Shape-Texture based Vision Transformer for Face Attribute Recognition with Limited Labeled Data

Kunyan LI, Jie ZHANG, Jinwu WEI, Xiaojie WANG, Shiguang SHAN

Frontiers of Computer Science, DOI: 10.1007/s11704-026-40839-5

2 Problem Statement and Original Ideas

⚠ Current Problem

Data Dependency: Face attribute recognition requires massive labeled data for effective model training, which is costly and time-consuming to acquire.

Limitation of Existing Methods: Methods like SSPL (Spatial-Semantic Patch Learning) fail to fully exploit shape and texture characteristics for recognizing attributes such as "Chubby" and "wrinkles."

Challenge: How to leverage unlabeled data effectively while learning intrinsic facial representations that capture both structural (shape) and appearance (texture) information?

Original Ideas

Core Insight: Exploit the natural correlation between facial features and attributes—3D shape (e.g., nose structure, face contour) and texture (e.g., wrinkles, skin tone) provide complementary cues for attribute recognition.

Dual-Tokenizer Design: Introduce separate tokenizers for shape and texture features, enabling disentangled learning of facial structure and appearance.

Efficiency Optimization: Use patch embeddings as inputs and dual shape-texture tokens as outputs, avoiding the high computational cost of recovering raw pixels from masked patches.

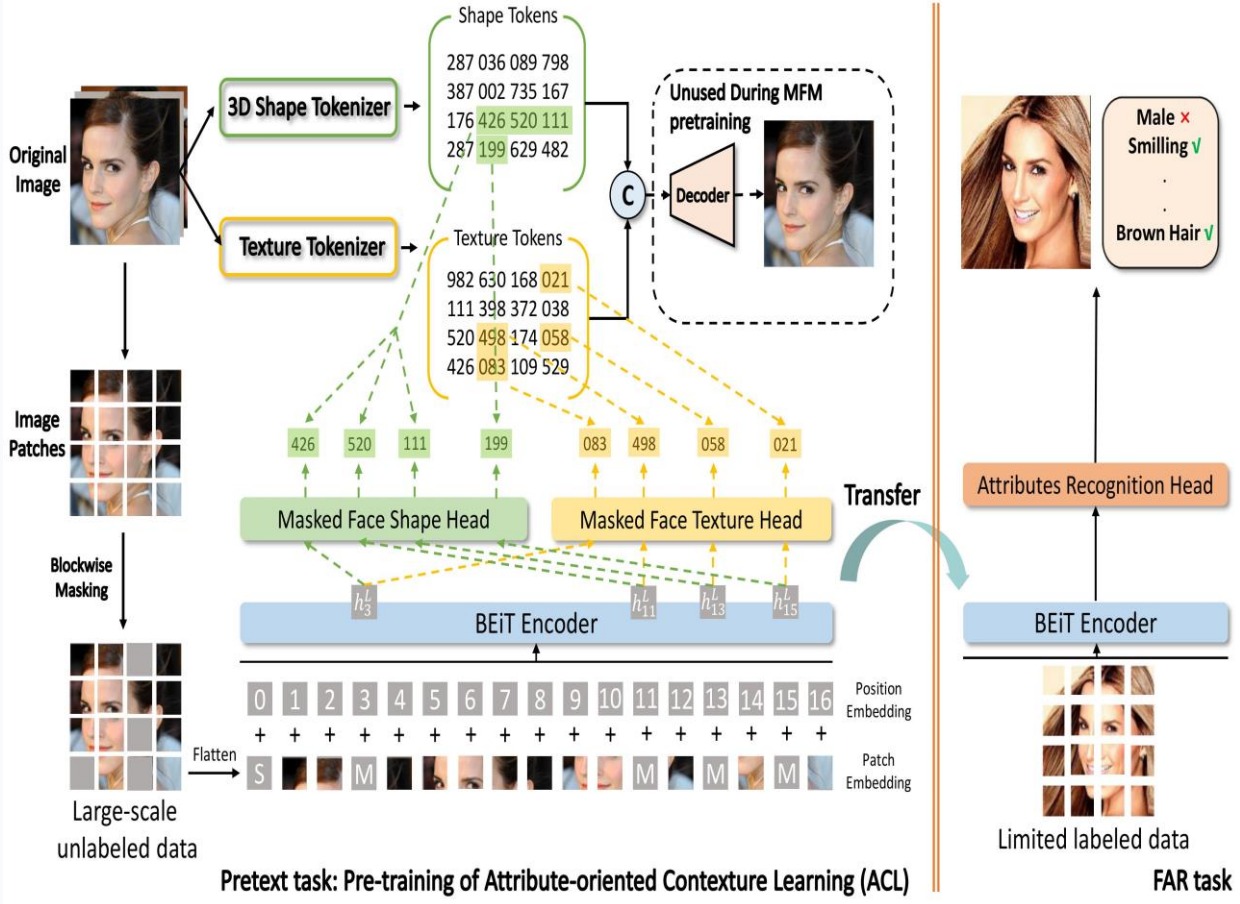

Figure 1: Overview of STViT framework with FTL and ACL pretext tasks

3 Main Experimental Results and Conclusions

Datasets and Evaluation

CelebA: 200,000 images with 40 facial attributes (gender, age, expressions, etc.)

LFWA: 13,000 images with same 40 attributes

Evaluation Protocol: Test with varying proportions of labeled training data (0.2% to 100%)

Key Results

CelebA (0.2% labels) +1.49% improvement over SSPL (86.67%)	**88.16 %**
LFWA (5% labels) +3.79% improvement over SSPL (78.68%)	**82.47 %**

Consistent Superiority: STViT outperforms all leading methods (DeepCluster, JigsawPuzzle, Rot, SSPL) across all label-scarce scenarios, demonstrating the efficacy of dual shape-texture tokenizers.

Conclusions

STViT achieves state-of-the-art performance for face attribute recognition with limited labels by learning intrinsic facial representations through shape-texture disentanglement. The dual-tokenizer approach effectively harnesses facial inherent information, showing overwhelming superiority in label-scarce scenarios. Future work will focus on improving interpretability of shape and texture contributions.

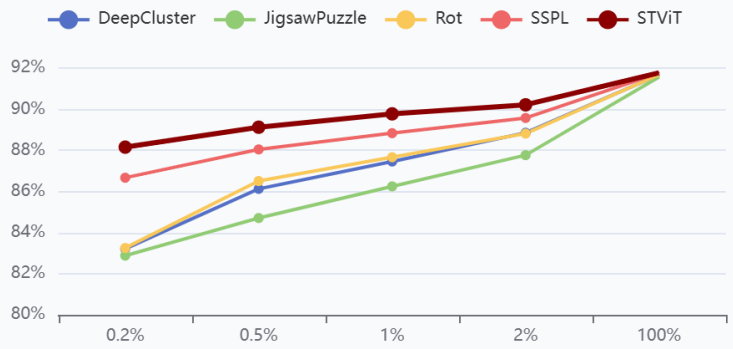

Figure 2: Accuracy comparison across different label proportions

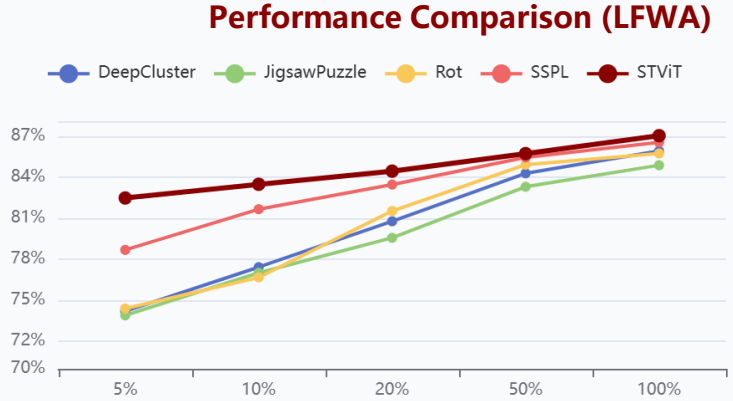

Figure 3: Accuracy comparison across different label proportions