Vol.2022-CG-185 No.9
2022/3/12

三次元情報を考慮した人物画像の意味的領域分割

奥山 裕大[†1,a)]　金森 由博[†1,b)]　遠藤 結城[†1,c)]　三谷 純[†1,d)]

概要：人物画像の各ピクセルに対して髪やシャツ, スカートなど写っているもののラベルを推定する意味的領域分割が盛んに研究されている. 現在の主流であるニューラルネットワークを用いた手法は RGB 画像のみを入力として, 主に色の違いに基づいて領域を判断している. しかし, 例えば同じ色のスーツの上下 (ジャケットとパンツ) など, 色情報のみでは衣服の境界を識別できない場合がある. そこで本研究では既存手法とは異なり, 人物に関する三次元情報を活用することで人物画像の意味的領域分割の精度を向上させる. 三次元情報としては法線マップに着目し, 人物画像から別途推定して利用する. 本研究では, Transformer に基づく最新のネットワークに基づき, RGB 画像のみを入力とするネットワークと法線情報を入出力に含むネットワークによるアンサンブルを検討した. 最終的な意味ラベルを決めるための Soft Voting の方法として, 個別ネットワークが出力する確率の単純平均と, 不確実性に基づく加重平均を検討した. 提案するアンサンブル手法により, RGB 画像のみを入力とした場合に比べて精度良く意味ラベルを推定できることを示す.

図 1　提案手法は三次元情報 (法線マップ) を考慮することで, 色の似た領域 (赤い点線部分) に対しても精度よく意味ラベルを推定できる.

1. はじめに

スマートフォンに搭載されるカメラの高性能化や, SNS の利用者の増加に伴い, 人物の写真を撮影する機会が増加している. 人物画像中の髪や肌, 衣服の領域が精度良く判別できれば, 特定の衣服の色を変えるなどの多様なアプリケーションが考えられる. こうした背景から, 人物画像の意味的領域分割が盛んに研究されている.

既存の人物画像の意味的領域分割の手法 [21] は, RGB 画像のみが入力である. しかし, 色情報のみに基づいて判断すると, 互いに似た色ながら異なる意味ラベルを持つ領域を区別することが難しい. 例えば**図 1** では, 入力画像中の

†1　現在, 筑波大学
Presently with University of Tsukuba
a)　yutaokuyam@gmail.com
b)c)d)　{kanamori,endo,mitani}@cs.tsukuba.ac.jp

人物のコートとストッキングが同じ黒色で, ベースライン手法ではコートのラベルが足の領域にまではみ出している.

そこで本研究では, 人物に由来する三次元情報, 特に, 形状の起伏を捉えやすい法線マップを考慮した意味的領域分割手法を提案する. 人物画像の意味的領域分割において, 三次元情報を活用した手法は本研究が初めてである. 新たな入力となる法線マップは, 人物画像を対象とした単眼 3D 復元手法 [15] を利用して生成する. これらの RGB 画像と法線マップの情報を活用するために, 以下の 3 つのネットワークを組み合わせたアンサンブル手法を提案する.

(1) RGB 画像のみを入力とするネットワーク
(2) RGB 画像と法線マップを入力とするネットワーク
(3) RGB 画像から意味ラベルと法線マップを同時に推定するネットワーク

これら 3 つのネットワークのバックボーンには Swin Transformer [10] を採用する. 個別ネットワークから出力される各クラスに対する確率に基づいて, Soft Voting により最終的な意味ラベルを出力する. Soft Voting の方法として, 個別ネットワークが出力する確率の単純平均と, 不確実性に基づく加重平均を検討する. (2) のネットワークについては, 法線マップのエンコーダの性能を引き出すために, 合成データを用いた対照学習により事前学習を行う. さらなる精度向上のため, 意味ラベルの境界をネットワークに認識させるための損失関数である Edge Loss を導入する. 図 1 に示す通り, RGB の情報のみを利用したベースラインでは区別の難しい, 黒色が連続したコートとストッキングの境界を提案手法では区別できることがわかる.

2. 関連研究

2.1 人物画像の意味的領域分割手法

人物画像の意味的領域分割は Yamaguchi らによって初めて提案された [19]. 最近の研究は深層学習に基づいている. Tangseng らの研究 [17] では,「同じ人物が T シャツとドレスを同時に身につけている可能性は低い」といった知識に基づき, 汎用的な意味的領域分割手法である FCN [11] に, 衣服の組み合わせを考慮するための CNN を, Outfit encoder として導入した. 他にも深層学習を用いた人物画像の意味的領域分割手法は複数提案されている [21] が, そのいずれも RGB 画像のみを入力としており, 三次元情報の活用に着目した手法は提案されていない. これに対して提案手法は RGB 画像から推定された法線情報も活用することで, さらなる精度の向上を図っている.

2.2 深度情報を用いた意味的領域分割手法

画像の意味的領域分割において, 人物画像を対象に三次元情報を考慮したのは本研究が初めてであるが, 関連する研究として, 室内画像を対象に深度情報を考慮した研究を紹介する. Seichter らの研究 [16] では RGB 画像に加えて深度マップを追加の入力とする手法を提案した. この手法では, 深度情報を考慮した特徴を抽出するために, RGB 画像のエンコーダで, 深度マップから抽出した特徴を足しあわせながら特徴抽出を行った. Wang と Neumann の研究 [18] では,「ピクセル間の深度値が似ているほど, そのピクセル間には関連があるはず」という仮説を立てた. この仮説に基づいて, ウィンドウ内の各ピクセルと中央のピクセルの深度値が近いものほど重みを大きくして畳み込みと平均プーリングを行う, Depth Aware Convolution Module, Depth Aware Average Pooling Module の 2 つのモジュールを提案した. しかし人物画像については室内画像と異なり, RGB 画像と深度情報の両方を含む公開データセットは存在しない. また, 人物画像は室内画像ほど深度の差がはっきりとつかないため, 深度マップは本タスクでは有効に働かないと考えられる. この検証結果は 4 節で述べる.

3. 提案手法

本研究では,「見た目の色では似た領域でも三次元情報があれば区別できるはず」という仮説に基づき, 三次元情報を活用した人物画像への意味的領域分割を提案する. 室内画像や風景画像と比べて人物画像は深度の差がつきにくいことから, 本研究では三次元情報として深度ではなく法線を用いる. 以降ではこれらの情報を用いたネットワーク構造およびその学習方法について説明する.

3.1 ネットワークの概要

本研究で提案するネットワークは, 意味的領域分割の最先端手法の一つである, Swin Transformer [10] に基づいている. Swin Transformer は 4 つの “stage” と呼ばれるブロックから構成される. 大規模 RGB データセットで事前訓練することにより, RGB 画像から様々なタスクに対して強力な特徴を抽出することができる. 加えて, Transformer 由来の利点として, 入力系列の関係性を学習できる [7].

本研究におけるネットワークの構成例を**図 2** に示す. この構成に含まれるネットワークは以下の 3 つである.

RGB → Label: RGB 画像のみを入力とするネットワーク
RGB + N → Label: RGB 画像と法線マップを入力とするネットワーク (3.2 節)
RGB → LabelN: RGB 画像から, 意味ラベルと法線マップを同時に推定するネットワーク (3.3 節)

これら 3 つのネットワークは個別に訓練される. 個別ネットワークから出力された各ピクセルにおける意味ラベルの確率に対し, 単純平均または不確実性に基づいた加重平均によって最終的な意味ラベルを得る (3.4 節). また前述の通り, Swin Transformer の性能を引き出すには事前学習が重要である. RGB 画像のエンコーダに対しては RGB 画像のみの大規模データセットで事前訓練されたモデルを

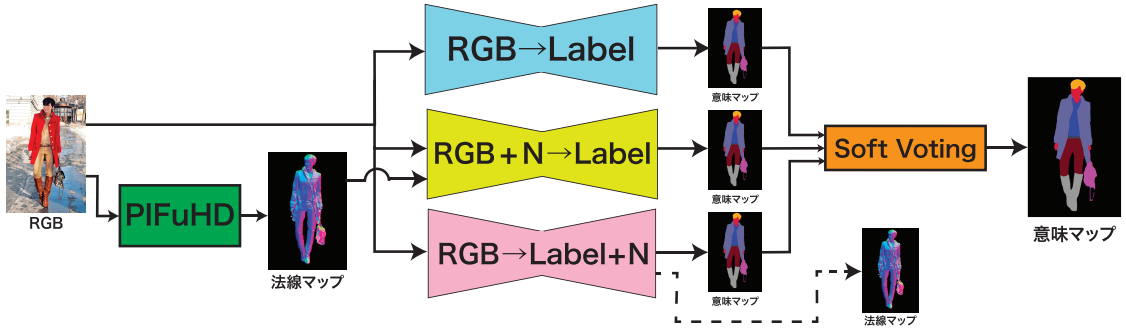

図 2 本研究のネットワーク構成例. RGB 画像と, 推定された法線マップが入力となる. 図中の RGB, N, Label はそれぞれ, RGB 画像, 法線マップ, 意味ラベルマップを表す.

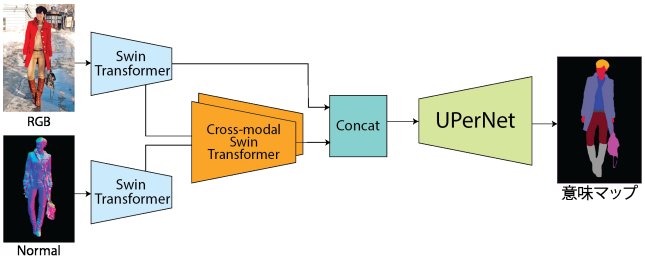

図 3 法線マップを追加の入力として与える手法のネットワーク図. 特徴抽出の段階のうち Stage 3 で RGB 画像, 法線マップから得られたクエリを交換する.

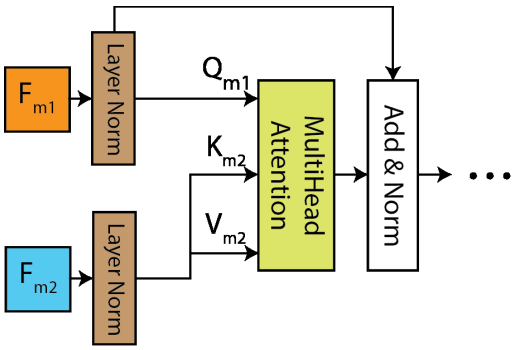

図 4 Cross Attention Module のネットワーク図. 2 種類の特徴量マップを受け取り, クエリを交換する.

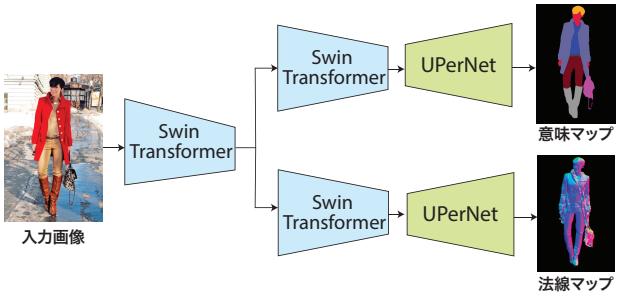

図 5 RGB 画像から意味ラベルの他に法線マップを同時に推定させる手法のネットワーク図.

用いる. 一方, (2) のネットワーク **RGB + N → Label** に含まれる法線マップのエンコーダについては, RGB 画像と法線マップとで性質が異なるため, 3DCG モデルから法線マップを生成し, 対照学習によって事前学習する (3.5 節). さらに, 意味ラベルの境界に着目した Edge Loss を導入する (3.6 節).

3.2 RGB + N → Label

図 3 に法線マップを追加の入力とするネットワーク構造を示す. このネットワークでは, Transformer が系列間の関係性を学習できることに着目し, RGB 画像と法線マップの関係を学習させることを狙う. まず, RGB 画像と法線マップそれぞれから, 別々のエンコーダで特徴を抽出する. その後に, 特徴抽出の際に得られたクエリをそれぞれのエンコーダで交換する. 学習の損失関数には, クロスエントロピー誤差と, Edge Loss を用いた. それぞれ重みは, 1.0 と 0.001 とした.

Cross Modal Attention Module

法線を追加の入力とするネットワークでは, RGB 画像と法線マップの間の Attention を計算する Cross Modal Attention Module [14] (**図 4**) を用いる. これにより RGB 画像と法線マップの関係を捉えた特徴の抽出を狙う.

Transformer で特徴抽出に用いられる Attention とは, 画像のある領域と他の領域との間にどれほど関連があるのかを表すスコアのことであり, 次元 d の 3 つのベクトル, クエリ $\mathbf{Q}$, キー $\mathbf{K}$, バリュー $\mathbf{V}$ から計算される. これらのベクトルは入力画像または特徴量マップを, パッチに分割して Embedding したものを一次元化して得られる.

$$Attention(\mathbf{Q},\mathbf{K},\mathbf{V}) = softmax(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}})\mathbf{V} \quad (1)$$

一方, Cross Modal Attention Module では, RGB 画像と法線マップをエンコードする際に得られたクエリ, キー, バリューから Attention を計算する. これらをそれぞれ $\mathbf{Q}_{RGB}, \mathbf{K}_{RGB}, \mathbf{V}_{RGB}, \mathbf{Q}_N, \mathbf{K}_N, \mathbf{V}_N$ と表す. これらのうち, $\mathbf{Q}$ を交換して Attention を計算する.

$$Attention_{RGB\to N} = softmax(\frac{\mathbf{Q}_{RGB}\mathbf{K}_N^T}{\sqrt{d}})\mathbf{V}_N \quad (2)$$

$$Attention_{N\to RGB} = softmax(\frac{\mathbf{Q}_N\mathbf{K}_{RGB}^T}{\sqrt{d}})\mathbf{V}_{RGB} \quad (3)$$

これらは Swin Transformer の Stage 3 で計算する.

3.3 RGB → Label + N

図 5 に, RGB 画像から意味ラベルと同時に法線マップを推定するネットワーク構造を示す. このネットワークでは, 入力として RGB 画像を受け取り, 意味ラベルと法線マップを出力するマルチタスク学習を行う. 初めの層で共通のネットワークで特徴を出力し, 途中でタスク固有のネットワークに分岐させる. これにより, 三次元情報を考慮した特徴を抽出させることを狙う. 具体的には, Stage 1 および Stage 2 までは共通とし, それ以降は意味ラベルと法線マップそれぞれについて, Swin Transformer とデコーダを追加して出力する. 意味ラベルを推定するデコーダの損失関数には, クロスエントロピー誤差と, Edge Loss, 法線マップを推定するデコーダの損失関数にはコサイン類似度を用いた. それぞれ重みは, 1.0 と 0.001, 0.0025 とした.

3.4 個別ネットワークの不確実性を考慮したアンサンブル

本手法で用いる 3 つのネットワークは, それぞれ識別に得意不得意がある (4.4 節参照). そのため, それぞれのネットワークからの出力をうまく組み合わせることができれば, 全体として精度の高い推定結果が得られると考えられる. 本研究では各ネットワークの出力のうち, 確信度の高い推定結果をより重視するアンサンブルの手法を提案する.

不確実性に基づいた重み付けの方法について説明する. 各ピクセルに出力される各意味ラベルの確率は, 差がつかない場合もあれば, はっきりと差がつく場合もある. 前者は推定の確信度が低く, 後者は高いものと見なし, 確信度の高いネットワークの出力を大きく重み付けする. 具体的には, 不確実さを表す *Softmax Entropy* [4] に基づいて重み付けする. ピクセル i の不確実さ U_i は, あるクラス c の確率を $P_i(c)$ として, 次式で計算される.

$$U_i = -\sum_c P_i(c) \log_2 P_i(c) \tag{4}$$

不確実さを可視化した例を**図 6** に示す. 意味ラベルの境界や首元で不確実さが大きくなっており, ラベルの判断に迷っていることがわかる. 各ピクセル i について, それぞれのネットワークから出力された確率をそれぞれ P_{RGB}, P_{RGBN}, P_{Multi}, 不確実さを U_{RGB}, U_{RGBN}, U_{Multi} とする. 重み付けされた最終的な確率 P'_i は, 次式で表される.

$$W_{RGB} = 1 - \frac{U_{RGB}}{U_{RGB} + U_{RGBN} + U_{Multi}} \tag{5}$$

$$W_{RGBN} = 1 - \frac{U_{RGBN}}{U_{RGB} + U_{RGBN} + U_{Multi}} \tag{6}$$

$$W_{Multi} = 1 - \frac{U_{Multi}}{U_{RGB} + U_{RGBN} + U_{Multi}} \tag{7}$$

$$P'_i = W_{RGB}P_{RGB} + W_{RGBN}P_{RGBN} + W_{Multi}P_{Multi} \tag{8}$$

図 6 不確実さの可視化例. 左から入力画像, 正解の意味ラベル, 推定した意味ラベル, 不確実さの可視化結果.

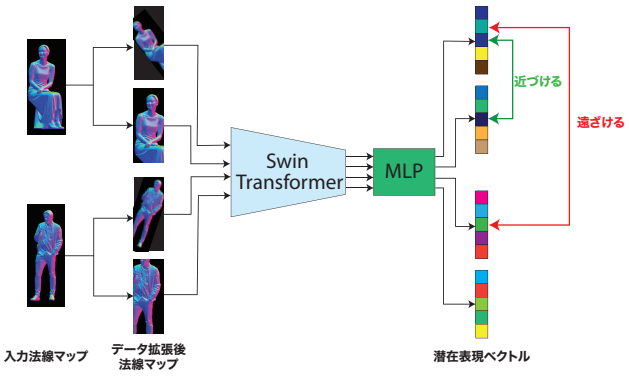

図 7 対照学習を行う手順の概略図. 同じ画像から得られた潜在表現は似るように, 異なる法線マップから得られた潜在表現は違うものになるように学習させる.

3.5 対照学習を用いた事前学習

法線マップのエンコーダのための事前学習について説明する. Transformer の性能を引き出すため, RGB 画像のエンコーダは, 約 1,400 万枚の RGB 画像からなる ImageNet-22K [3] を用いて事前訓練された重みで初期化している. しかし法線マップのエンコーダに対しては, RGB 画像と法線マップの見た目が大きく異なるため, RGB 画像で事前学習した重みを用いても有効な特徴を抽出できない (4.4 節参照). そこで, 3D 人物モデルをレンダリングして作成した法線マップのデータセットを用いて, 自己教師あり学習の一つである対照学習によって事前学習する. 対照学習とは, 「似た画像は潜在変数空間内で距離が近くなるはず」という仮説に基づいて行われる距離学習 [6] の一種である. 対照学習の有効性は画像関連のタスクでも報告されており [1], [13], 本研究では文献 [1] に倣って事前学習を行う.

本研究での対照学習の概略を**図 7** に示す. まず, N 枚の入力画像のミニバッチに対して, 回転やクロップなどのデータ拡張を 2 通り適用して, 全部で $2N$ 枚の法線マップを得る. これを Swin Transformer に入力して特徴量を抽出したのち, 多層パーセプトロンに入力して最終的に 128 次元の潜在変数ベクトルを $2N$ 個得る. 対照学習のための

図 8 エッジマップの例. 左から入力の RGB 画像, 正解の意味ラベル, 推定された意味ラベル, 正解のエッジマップ, 推定された意味ラベルから作られたエッジマップ.

表 1 ベースライン手法と提案手法の定量比較. 各指標の最良値を太字で示す.

	mIoU ↑	mAcc ↑	aAcc ↑
ベースライン	72.94	83.01	97.03
提案手法	**74.23**	**83.79**	**97.21**

損失 $\mathcal{L}_{cont}$ は, $2N$ 個の潜在変数ベクトル $\mathbf{z}$ から計算される. 同じ画像から得られた潜在変数ベクトルの組を $\mathbf{z}_i, \mathbf{z}_j$, コサイン類似度を求める関数を f として, $\mathbf{z}_i, \mathbf{z}_j$ の組の損失は式 (9) で表される. これをバッチ内の各画像について計算して, 最終的な損失 $\mathcal{L}_{cont}$ は式 (10) で計算される.

$$l(i,j) = -\log \frac{\exp(f(\mathbf{z}_i, \mathbf{z}_j))}{\sum_k^{2N} \mathbb{1}_{i,k} \exp(f(\mathbf{z}_i, \mathbf{z}_k))}$$
$$\mathbb{1}_{i,k} = \begin{cases} 1 & i \neq k \\ 0 & i = k \end{cases} \quad (9)$$

$$\mathcal{L}_{cont} = \frac{1}{2N} \sum_{k=1}^{N} [l(2k-1, 2k) + l(2k, 2k-1)] \quad (10)$$

3.6 Edge Loss

ネットワークを意味ラベルの境界に注目させるために, エッジについての損失関数である Edge Loss [2] を追加した. 具体的には, 正解および推定された意味ラベルマップの両方にラプラシアンフィルタを適用してエッジマップを作成し, エッジマップに関するクロスエントロピー誤差を損失とした. **図 8** にエッジマップの例を示す.

4. 実験

4.1 実験環境

提案手法を Python および PyTorch を用いて実装し, NVIDIA RTX A5000 上で学習, 推論を行った. 最適化には AdamW [12] を用いた. 学習率は初期値を 0.0004 に設定し, 5 エポック連続して検証データに対する評価指標が改善しなかった場合に 0.9 倍して減衰させた. 25 エポック連続して評価指標が改善しなかった場合に学習を終了させた. 評価には層化抽出法 [8] を用いた 10 分割交差検証を用いた. 学習にはデータセット 1 分割ごとに約 2 日を要した. 本稿に示す結果画像は, 10 分割交差検証で得られる 10 セット分の出力を平均して意味ラベルを決め, 紙面の都合で背景部分をトリミングしたものである.

4.2 データセット

本研究で用いた RGB 人物画像データセットは UTFPR-SBD3 [5] である. このデータセットは既存の人物画像のデータセット [9], [19], [20] を統合し, 間違ったラベルを修正, 新たに 133 枚の画像を追加して作成された. 画像の総数は 4,500 枚である. 交差検証での 1 分割の内訳は, 訓練データ 4,091 枚, テストデータ 409 枚とした. 人物画像の法線マップについては, 一枚の RGB 画像から人物 3D モデルを再構成する手法である PIFuHD [15] によって人物 3D モデルを推定したのち, 平行投影により法線マップを得た.

4.3 ベースライン手法との比較

定量比較

ベースライン手法と提案手法について, 定量評価指標による評価を行った. ベースライン手法は, RGB 画像から意味ラベルを推定するネットワークである. 評価指標として mIoU, mAcc, aAcc を用いた. ここまで 3 つのネットワークを検討してきたが, 後述する ablation study の結果より, **RGB → Label** と **RGB + N → Label** の 2 つのネットワーク出力の単純平均が最良と判明したため, 以降ではこの組み合わせを提案手法として評価する. **表 1** に示す通り, 各評価指標の値において提案手法がベースライン手法を上回ったことがわかる. また, **表 2** にベースラインと提案手法のクラスごとの IoU の値を示す. 多くのクラスで IoU が改善した一方, ネックウェアと靴下で顕著に悪化が見られた. これらのクラスは, **RGB + N → Label** でも, **RGB → Label** に比べて顕著に精度が悪化したため, 法線マップを用いることで識別が不利になるクラスと考えられる. この悪化の原因の考察は, 4.4 節で行う.

定性比較

図 9 に, ベースラインと提案手法の出力結果の比較を示す. 上段の結果は, コートの襟周りの領域とシャツが白色で連続しているために, RGB 画像のみを入力とするベースラインではシャツだと誤っている. 一方, 法線情報を利用している提案手法では, 正しくコートとシャツの領域を分割できている. 下段の結果も同様に, コートの襟周りとシャツが黒色で連続している領域をベースライン手法では区別できていない. 一方, 提案手法ではコートとシャツを正しく識別できている領域が増えている. これは上段, 下段共に法線マップ上で胸元のコートとシャツの境界が現れていることが, 識別に有利に働いたと考えられる.

表 2 ベースラインと提案手法のクラスごとの IoU. “BG” は “Background”, “R/J” は “Rompers/Jumpsuit” の略. 提案手法の実験条件については 4.3 節参照.

	BG	Bag	Belt	Coat	Dress	Footwear	Hair	HeadWear	R/J	Neckwear	Shorts	Skin	Skirt	Socks	Stocking	EyeWear	Sweater	Shirt	Pants
ベースライン	98.97	82.51	62.79	76.69	75.51	81.29	82.35	82.29	43.90	**54.18**	81.81	88.08	**79.72**	**52.92**	67.77	72.32	42.95	76.46	83.53
提案手法	**99.03**	**83.79**	**65.78**	**76.73**	**76.24**	**82.21**	**83.03**	**83.09**	**45.73**	53.77	**83.66**	**88.64**	79.62	52.41	**74.06**	**73.62**	**44.76**	**78.86**	**85.38**

図 9 定性評価結果. 左から入力 RGB 画像, 入力法線マップ, 正解意味ラベル, ベースライン, 提案手法の出力結果.

表 3 Edge loss の有無による定量比較.

実験条件	mIoU ↑	mAcc ↑	aAcc ↑
ベースライン	72.94	83.01	97.03
ベースライン + Edge Loss	**73.28**	**83.19**	**97.07**

表 4 対照学習を用いた事前学習の有無による定量比較.

実験条件	mIoU ↑	mAcc ↑	aAcc ↑
事前学習なし	72.43	82.64	96.99
事前学習あり	**73.28**	**83.28**	**97.10**

4.4 Ablation Study

Edge Loss の有効性についての検証

ベースライン手法に対する Edge Loss の有無の比較によって, Edge Loss の有効性を検証した. **表 3** に結果画像に対する各評価指標の値を示す. 表より, Edge Loss による改善が読み取れる.

対照学習を用いた事前学習の有効性についての検証

ネットワーク **RGB + N → Label** の法線マップのエンコーダに対する, 対照学習を用いた事前学習の有無の比較によって, 対照学習の有効性を検証した. **表 4** の実験結果より, 対照学習が有効に働いたことがわかる.

表 5 本手法で用いる各ネットワークの定量比較.

ネットワーク	mIoU ↑	mAcc ↑	aAcc ↑
RGB → Label	**73.28**	83.19	97.07
RGB + N → Label	73.25	**83.28**	**97.10**
RGB → Label + N	72.19	82.25	96.96

個別ネットワークの評価

表 5 に個別のネットワークを定量評価した結果を示す. 表より, mIoU においては **RGB → Label** が, mAcc, aAcc においては **RGB + N → Label** が最良だとわかる.

加えて, 個別ネットワークの識別の傾向を確認するために, クラスごとの mIoU の比較と, 混同行列による評価を行った. **表 6** に個別ネットワークのクラスごとの IoU を示し, **図 10**, 11, 12 に, 個別ネットワーク同士の混同行列の差分を示す. ここで示す混同行列の値は, 値を各行で正規化したもので, 単位はパーセントポイントである.

表 6 より, **RGB + N → Label** は **RGB → Label** と比べて, ベルトやロンパース/ジャンプスーツ, ショートパンツで改善が見られた. 一方, 靴下やセーターで悪化した. 図 10 のベルトの列を見ると, 他のクラスの値が下がっていない. これは, ベルトの形状を捉えたことで新たに識別できるようになったためと考えられる. 同様に, 図 10 のロンパース/ジャンプスーツの行を確認すると, シャツやパンツと誤って識別していたピクセルが減っている. これは, シャツやパンツに比べて, ロンパース/ジャンプスーツはゆったりとした衣服であるため, 大きなシワやヨレの形状を, 法線で上手く捉えられたからであると考えられる. ショートパンツが改善した理由は, スカートと誤った識別をしているピクセルが減少している影響が大きい. これは, ショートパンツを身につけた際に, 足の形がスカートに比べて現れることを捉えた結果だと考えられる. 一方, **RGB + N → Label** では, ネックウェアやセーター, 靴下で顕著な精度の悪化が見られた. ネックウェアは, 例えばマフラーやネクタイなど, 首の周りに身につける衣服のことである. マフラーやネクタイは生地が薄く風でなびいたり, 比較的狭い領域に複雑な形状をしていたりするため, 他の衣服に比べて変形が大きく 3D モデルの再構成に失敗しやすい. これによる法線の欠損や乱れが, 推定に不利に働いていると考えられる. セーターが悪化した理由は, コートと誤った識別が増えた結果だと考えられる. セーターはコートのように上に羽織る場合が多く, 形状的に似ることが多いためと考えられる. また, 靴下で悪化した原因はストッキングとの誤った識別が増えたためである. これは, 靴下とストッ

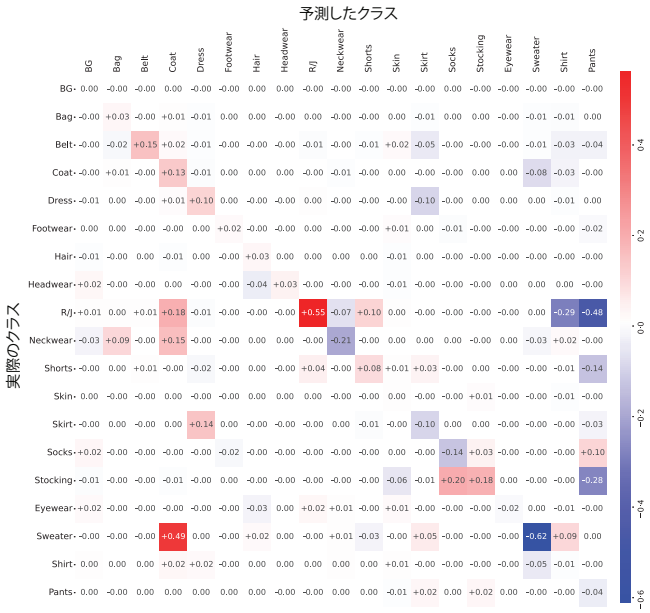

図 10 **RGB + N → Label** の混同行列から **RGB → Label** の混同行列を引いた結果.

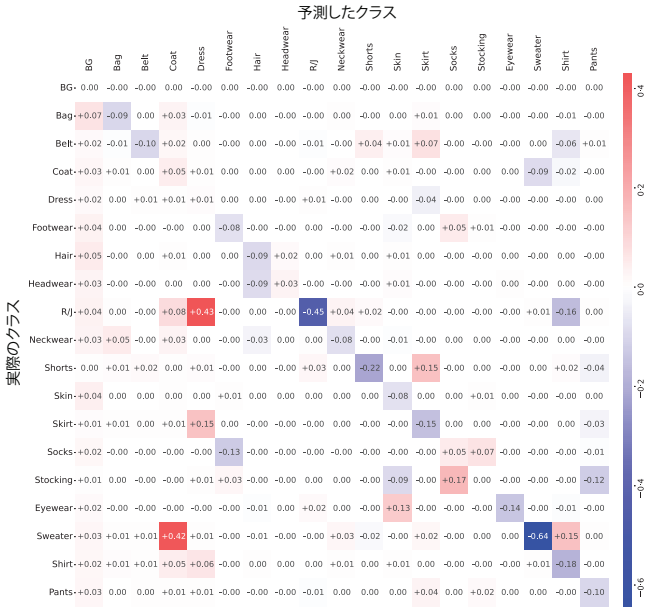

図 11 **RGB → Label + N** の混同行列から **RGB → Label** の混同行列を引いた結果.

キングはどちらも肌にぴったりと張り付く衣服であるため, 形状の上で判別しづらいことが影響していると考えられる.

一方, 表 5, 6 を見ると, **RGB → Label + N** は他のネットワークに対して全ての評価指標で劣っている. 混同行列の差分の**図 11**, 12 を確認しても, 他のネットワークと比べて精度が上がっているとは言えない. この理由として, **RGB + N → Label** ほど明示的に法線情報を考慮しておらず形状を考慮することによる精度向上の恩恵を受けられていない一方, 意味ラベル推定と法線マップ推定のネットワークを一部共有しており RGB 画像から十分に特徴抽出を行えていないためだと考えられる.

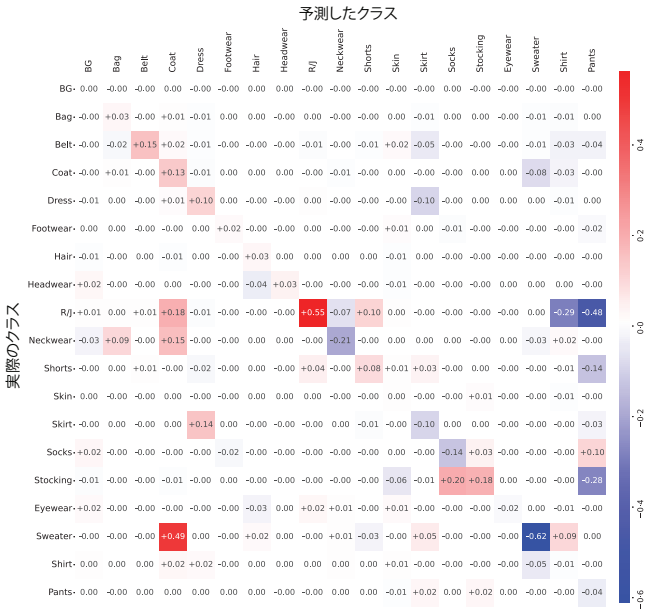

図 12 **RGB + N → Label** の混同行列から **RGB → Label + N** の混同行列を引いた結果.

個別ネットワークのアンサンブル方法の検討

個別ネットワークの組み合わせ方や Soft Voting の方法について検討した. **表 7** に実験条件, **表 8** に実験結果を示す. 2 つのネットワークの組み合わせについては, 前述の評価結果を踏まえ, 有効だと考えられる **RGB → Label** と **RGB + N → Label** の組み合わせのみ検討した. 結果として, 個別ネットワークを 2 つ用いて単純平均を行った結果が最良となった. 不確実性を考慮した加重平均については, ネットワークを 3 つ用いた場合 (実験条件 1,2), 評価指標の値が僅かに向上した. 一方, ネットワークを 2 つ用いた場合 (実験条件 3, 4), 単純平均の方がよい結果となった.

深度マップではなく法線マップを採用した妥当性の検証

三次元情報として, 法線マップの代わりに深度マップを入力した場合と, 法線マップを入力した場合を比較し, 法線マップを使う妥当性を検証した. **表 9** に示す通り, 法線マップを用いた結果が, 深度マップを用いた場合よりも全ての評価指標において上回った.

5. 結論

本研究では, 人物画像に関する意味的領域分割において, 三次元情報を考慮した初の手法を提案した. 三次元情報として法線マップに着目し, 人物画像を対象とした単眼 3D 復元手法 [15] を利用して法線マップを生成した. 意味的領域分割の最先端汎用ネットワークである Swin Transformer [10] に基づいて, RGB 画像のみを入力としたネットワークや法線情報を入出力に含むネットワークのアンサンブル手法を検討した. 最終的な意味ラベルを決定するための Soft Voting の方法として, 単純平均と不確実性に基づく加重平均を検討した. また, 法線マップのエンコーダには, 法線

表 6 個別ネットワークのクラスごとの IoU.

	BG	Bag	Belt	Coat	Dress	Footwear	Hair	HeadWear	R/J	Neckwear	Shorts	Skin	Skirt	Socks	Stocking	EyeWear	Sweater	Shirt	Pants
RGB→Label	98.97	82.85	63.59	**75.97**	**75.62**	81.37	82.34	**82.22**	41.13	**53.99**	81.73	88.06	**79.56**	**53.29**	**71.26**	**72.87**	**45.97**	77.55	83.92
RGB+N→Label	**99.00**	**83.00**	**64.67**	75.72	75.56	**81.61**	**82.56**	82.12	**44.67**	52.94	**83.00**	**88.23**	79.15	49.59	71.19	**72.87**	42.94	**78.01**	**84.96**
RGB→Label+N	98.94	82.12	62.12	75.12	74.62	80.83	81.87	81.41	42.85	50.57	81.42	87.73	78.07	50.01	69.47	71.90	41.30	76.94	84.21

表 7 アンサンブル方法の検討のための実験条件.

	組み合わせるネットワーク	ラベルの出力方法
実験条件 1	**RGB→Label**, **RGB+N→Label**, **RGB→Label+N**	不確実性に基づく加重平均
実験条件 2	**RGB→Label**, **RGB+N→Label**, **RGB→Label+N**	単純平均
実験条件 3	**RGB→Label**, **RGB+ N→Label**	不確実性に基づく加重平均
実験条件 4	**RGB→Label**, **RGB+N→Label**	単純平均

表 8 アンサンブル方法の検討実験の結果.

実験条件	mIoU ↑	mAcc ↑	aAcc ↑
実験条件 1	74.05	83.62	**97.21**
実験条件 2	74.02	83.62	97.20
実験条件 3	74.22	83.78	97.20
実験条件 4	**74.23**	**83.79**	**97.21**

表 9 深度マップではなく法線マップを採用した妥当性の検証結果.

実験条件	mIoU ↑	mAcc ↑	aAcc ↑
深度マップを入力	73.91	83.58	97.19
法線マップを入力	**74.23**	**83.79**	**97.21**

マップから有用な特徴を抽出するために, 対照学習を用いた自己教師あり学習を行った. さらに, 意味ラベルの境界の推定精度を向上させるために, Edge Loss を導入した. これらの工夫について有効な組み合わせを検討し, RGB 画像のみから意味ラベルを推論するベースラインよりも精度を改善できることを示した. 今後の課題として, 法線マップとは異なる三次元情報の利用が挙げられる. 例えばアンビエントオクルージョンマップや曲率マップなど, 人物 3D モデルならではの情報を有効に活用することで精度を向上させられる可能性がある.

参考文献

[1] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pp. 1597–1607. PMLR, 2020.

[2] Yifu Chen, Arnaud Dapogny, and Matthieu Cord. SEMEDA: Enhancing segmentation precision with semantic edge aware loss. *Pattern Recognition*, Vol. 108, p. 107557, 2020.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR 2009*, pp. 248–255, 2009.

[4] Rebecca Hwa. Sample selection for statistical parsing. *Computational Linguistics*, Vol. 30, No. 3, pp. 253–276, 2004.

[5] Andrei De Souza Inacio and Heitor Silvério Lopes. EPYNET: Efficient pyramidal network for clothing segmentation. *IEEE Access*, Vol. 8, pp. 187882–187892, 2020.

[6] Mahmut Kaya and Hasan Sakir Bilge. Deep metric learning: A survey. *Symmetry*, Vol. 11, No. 9, p. 1066, 2019.

[7] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *arXiv preprint arXiv:2101.01169*, 2021.

[8] Edo Liberty, Kevin Lang, and Konstantin Shmakov. Stratified sampling meets machine learning. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *ICML 2016*, Vol. 48 of *Proceedings of Machine Learning Research*, pp. 2320–2329. PMLR, 2016.

[9] Si Liu, Jiashi Feng, Csaba Domokos, Hui Xu, Junshi Huang, Zhenzhen Hu, and Shuicheng Yan. Fashion parsing with weak color-category labels. *IEEE Transactions on Multimedia*, Vol. 16, No. 1, pp. 253–265, 2013.

[10] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV 2021*, pp. 10012–10022, 2021.

[11] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR 2015*, pp. 3431–3440. IEEE, 2015.

[12] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR 2019*. OpenReview.net, 2019.

[13] Taesung Park, Alexei A. Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *ECCV 2020 Proceedings, Part IX*, Vol. 12354 of *Lecture Notes in Computer Science*, pp. 319–345. Springer, 2020.

[14] Bin Ren, Hao Tang, Fanyang Meng, Runwei Ding, Ling Shao, Philip HS Torr, and Nicu Sebe. Cloth interactive transformer for virtual try-on. *arXiv:2104.05519*, 2021.

[15] Shunsuke Saito, Tomas Simon, Jason M. Saragih, and Hanbyul Joo. PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization. In *CVPR 2020*, pp. 81–90. Computer Vision Foundation / IEEE, 2020.

[16] Daniel Seichter, Mona Köhler, Benjamin Lewandowski, Tim Wengefeld, and Horst-Michael Gross. Efficient RGB-D semantic segmentation for indoor scene analysis. In *ICRA*, pp. 13525–13531, 2021.

[17] Pongsate Tangseng, Zhipeng Wu, and Kota Yamaguchi. Looking at outfit to parse clothing. *arXiv preprint arXiv:1703.01386*, 2017.

[18] Weiyue Wang and Ulrich Neumann. Depth-aware CNN for RGB-D segmentation. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *ECCV 2018*, Vol. 11215 of *Lecture Notes in Computer Science*, pp. 144–161. Springer, 2018.

[19] Kota Yamaguchi, M Hadi Kiapour, Luis E Ortiz, and Tamara L Berg. Parsing clothing in fashion photographs. In *CVPR 2012*, pp. 3570–3577. IEEE, 2012.

[20] Wei Yang, Ping Luo, and Liang Lin. Clothing co-parsing by joint image segmentation and labeling. In *CVPR 2014*, pp. 3182–3189. IEEE Computer Society, 2014.

[21] Xingxing Zou and Waikeung Wong. fAshIon after fashion: A report of AI in fashion. *arXiv preprint arXiv:2105.03050*, 2021.