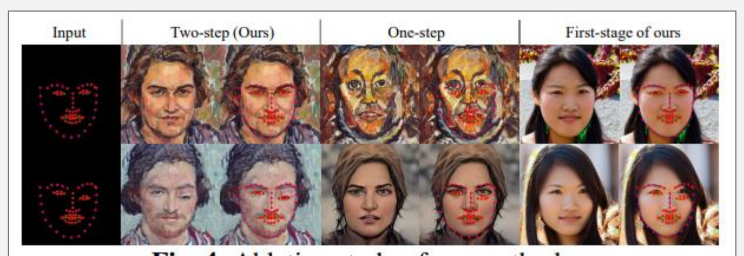


Fig. 4. Ablation study of our method.

- Illustrates the results of an ablation study, evidencing the effectiveness of the two-stage training approach in maintaining alignment between generated images and input landmarks, even with exaggerated modifications.

Conclusions

- Introduced a novel two-stage approach for generating synthetic, multi-domain facial landmark data.
- The approach effectively handles exaggerated landmarks and diverse styles, validated by improved accuracy in multi-domain landmark detection.