テキストタグを併用した画像の品質評価モデルの提案

杉浦敦之，山内悠嗣（中部大学）

1. 研究背景，目的

・スマートフォンや画像生成AIの普及に伴い，
デジタル画像が急増

→ **画像の品質を自動的に評価する技術の需要が高まる**

・色，構図などの視覚的特徴のみでは
人間の感性に近い品質評価を捉えることが困難

・画像内容を記述したテキストタグを併用
・視覚情報，言語情報を考慮した品質評価を提案

2. 提案手法

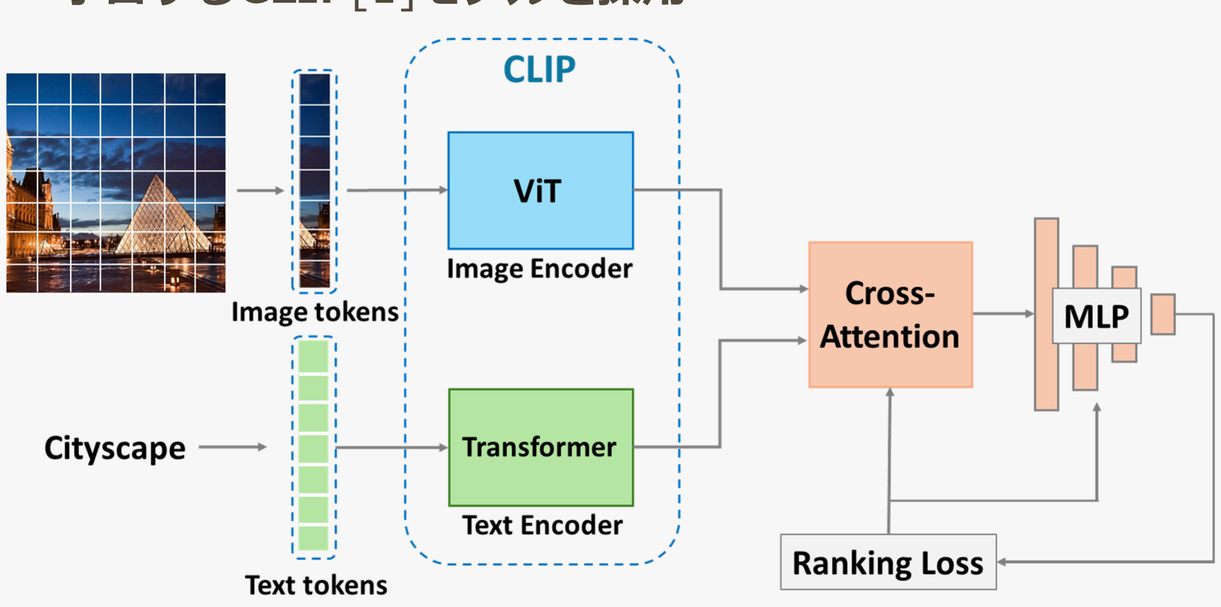
マルチモーダル学習＋ランキング学習

マルチモーダル学習

・画像品質，画像とテキストの意味的一致性を同時に評価

→ 単一の情報だけでは捉えきれない，より複雑な情報を理解

・基盤として，画像とテキストをマルチモーダルに学習するCLIP[1]モデルを採用



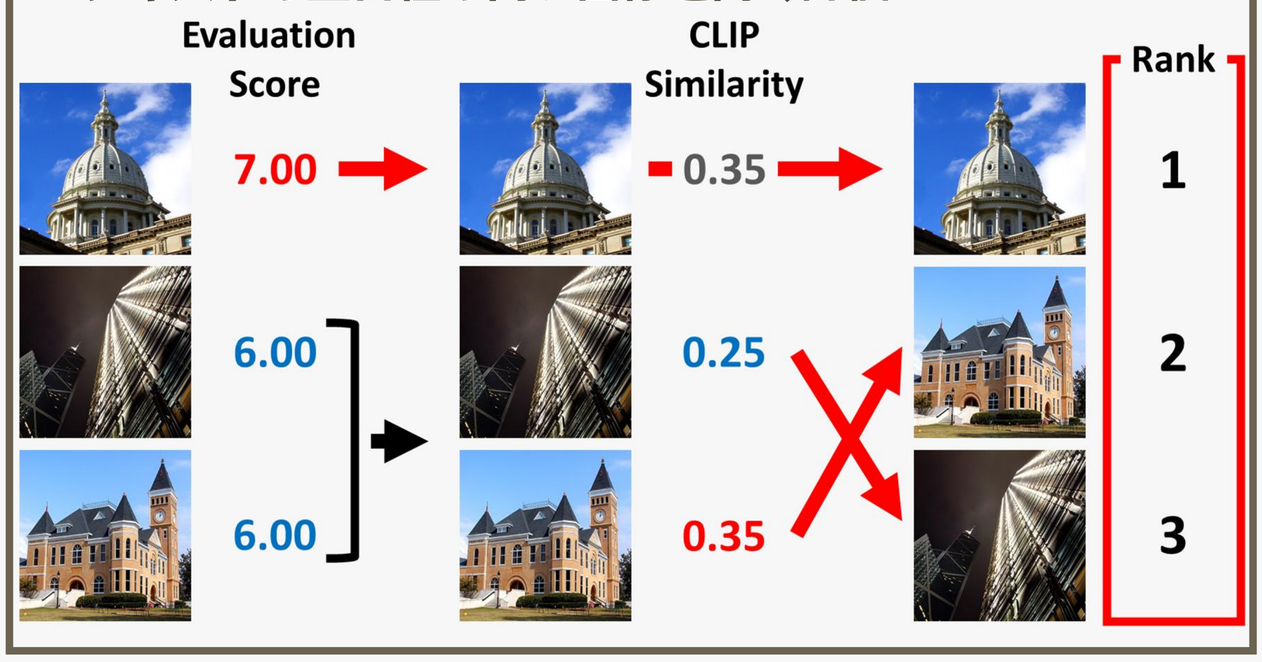
提案手法の流れ

1. **画像**と**画像に対応したテキストタグ**をCLIPモデルに入力
2. CLIPモデルで各モーダルの特徴量を抽出
3. **Cross-Attention**で各特徴量を統合
4. 統合後の特徴量を**MLP**に入力，品質スコアを推定

ランキング学習

・２つの指標を用いたランキング情報を付与
画像の品質スコア＋CLIP類似度

・順位関係に基づいた学習により，品質に加え，テキストの整合性の高い画像を高く評価



3. 実験結果

各手法における画像の品質評価精度を比較

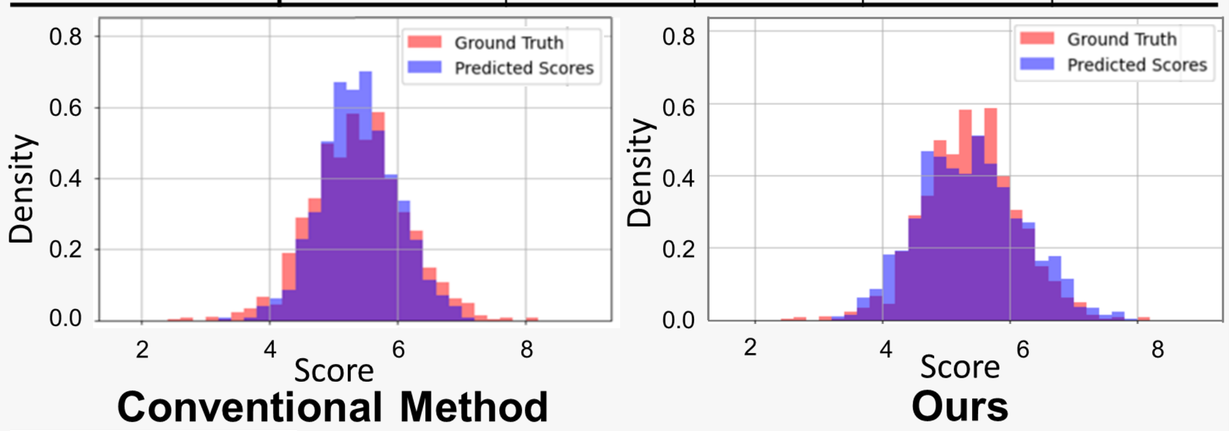
・従来手法[2]
・CNNベースであり，視覚特徴を基に品質スコアを回帰予測するモデル

・提案手法

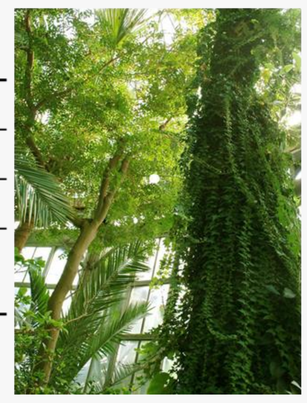
AVAデータセット[3]での実験結果

・人間の感覚を表す品質スコアを正解スコアとして，
品質評価精度と予測分布を比較

Method	Error ↓	LCC ↑ (mean)	SRCC ↑ (mean)	KL Divergence ↓	EMD ↓
Conventional	0.478(±0.36)	0.596	0.556	0.295	0.106
Ours	0.482(±0.15)	0.684	0.620	0.138	0.094



TextTag	Birds
Conventional	5.36
Ours	6.62
Human Evaluation	6.84

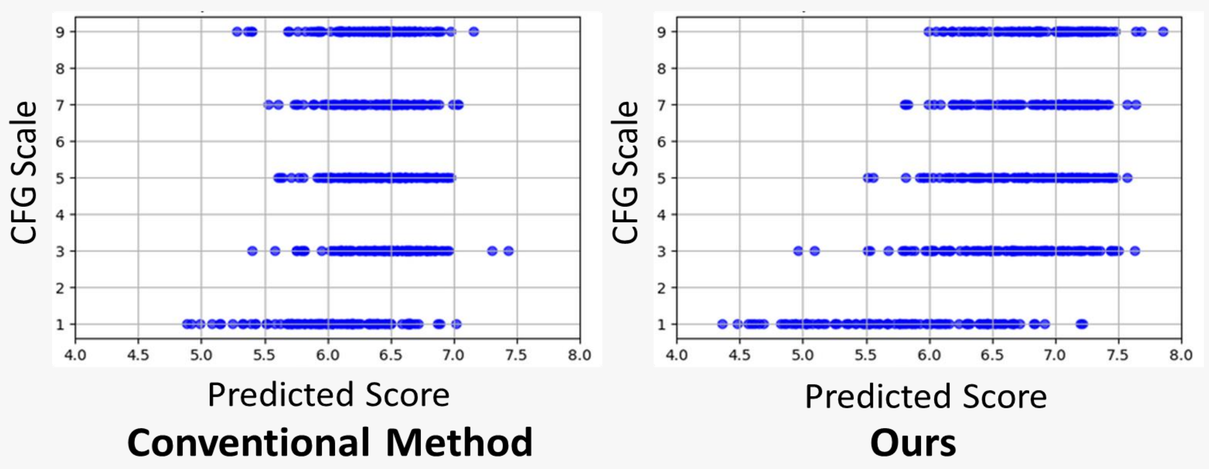


TextTag	Nature
Conventional	4.98
Ours	5.91
Human Evaluation	5.87

より人間の評価に近い予測が可能

生成画像データセットでの実験結果

・CFG Scale：プロンプトに対する画像生成の忠実度を調整するパラメータ



従来手法より豊かな評価が可能

4. 今後の展望，参考文献

・人間の評価に近い画像の品質評価モデルの構築
・画像に対する品質評価の説明性の出力

[1] Radford, *et al.* "Learning transferable visual models from natural language supervision." ICML, 2021.
[2] Talebi, *et al.* "NIMA: Neural image assessment." IEEE transactions on image processing, 2018.
[3] Murray, *et al.* and Florent Perronnin. "AVA: A large-scale database for aesthetic visual analysis." CVPR, 2012.