ABCI 3.0開発加速利用 （2024年度）成果概要（公開用）

課題名：

高解像度人間基盤モデルの構築と学習

実施時期：2025.1-2025.3
所属機関名：産業技術総合研究所
代表者氏名：吉安祐介

成果概要：

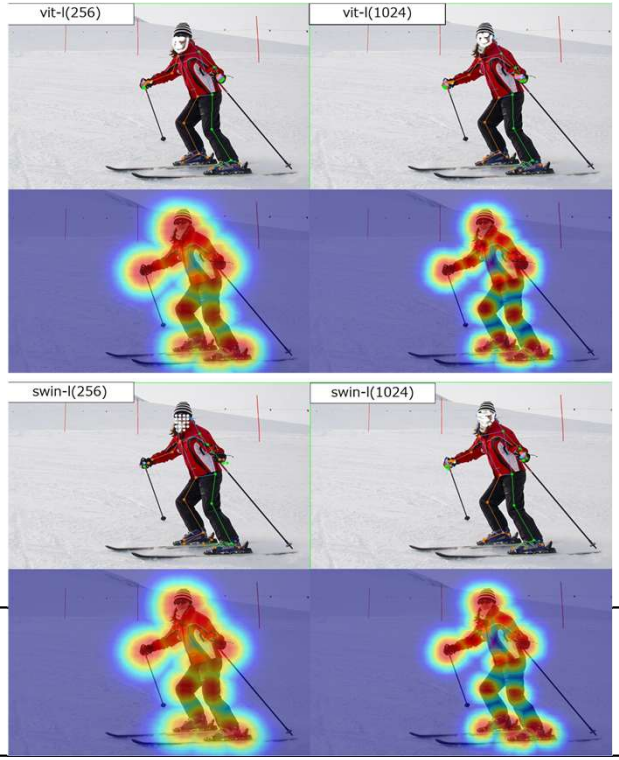
高解像度画像（1024x768）を用いて、手や顔を含む全身のポーズ推定モデルを学習・構築することで高い認識性能を獲得することができた。とりわけ、12種類のデータセットを用いてViT-Lモデルを学習することで、全身ポーズ推定モデルの評価を行うCOCOWholebody ベンチマークにおいてトップ性能を得ることができた。また、CNNベースのRTMPoseやSSMベースのViMとVmambaも高解像度画像を用いた学習を行った結果、低解像度画像（256x192）を入力とするモデルよりも大きく性能が向上した。

成果のポイント：

高解像度人間画像基盤モデルSapiens[ECCV2024]は、三億枚の人間画像データセットを用いて億単位のパラメータを持つViTモデルを事前学習することで、手指の姿勢や顔のランドマークを含む全身ポーズ推定など、人間の身体に関する詳細な認識を高性能に実現できる。しかしながら、計算コストが大きくリアルタイムアプリケーションに利用することが難しいなどの課題がある。本研究開発では、リアルタイム実行可能でより軽量な高解像度人間基盤モデルの構築を目指す。

高解像度画像（1024x768）を含む12種類のデータセットを用いて、ViT-Lに基づく全身のポーズ推定モデル学習することで、COCOWholebody データセットを用いた全身ポーズ推定ベンチマークにおいて、Sapiensの同サイズモデルの性能を大きく上回った。また、右表のようにCNNベースのRTMPoseやSSMベースのViMとVmambaについても高解像度画像を用いた学習を行った結果、低解像度画像（256x192）を入力とするモデルよりも大きな性能の向上が見られた。実際、高解像度画像を入力することで右図のように顔や手の指のキーポイント推定誤差を低減できる。今後は、学習効率や推論スピードの向上を図るため、トークン削減技術などのモデル効率化技術を研究する。

Method	WholeBodyAP 256x192	WB AP 1024x768
Rtmpose-l	0.63	0.68
ViM-S	0.52	0.63
VMamba	0.58	0.67
Swin-B	0.55	0.66
ViT-S	0.55	0.67
ViT-B	0.59	0.68
ViT-L	0.64	0.71
ViT-L 12データセット	-	**0.72**
Sapiens 0.3B	-	0.63



成果についてより詳細な情報を提供しているWebページ、発表論文などの情報：