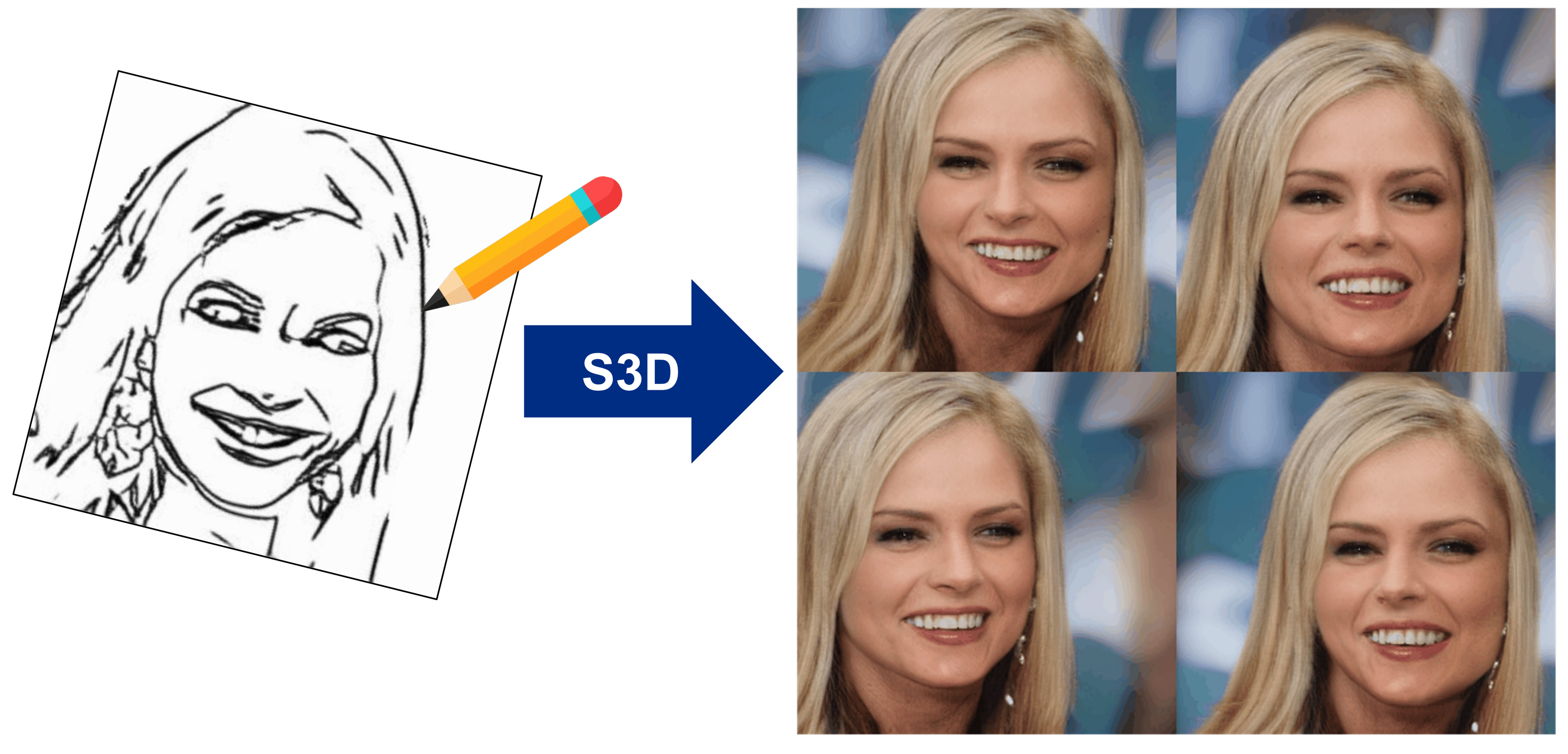
S3D: Sketch-Driven 3D Model Generation

Hail Song[1*] Wonsik Shin[2*] Naeun Lee[2] Soomin Chung[2] Nojun Kwak[2†] Woontack Woo[1†]

[1]KAIST [2]Seoul National University *Equal contribution.

Paper Code

Introduction



Input Sketch Image

Generated face images with novel views

Motivation : From 2D sketch to 3D model

- **Challenge**: Traditional 3D generation requires multi-view images or complex setups
- **Problem**: Sketches are highly abstract and sparse – no existing method can generate 3D human faces from sketches
- **Opportunity**: Sketches are the most accessible visual input for users

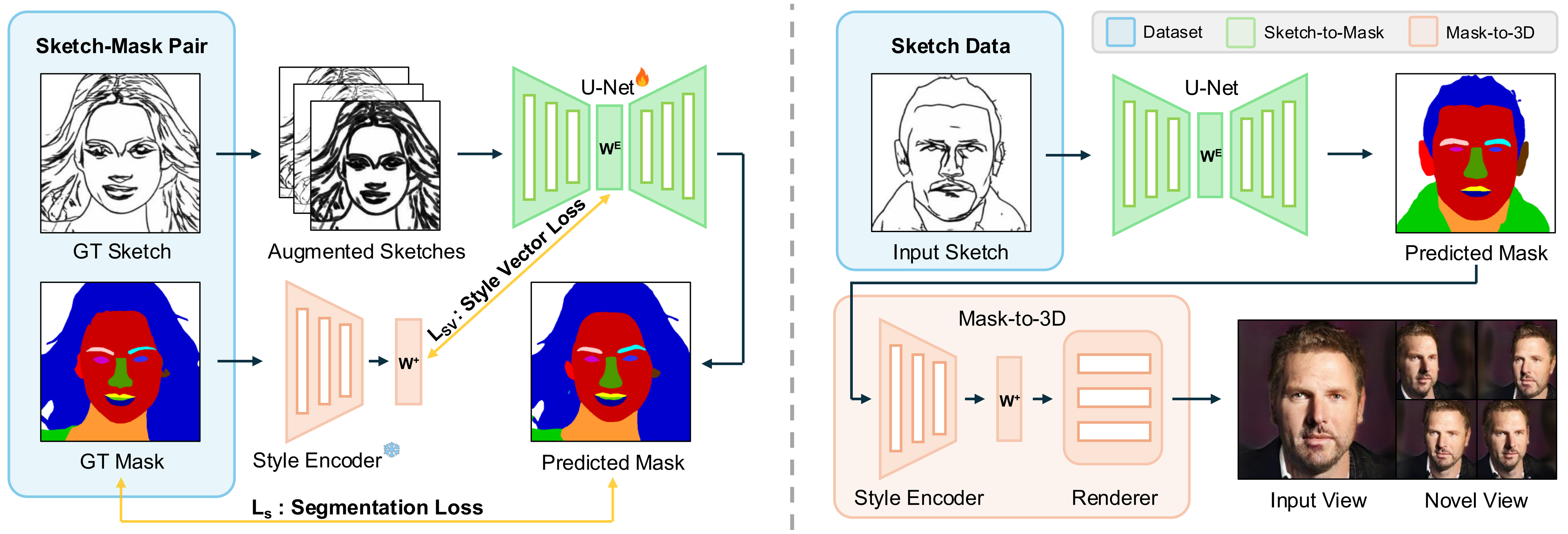
Why Sketches Matter?

- **Accessibility**: Everyone can draw simple sketches!
- **Applications**: Forensic reconstruction, Avatar creation, Rapid prototyping of 3D contents

Key Contributions

- **First end-to-end sketch-to-3D** human face generation pipeline
- **Novel Style Vector Loss** bridging sketch-3D domain gap
- **Potential real-world impact** in forensics, film, gaming, and virtual reality

Method

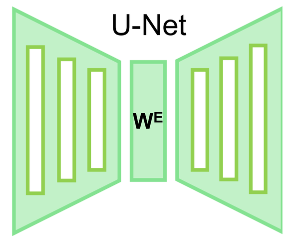


(a) Training Process (b) Inference Process

S3D: Two-Stage Pipeline

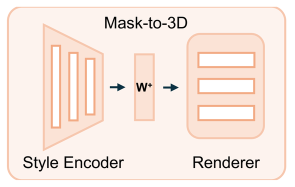
1. Sketch-to-Mask Module

- **U-Net architecture** converts 512×512 sketches to segmentation masks
- 7 encoder-decoder pairs with bottleneck matching 3D module dimensions (7x512)



2. Mask-to-3D Module

- **Tri-plane-based network** (pix2pix3D) generates 3D models from masks
- **Novel view synthesis** with volumetric rendering



Learning Objective

$$L_{total} = \underbrace{L_{SV}}_{\text{Style Vector loss}} + \underbrace{L_{CE} + L_{Dice}}_{\text{Segmentation loss}}$$

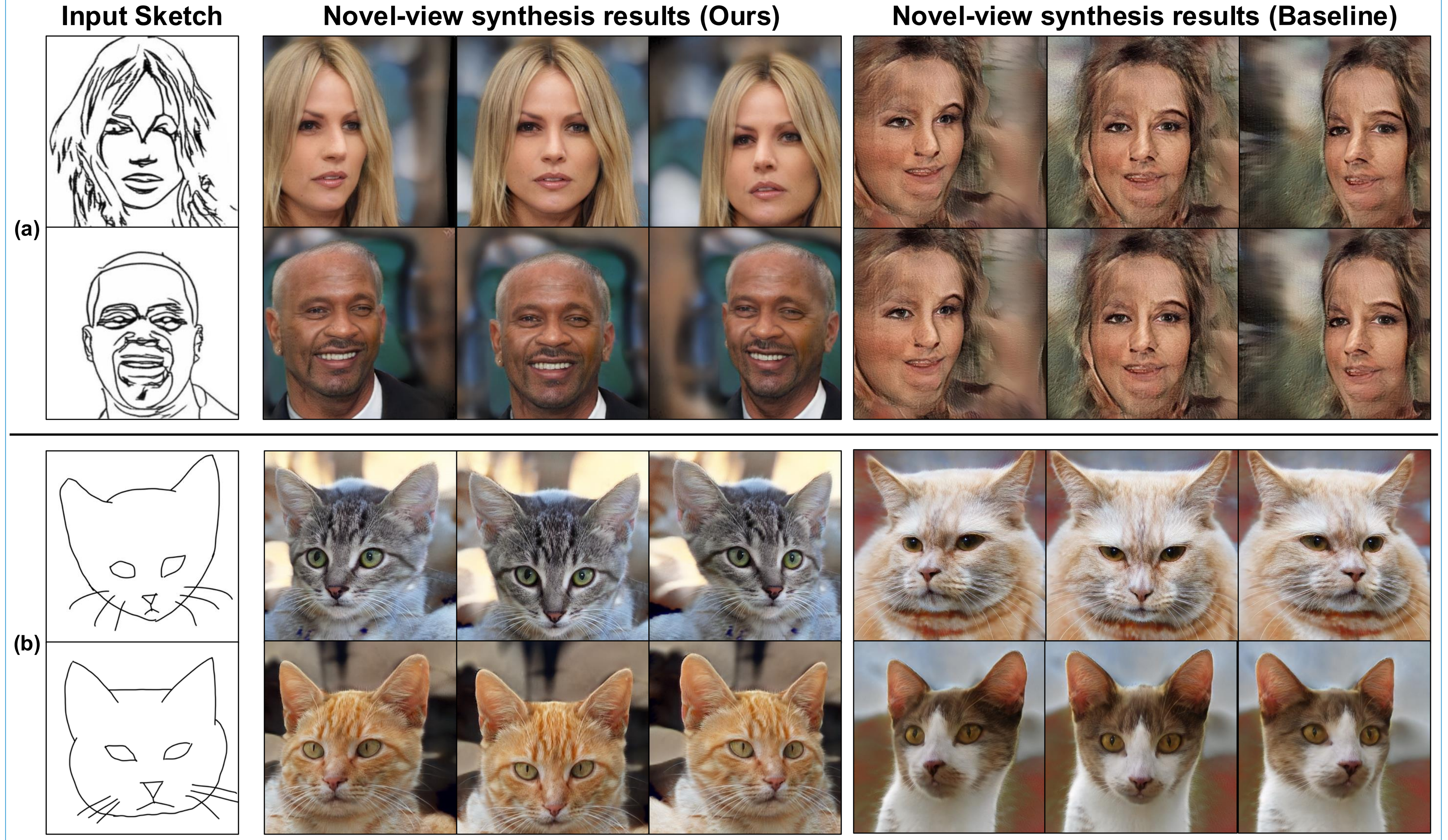
- $L_{SV} = \|w^+ - w^E\|_2^2$
  - Aligning U-Net bottleneck features (w^E) with 3D module's style vectors (w^+) to bridge domain gap
- Cross-entropy (L_{CE}) for pixel-wise accuracy + Dice (L_{Dice}) for shape consistency

Datasets

- Human: Multi-Modal-CelebA-HQ dataset (sketches generated via Photocopy filter + sketch simplification)
- Cat: AFHQ with sketches extracted using PidiNet edge detection

Experimental Results

Qualitative Results

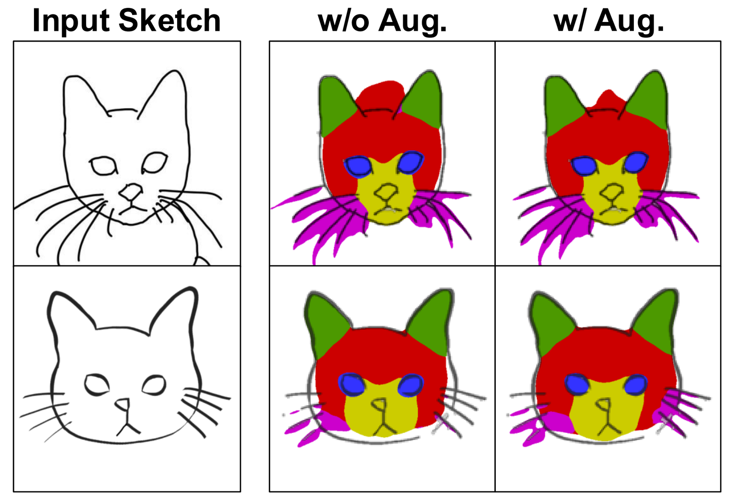


Quantitative Results

Dataset	Method	FID ↓	KID ↓	FVV ↓
CelebA	pix2pix3D	232.81	0.3142	0.20
	S3D (Ours)	**21.71**	**0.0065**	**0.18**
AFHQ	pix2pix3D	27.36	0.0054	–
	S3D (Ours)	**23.86**	**0.0047**	–

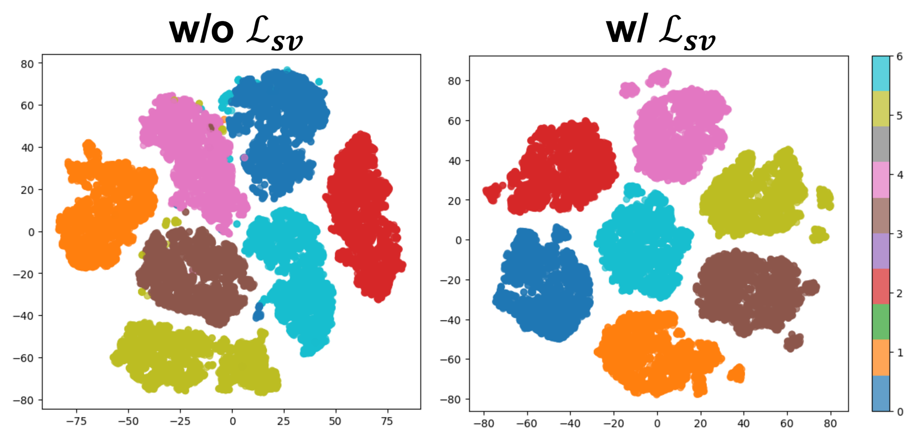
Augmentation Strategy

- 25% dilation, 25% erosion



Ablation Study

t-SNE visualization of style embeddings



Quantitative results

Dataset	Method	mIoU ↑	mAP ↑
CelebA	w/o L_{SV}	0.692	0.793
	w/ L_{SV}	**0.698**	**0.823**
AFHQ	w/o L_{SV}	0.804	0.884
	w/ L_{SV}	**0.807**	**0.890**