令和 3 年度修士論文

深層学習を用いた遮蔽あり単一画像からの人体 3 次元形状推定

宮崎大学大学院 工学研究科 工学専攻

機械・情報系コース 情報システム工学分野

学籍番号 T2003531 森慎太郎

指導教員 椋木 雅之 教授

令和 4 年 1 月 24 日

概要

本研究では、遮蔽ありの単一画像から画像内の人体の 3 次元形状を推定することを目指す。そのために、深層学習を用いた前景抽出、画像補完、3 次元形状推定の 3 つの技術を組み合わせて一般背景下で撮影された画像からの 3 次元形状推定処理を実現する。

本研究では、3 次元形状推定手法として深層学習を用いた PIFu という手法を用いる。この手法は 1 枚または少数の画像から画像に写っていない面も含めた全周囲の形状を推定して 3 次元モデルを生成できる。また、人物像の服の色や模様を反映したテクスチャも 3 次元モデルに付与できる。

PIFu による 3 次元形状推定には、元画像だけでなく画像内の人体領域を示すシルエット画像が必要である。一般背景下で撮影された画像から前景となる人体領域を抽出するための前景抽出手法として、深層学習の一種である$U^2 - Net$を用いる。学習済みの$U^2 - Net$により、一般背景下であっても人体領域を良好に抽出できる。

人体の一部が遮蔽されている場合、前景抽出結果のシルエット画像でも遮蔽部分は欠損している。PIFu による 3 次元形状推定では、このような欠損があると 3 次元形状推定が適切に行えない、これに対処するために、深層学習を用いた画像補完法である GMCNN を用いる。GMCNN により元画像を画像補完した場合、欠損部分が背景に近いテクスチャにより補完される傾向があった。そのため、本研究では元画像とは別に、2 値画像であるシルエット画像の欠損部分に対して画像補完を適用し、適切なシルエット画像を得る。

実験では、提案手法による 3 次元形状推定結果を示した。人体の腕や足の途中部分が遮蔽されている場合は、シルエット画像の補完が適切に行え、違和感のない 3 次元形状推定結果が得られた。一方、腕先や足先の部分が遮蔽されている場合は、画像補完により先端部分まで補完することができなかった。よりよい画像補完手法の導入や、3 次元形状推定結果の定量的評価、人体以外の種々の物体への 3 次元形状推定対象の拡大が今後の課題である。

目次

1. はじめに

近年、深層学習の技術が発展し、特に画像認識や音声認識といった分野で大きな成果を生み出している。これに伴い、従来の画像処理、画像生成、コンピュータビジョン手法に代わり、深層学習を用いた手法が種々開発されている。

例えば、画像からの 3 次元形状復元においては、従来は画像を撮影したカメラと 3 次元点の幾何学的関係を数学的に求める Shape from X と総称される手法が用いられていた。画像は 3 次元世界の像を 2 次元に投影したものであり、奥行き情報を失った 2 次元の画像から、元の 3 次元世界を正確に復元することは数学的に解けない不良設定問題となる。不良設定問題を数学的に解けるようにするために、従来は、多数の画像を利用したり、対象物体の特性に関して仮定を置くなどの工夫をしていた。一方、人間は単一の画像からでも 3 次元世界の情報をある程度得ることができる。深層学習による画像からの 3 次元形状推定では、画像とその画像内に撮影された対象の 3 次元形状の組を多数学習することで、人間と同様に、単一または少数の画像から画像内の物体の 3 次元形状を推定することが可能となっている。

他の例として、画像補完では、従来、補完領域の周りの画素値を補完領域内に伝播して補完する手法や補完領域と類似した部分画像を用いて補完領域を埋める手法などがあった。これらの画像処理的な手法は、補完領域の被写体の種類に依存せず適用可能という利点はあるが、補完領域が大きくなると精度が落ちる、処理に時間がかかるといった問題があった。これに対して、深層学習を利用した画像補完手法では、補完領域の被写体の種類に応じて事前に学習を行う必要はあるが、補完領域が比較的大きくても精度よく補完が行える。また、学習には時間がかかるが、学習済みのモデルがあれば補完処理自体は高速に行える。特に、敵対的学習(GAN)を用いた手法は、精度良く画像補完が行えるため、多くの手法が開発されている。

他にも、前景抽出では、従来、画素値の閾値処理により、前景と背景を分ける手法が一般的であったが、深層学習を用いれば、前景となる対象を学習することにより、精度の高い結果を得ることができるようになっている。

本研究では、深層学習に基づく 3 次元形状推定、画像補完、前景抽出の 3 つの手法を組み合わせることにより、これまで実現されていなかったような画像からの 3 次元形状推定を目指す。具体的には、一般背景下で撮影された単一画像から、画像内の人体の 3 次元形状を推定する。画像内の人体は、他の物体との重なり等により一部が遮蔽されていてもよいものとする。このような処理が実現できれば、例えば、集合写真内の人物を 3 次元形状推定し別角度から撮影した画像を生成したり、防犯カメラの一部遮蔽された人物像から任意視点の見え方を推定するといった応用が可能となる。

本研究では、深層学習に基づく単一画像からの人体 3 次元形状推定手法を基に、深層学習による、前景抽出を組み合わせることで、一般背景から人体 3 次元形状推定に必要なシルエット画像を自動で得る。また、遮蔽がある場合を考慮して、深層学習による画像補完を適用する。この際、元画像だけでなく、シルエット画像にも画像補完を適用することで、人体 3 次元形状推定の精度が向上することを示す。また、これらの手法を一貫した処理として構成することで、入力画像から人体 3 次元形状推定までを自動で行えるようにする。

以下、2 章では、本研究に関連する 3 つの要素技術について述べる。3 章では、提案手法の具体的な処理手順を述べる。4 章では、実験を行い、定性的に人体 3 次元形状推定結果を評価する。5 章はまとめとする。

2. 3 次元形状推定の要素技術

2.1 3 次元形状推定

3 次元形状推定とは、単一または複数の画像が与えられた際に、画像内に撮影された 3 次元世界の幾何学的な情報を推定する処理である。

従来は、Shape from X と総称される手法が主に用いられていた。例えば、陰影情報から物体表面の法線方向を求めて、表面の形状を復元する Shape from Shading や、レンズカメラの合焦距離からのズレとピンぼけのボケ幅の関係から物体とカメラとの距離を推定する Shape from Defocus などの手法が知られている。他にも、移動するカメラから撮影した複数の画像から、それらの画像を撮影したカメラの位置・姿勢と撮影された物体の 3 次元形状を同時に復元する Shape from Motion(Structure from Motion)や、複数の画像から得られる物体シルエットとカメラの射影中心が形成する錐体の共通部分により 3 次元形状を求める Shape from Silhouette などがある。これらの手法は多数の画像を利用したり、対象物体に条件をつけることで、本来 2 次元の画像からは復元できない 3 次元的な情報を、数学的に解ける問題設定に持ち込んで解いている。一方で、対象物体が完全拡散反射をする、撮影時のカメラや光源の位置が既知である、カメラのレンズ特性が既知である、3 次元世界で同一の点が画像間で対応点として与えられるといった制限があり、適用範囲が限られる。

これに対して、近年、深層学習を用いた手法が多く提案され、単一画像や少数の画像から、より一般的な状況で 3 次元形状を推定することが可能となってきている。例えば、PIFu[1] は、深層学習を用いて 1 枚もしくは少数枚の画像から、画像内の人体の 3 次元形状を推定することができる。図 1 に PIFu による 3 次元形状推定の例を示す。上段の入力画像とシルエット画像から、下段の 3 次元モデルが生成される。入力の画像には含まれない、人体背景の 3 次元形状やテクスチャも復元されている。多くの同種の手法では、推定した 3 次元形状を含むボクセル空間を直接出力するのに対し、PIFu は、 3 次元空間の各点が人体の内部なのか外部なのかを判定する判定器を学習により構築する。これにより、学習に必要なメモリ容量を削減でき、高解像度の 3 次元形状推定が行える。また、同様の考え方で人体表面のテクスチャも推定した 3 次元形状に同時に付与することができる。さらに、PIFu は 1 枚の画像からでも形状復元を行えるが、複数枚の画像が与えられた場合には、推定精度をより向上させることができる。

深層学習を用いた 3 次元形状推定では、入力される画像と出力される 3 次元形状の組を学習データとして与えて学習する必要がある。形状推定の結果は、学習データに依存するので、例えば 3 次元形状推定を行うには、人体のデータセットが必要となる。また、学習には多くの処理時間を要するという欠点もある。しかし、近年、多くのデータで学習した学習済みモデルが提供されており、比較的簡単に 3 次元形状の推定が行えるようになっている。

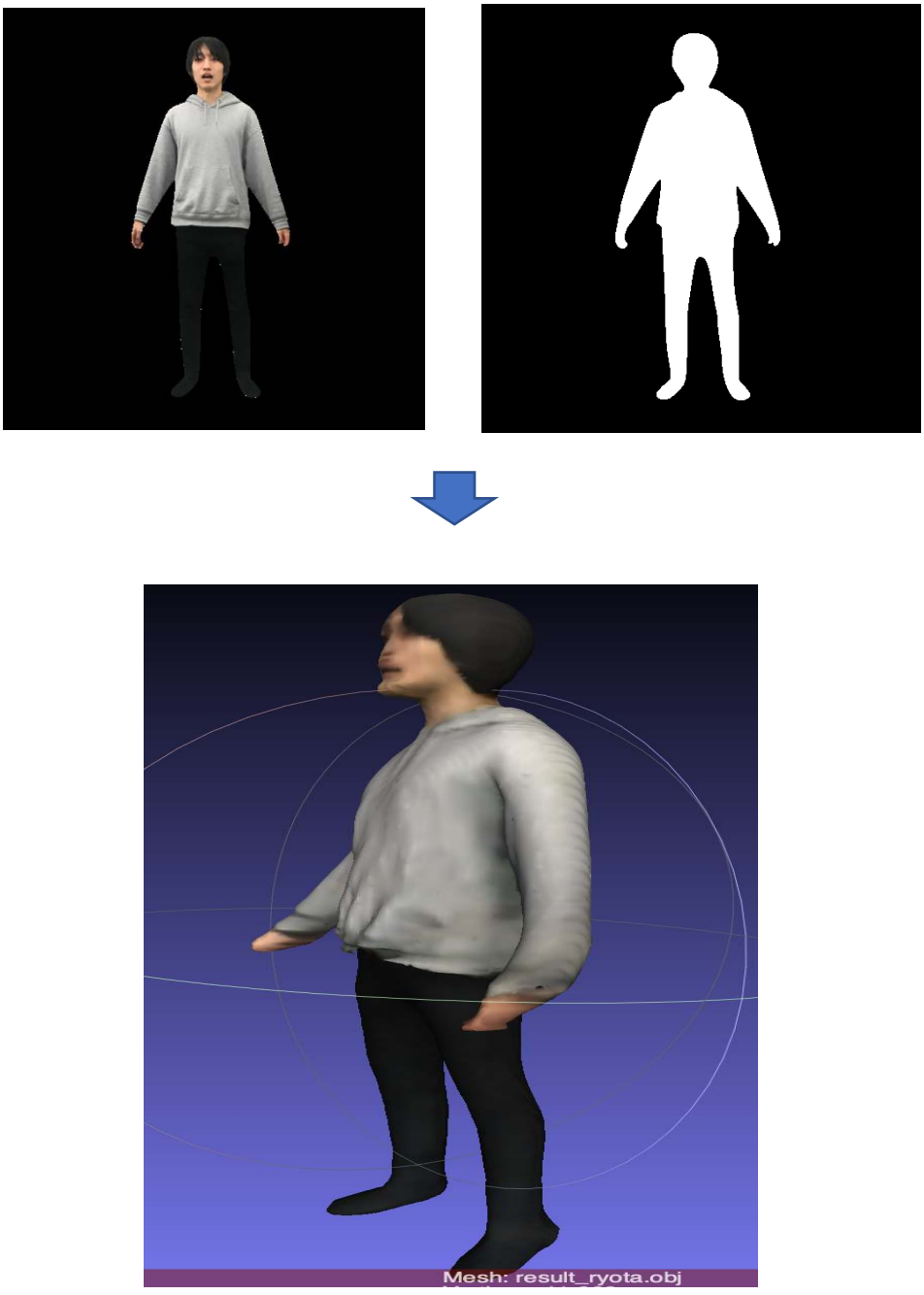


図 1　PIFu の実行結果

2.2　画像補完

画像補完とは、与えられた画像の一部の領域(マスク領域)に対して、マスク領域以外の画像情報を用いて、自然な画像に見えるよう画素値を自動で補完する技術である。画像に写り込んだ望まない物体の削除や、ノイズ等により一部が欠損した画像の復元に用いられる。現在研究されている手法は、大まかに 3 つに分けられる。

1 つ目は、拡散ベースの手法である。この手法では、マスク領域の周辺画素値をマスク領域内に徐々に伝播して補完を行う。周辺画素の画素値の勾配情報なども考慮して伝播を行うことで、周囲との連続性を保った画像補完が行える。しかし、マスク領域が大きい場合は、単に周囲と連続した画素値を補完するのでは、自然な画像補完結果が得られないことがある。

2 つ目は、パッチベースの手法である。この手法は、マスク領域周辺と類似した領域を画像内から探し、マスク領域に貼り付けることで補完を行う。画像内の被写体が、果物の盛り合わせや都市のビル群、岩山や湖といった類似した物体群である場合は、マスク領域が比較的大きくても自然な補完が行える。しかし、画像内にマスク領域と類似した領域がないと、適切な補完が行えない。例えば、1 人の人物の顔を大きく撮影した画像に対して、口の部分をマスク領域とした場合、画像内に他に類似した領域がないため、自然な補完とならない。

3 つ目は、学習ベースの手法である。この手法では、まず、入力画像とマスク領域に対してマスク領域の望ましい補完結果の組を学習データとして与え、多数の学習データを用いて学習する。得られた学習済みモデルに、任意の入力画像とマスク領域を与えることで、マスク領域を適切に補完できる。パッチベースの手法では、画像内にマスク領域と類似した領域が含まれている必要があるが、学習ベースの手法では、多数のデータを学習することで学習済みモデルとして適切な補完に必要な情報が保持されるため、マスク領域と類似した領域が含まれてなくてもよい。従来の学習ベースの手法では、マスク領域がぼやけて不自然となる傾向があった。これに対して、深層学習に敵対的学習(GAN)を組み合わせた手法が提案され、精度良く補完が行えるようになっている。例えば、GMCNN[2]は、GAN を取り入れた深層学習に基づく高速、高精度な画像補完法である。図 2 に GMCNN による画像補完の例を示す。図中の左の画像の白い四角の部分がマスク領域、右の画像がマスク領域を GMCNN により補完した結果である。自然な補完が行えていることがわかる。この手法では、3 つの畳み込みネットワークにより異なる粒度で画像特徴を抽出し、補完画像を生成する。補完画像を、画像全体の自然さを評価するグローバル判別機とマスク領域の自然さを評価するローカル判別器にかけ、GAN の枠組みでより自然な補完が行えるよう学習する。さらに、補完画像を学習済みの畳み込みネットワーク VGG19 に与え、得られる特徴量から補完画像の自然さを評価することで、マスク領域とその周囲とを違和感なく補完できるよう学習する。

深層学習を用いた学習ベースの画像補完でも、補完結果は学習データに依存する。多様な

画像を学習することで、多様な画像の補完が行えるが、学習に用いていない未知の物体が写っている画像の補完は適切に行えないことがある。また、多くの学習データを学習するには、多くの処理時間を要する。一方で、学習済みのモデルがあれば、補完処理自体は比較的高速に行える。

図 2　GMCNN の結果

2.3　前景抽出

前景抽出とは、ある対象物体が主に写った画像内から、その対象物体の領域（前景）部分を抜き出す処理である。画像内の前景以外の部分を背景と呼ぶ。本研究において、前景は人体領域である。前景抽出手法は、大きく３つに分けられる。

１つ目は、色やテクスチャに基づく手法である。前景または背景に色やテクスチャの一様性を仮定し、一様な特徴をもつ領域を前景または背景として抜き出す。例えば、青や緑の単一色の背景にいる人物を撮影した画像に対して、青や緑以外の部分を前景として取り出すことができる。このような手法は、単純な処理なので高速に処理が行えるため、クロマキー合成として放送現場などで実利用されている。しかし、背景が一様な特殊な環境で撮影した画像に対してのみ適用可能な手法である。

２つ目は、奥行き情報に基づく手法である。RGBD カメラやステレオカメラを用いることで、各画素について色情報だけなく、カメラから被写体までの距離(奥行き情報)が得られる。前景は、背景に対して前方にあると仮定し、カメラからの距離が近い画素の連続領域を前景として抽出する。奥行き情報が得られれば、複雑な背景や撮影時の照明変化があっても適切に前景を抽出できるが、撮影時に特殊なカメラが必要であり、一般の画像には適用できない。

３つ目は、物体認識に基づく手法である。この手法は、前景の物体がどのようなものであるか、あらかじめ処理の組み合わせや対象物体の画像の学習により与えておく。入力された画像内で、物体を探すことで、前景を抽出する。前景がどんな物体であるか、事前に与える必要があり、対象物体が限定されるが、どのような背景であるかに関わらず、一般の画像から前景を抽出できる。近年は、深層学習を用いた前景抽出法も多く提案されている。例えば、U^2-Net[3]は深層学習を利用した前景抽出方法で、人や猫などの画像を入力として、前景と背景を分離するためのアルファ値を計算することできる。U^2-Netは深層学習を利用した前景抽出で、既存の U-Net[4]を改良したものである。ネストさせることで、局所的な特徴と大域的な特徴を効率的に抽出をすることができる。

3． 遮蔽あり単一画像からの 3 次元形状推定手法

3.1 処理の流れ

本研究では、一般背景下で撮影された 1 枚の人物画像から、画像内の人物を 3 次元形状推定し、3 次元モデルを生成する。画像内の人物は他の物体等との重なりで一部遮断されて良いものとする。

入力として、形状推定の対象となる人物が写された元画像と、その人物像の中で一部遮蔽されている部分を示すマスク領域が与えられる。出力は、推定された人体の 3 次元モデルである。 3 次元モデルは、メッシュモデルで表現される。

3 次元形状推定には、画像内の人物像の領域を示すシルエット画像が必要となる。一般背景下で撮影された元画像から、シルエット画像を得るために、前景抽出を適用する。また、画像内の人物像の一部が遮蔽されたままだと、正しく 3 次元形状推定が行えない。これに対処するために、遮蔽部分に画像補完を適用する。

以下に処理の流れを示す。

1.シルエット抽出

2.遮蔽部分の補完

3.人体 3 次元形状推定

1.で、元画像から人物像の領域を前景として抽出したシルエット画像を生成する。シルエット画像は、人物領域が 1 、背景が 0 の画素値をもつ 2 値画像である。2.では、遮蔽部分を補完する。この際、元画像だけでなくシルエット画像にも画像補完の処理を適用することを提案する。補完した元画像、シルエット画像をそれぞれ補完元画像、補完シルエット画像と呼ぶ。3.では、補完元画像と補完シルエット画像を元に、 3 次元形状推定を行う。

3.2　シルエット抽出

シルエット抽出では、一般背景下で撮影された元画像から、3 次元形状推定の対象となる人物像の領域をシルエット画像として抽出する。

本研究では、シルエット抽出にU^2-Netを利用する。U^2-Netの構造を図 3 に示す。U^2-Netは、エンコーダ・デコーダ構造を持つ U-Net を拡張した構造となっている。入力画像はエンコーダに与えられる。図左側のエンコーダでは、画像の解像度を下げながら、情報を集約しつつ特徴抽出を行う。その後、図右側のデコーダにより、抽出した特徴量を利用して解像度を上げながら、最終的に所望の出力(本研究の場合、人体のシルエット画像)を得る。解像度を上げる際、詳細部分の情報を補うため、エンコーダの途中結果も利用する。このような構造を U-Net 構造と呼ぶ。U^2-Netでは、エンコーダ・デコーダの各段階の処理においても U-Net 構造を用いる 2 重にネストした U-Net 構造を採用している。これにより局所的な特徴から大域的な特徴まで、様々な粒度での特徴量抽出が行え、適切なシルエット画像を生成可能となる。

人体像の抽出のためには、人体画像を用いて学習する必要がある。本研究では、人物画像の学習済みモデル[5]を利用する。学習済みのU^2-Netに元画像を与えると、前景らしさ（アルファ値）を画素値としてもつ画像が出力される。この画像に対して、閾値処理で 2 値化することで、シルエット画像を得る。この時点では、人物の一部が遮蔽されていると、その部分が欠損したシルエット画像が得られる。

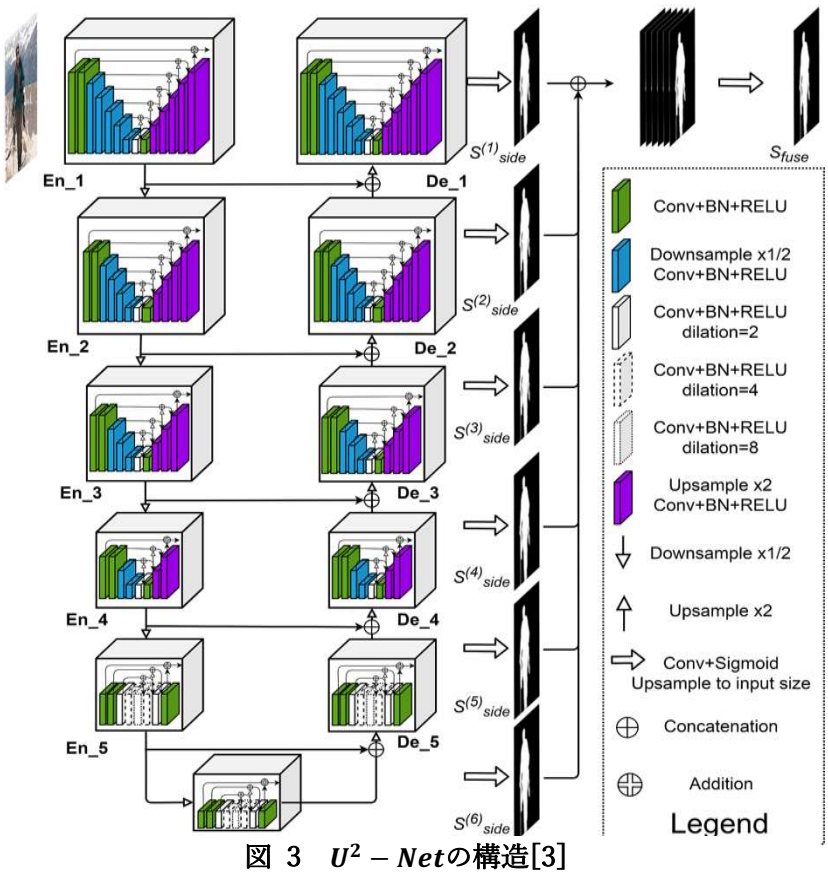


図 3　$U^2 - Net$の構造[3]

3.3　遮蔽部分の補完

本研究では、遮蔽により一部欠損した人体像を補完するために、画像補完を適用する。画像補完手法としては、Generative Multi-column Convolutional Neural Networks(GMCNN)を利用する。

GMCNN のネットワーク構造を図 4 に示す。入力として、欠損した画像と欠損部分のマスク領域を与える。最初の 3 つのサブネットワークは全て畳み込みニューラルネットワーク(CNN)で構成されている。この 3 つのネットワークは、並列に別れており、それぞれフィルタサイズが異なる。入力画像 2 つをこのネットワークに与え、通常の畳み込みを行った後に逆畳み込みを行う。この処理によって、様々なレベルの特徴量を取ってくることができる。この 3 つのネットワークの出力を目的の画像サイズになるように復元し、連結をする。そして、共通のデコーダーに先程生成したものを入力として与える。その結果を、improved Wasserstein GAN[6]に与えることによって、精度の高い画像補完を行なっている。

GMCNN では、学習した対象と同じ種類の物体の補完は精度良く行える。しかし手元で学習を行なったところ良い精度の画像補完が行えなかった。これに対して、2 値画像であるシルエット画像に対する画像補完は、比較的良好な結果を示した。遮蔽により一部欠損した元画像を画像補完した上でシルエット抽出を適用する手順も考えられるが、この場合、上記の理由から精度良く画像補完、シルエット抽出が行えない。そこで本研究では、学習済みモデル[7]を利用し、シルエット抽出により得られたシルエット画像と、元画像それぞれに対して画像補完を適用し、遮蔽部分を補完した補完元画像、補完シルエット画像を得る。

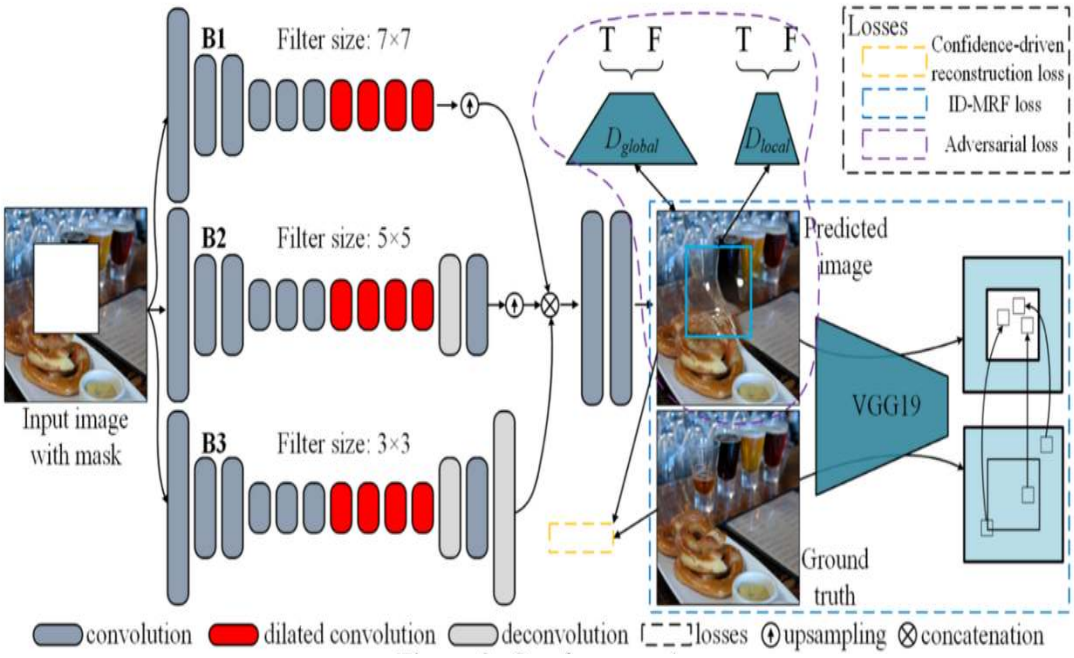


図 4　GMCNN のネットワーク構造[2]

3.4　人体 3 次元形状推定

本研究では、 3 次元形状推定手法として Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization(PIFu)を使用する。PIFu は、単一の画像から人体全周囲 360 度を推定して、 3 次元モデルを生成できる。また、人物像の服の色や模様を反映したテクスチャを 3 次元モデルに付与できる。

図 5 に PIFu の概略を示す。PIFu は、 3 次元形状を推定する部分(Surface Reconstruction:SR)と表面テクスチャを推定する部分(Texture Inferrence:TI)の 2 つで構成されている。

3 次元形状を推定する SR では、エンコーダ(Image encoder)を用いて、入力画像の各点の特徴量を抽出する。次に、得られた特徴量を使って、その点のある奥行きでの 3 次元位置が、人体の内部か外部かを判定する判定器を学習する。学習の際には、人体表面付近の奥行きを使うことが有効と考えられるが、人体表面付近のみでは過適合してしまう可能性がある。PIFu では、人体の表面を中心に正規分布状に重みつけられたサンプリング結果と、空間内を一様にサンプリングした結果とを 16:1 の割合で混合したものを学習データとして用いる。3 次元モデル生成時には、画像の各点について、判定器の出力が 0.5 になる奥行き位置を人体の表面とする。

テクスチャを推定する TI も同様に、エンコーダを用いて入力画像の各点の特徴量を抽出する。得られた特徴量を使って、その点のある奥行きでの 3 次元位置での色(RGB 値)を求めるテクスチャ関数を学習する。この際、物体表面以外の奥行きでは RGB 値はつかないため、物体表面の形状情報も必要となる。そのため、テクスチャ関数の学習時には、 3 次元形状を推定する SR の判定器も同時に利用する。また、物体表面のみではなく、一定距離の範囲内で RGB 値を推定するように学習する。

PIFu による 3 次元形状推定では、元画像と元画像中の人物領域を示すシルエット画像が必要となる。本研究では、元画像、シルエット画像として、3.3 節の方法で画像補完した補完元画像、補完シルエット画像を用いる。

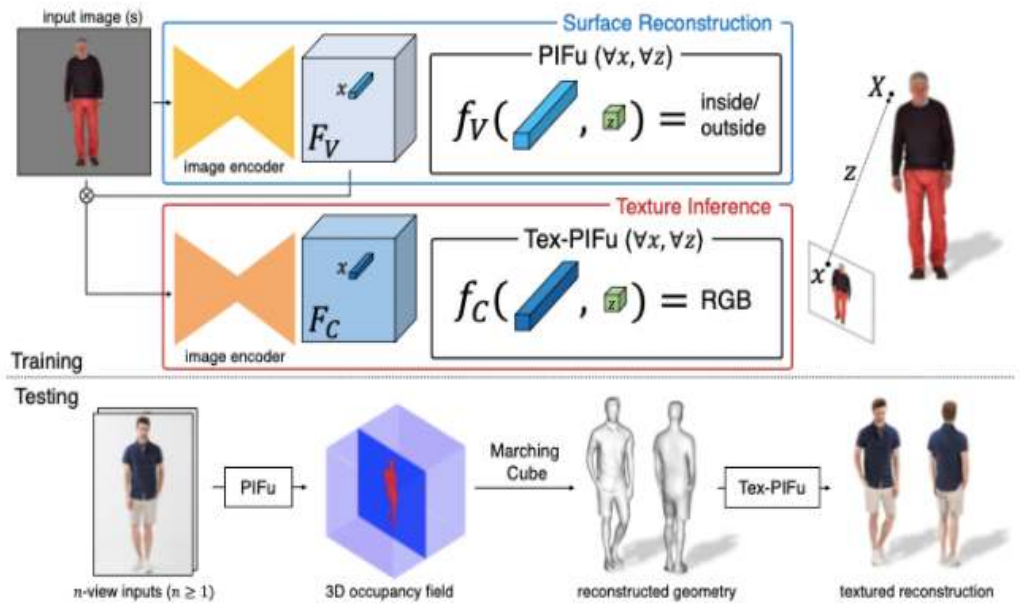

input image (s)
Surface Reconstruction
PIFu (∀x, ∀z)
inside/
outside
image encoder
Texture Inference
Tex-PIFu (∀x, ∀z)
RGB
image encoder
Training
Testing
PIFu
Marching
Cube
Tex-PIFu
n-view inputs (n ≥ 1)
3D occupancy field
reconstructed geometry
textured reconstruction

図 5　PIFu の概要[1]

4. 実験

4.1 実験設定

提案手法の評価のために、種々の画像を与えて処理結果を評価する。まず、シルエット抽出、遮蔽部分の補完、人体 3 次元形状推定のそれぞれの処理について評価する。その後、提案手法を使って遮蔽のある単一画像から人物の 3 次元形状を推定した結果を示す。

各処理では、学習済みの重みを与えて使用する。実装上、各処理で入力となる画像の大きさが決められているため、その条件に合うよう適宜、画像の拡大・縮小を行った。縦横比が条件に合わない場合には白画素の領域を追加(パディング)した。

4.2　シルエット抽出の評価実験

シルエット抽出の評価実験には、以下 5 種類の異なる画像(画像 1~5)を用いた。

画像 1:背景が単純な室内画像
画像 2:背景が複雑な室内画像
画像 3:背景が複雑な屋外画像
画像 4:背景が単純な屋外画像
画像 5:背景の一部が人体と似ている屋外画像

図 6 にシルエット抽出の結果を示す。各画像について、左が元画像、右がシルエット抽出の結果である。本研究では、人が持っているものも人体の一部とみなす。

(a) 画像 1

図 6　シルエット抽出結果

(b)　画像 2

(c)画像 3

(d)画像 4

図 6 つづき　シルエット抽出結果

（e）画像 5

図 6　つづきシルエット抽出結果

精度がよいのが、画像 1 と画像 2 である。画像 1 では、人体の服と背景の壁等の色が大きく離れている。画像 2 では、背景の壁に模様がありやや複雑であるが、人物の服の色とは大きく離れている。このように、人体と背景の色が大きく離れていると、前景と背景を分けることが簡単となり、よりくっきりと前景抽出ができる。それに対して、画像 3 では背景の木やネットと重なった銃の中央から先端部分の範囲が抽出されていない。複雑な背景に前景が溶け込んでしまったため、抽出されなかったと考えられる。画像 4 については、背景自体は比較的単純であるが、銃の先端部分が一部欠けている。画像 5 については、肌の色や足先など背景に類似した部分があるが、概ね正確に前景抽出できている。これらの結果から、本研究で適用した$U^2 - Net$による前景抽出は、実画像からのシルエット抽出に有効であることが分かった。

4.3　遮蔽部分の補完の評価実験

本研究では、遮蔽部分の補完のために、元画像および元画像から抽出したシルエット画像の両方に対して画像補完を適用する。いづれに対しても同じ学習モデルを使用している。元画像としては、シルエット画像の補完結果が良好であった前節の画像５を用いる。マスク領域(遮蔽部分)の場所により補完結果が影響を受けるか調査するために、実験では、太もも、肩、足先の３パターンの場所にマスク領域を設定して画像補完を適用する。図 7,8,9 に画像補完結果を示す。各図左上(a)が元画像で、その中の白い部分がマスク領域である。各図右上は元画像を補完した結果(補完元画像)である。各図左下(c)は(b)のシルエット画像、各図右下(d)が前節の手法で元画像から抽出したシルエット画像に画像補完を適用した結果(補完シルエット画像)である。

（ａ)元画像とマスク領域

（ｂ)補完元画像

(c)(ｂ)のシルエット画像

（ｄ)補完シルエット画像

図 7　画像補完結果(画像５:太もも)

（ａ）元画像とマスク領域

（ｂ）補完元画像

(c)（ｂ）のシルエット画像

（ｄ）補完シルエット画像

図 8 画像補完結果(画像 5:肩)

（ａ）元画像とマスク領域

（ｂ）補完元画像

(c)（ｂ）のシルエット画像

（ｄ）補完シルエット画像

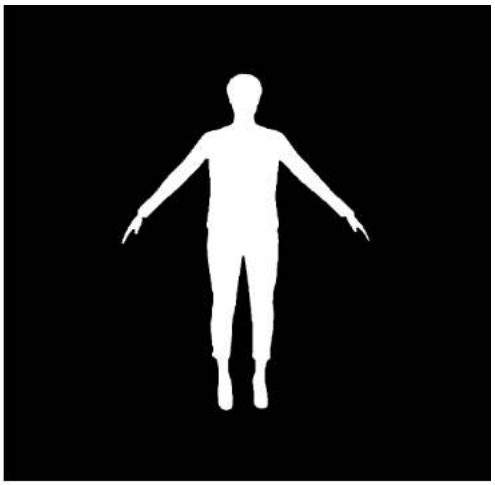
図 9 画像補完結果(画像 5:足先)

図 7 では左足の大部分をマスク領域として設定している。(b)補完元画像では、マスク領域がほぼ背景の色、テクスチャで補完されている。そのため、(b)からシルエット抽出した(c)では、マスク領域を補完した部分が前景として抽出できていない。一方、元画像と別に、シルエット画像に対して画像補完を適用した(d)では、マスク領域が違和感なく補完されている。このように、元画像とシルエット直像の両方に画像補完を利用することで、望ましいシルエットを抽出することが可能である。図 8 は肩の部分をマスク領域として設定しているが、図 7 と同様に適切な補完シルエット画像が得られている。一方、左足の足先部分をマスク領域として設定した図 9 では、補完シルエット画像の足先部分が欠けたままとなっている。マスク領域の大きさが同程度でも、手先や足先などの先端部分がマスク領域となっていると、選切に補完されない傾向がある。適切なシルエットを得るためには、足や腕のつながっている一部分が遮蔽されていることが前提条件になると考えられる。

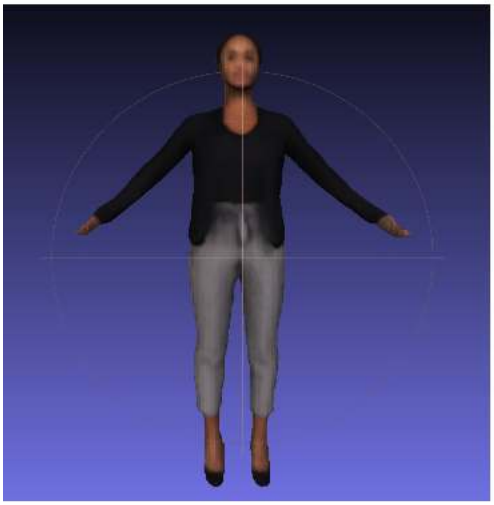
4.4　人体 3 次元形状推定の評価実験

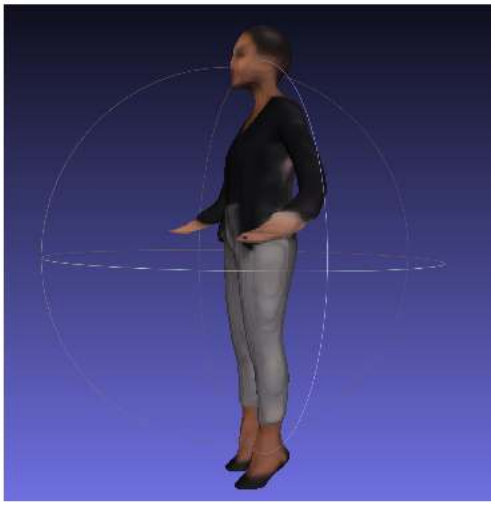
本研究では、PIFu を用いて人体 3 次元形状推定を行う。PIFu は入力として 3 次元形状推定したい元画像とその中の人体の存在範囲を示すシルエット画像を必要としている。本実験では、遮蔽によって画像に欠損がある場合、PlFu による 3 次元形状推定の結果にどの程度影響するかについて評価する。元画像として、RenderPeople[8]の人体 3 次元モデルを描画してキャプチャしたものを用いる。シルエット画像およびマスク領域は手動で与えている。図 10 に欠損(遮蔽)がない場合の結果、図 11 に欠損(遮蔽)がある場合の結果を示す。各図の上段左(a)が元画像、上段右(b)がシルエット画像、中段下段が 3 次元形状推定結果の 3 次元モデルを様々な方向から表示した結果である。図 11 で欠損は右足の脛の部分に設定している。

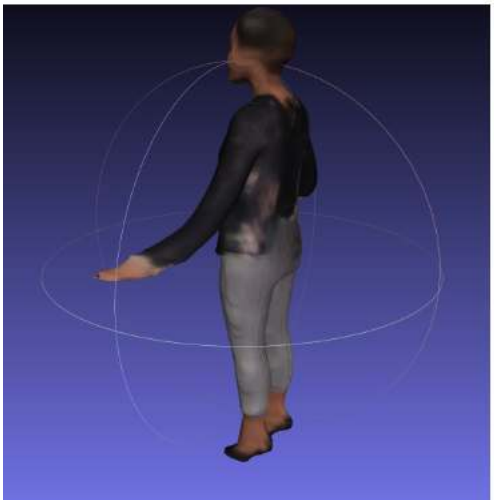
(a)　元画像



(b)　シルエット画像







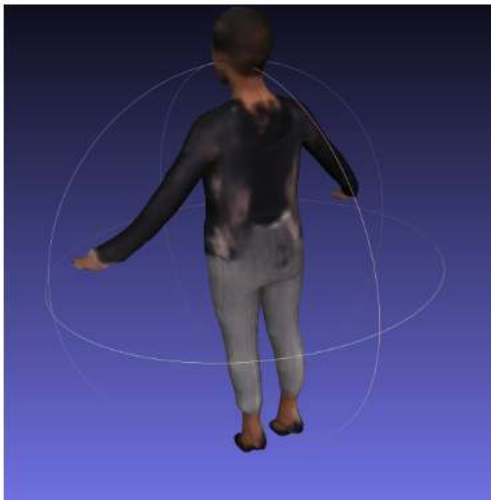
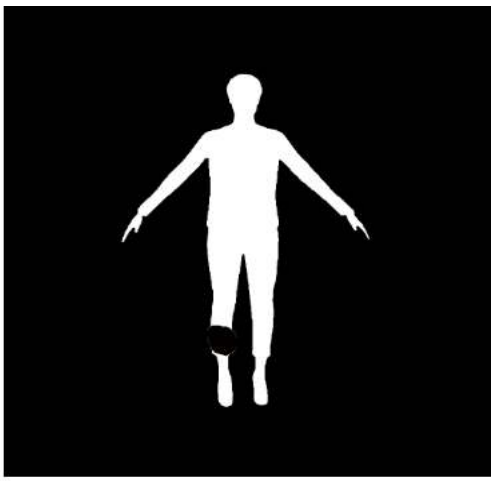
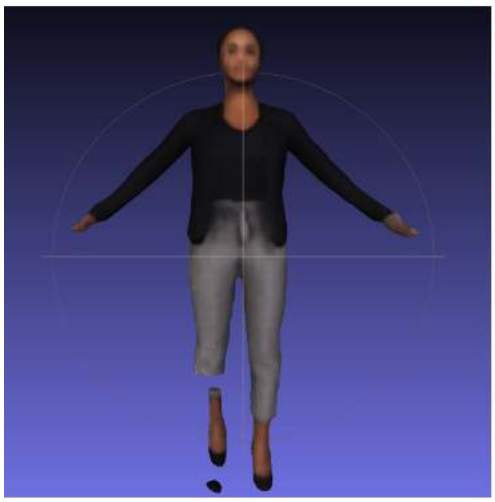


図 10　3 次元形状推定結果(欠損なし)

(a)　元画像



(b)　シルエット画像



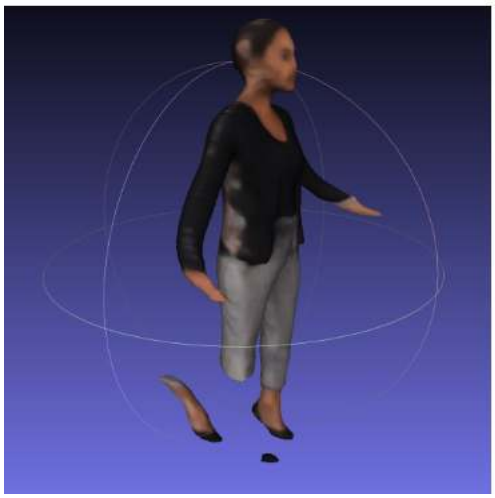


図 11　3 次元形状推定結果(欠損あり)

欠損がない図 10 では、いずれの部分も違和感なく 3 次元形状を推定できている。元画像には写っていない、側面、背面についても大きな違和感はない。一方、欠損がある図 11 では、脛より下の足先部分が人体から離れた位置に、脛より上の部分とズレて 3 次元形状推定されている。このように、欠損がある元像をそのまま PIFu により 3 次元形状推定すると、適切な結果が得られず、補完が必要であることが分かる。

4.5　提案手法の結果

図 12,13,14 に提案手法による 3 次元形状推定結果を示す。各図の上段(a)が元画像と設定したマスク領域、上段右(b)が補完シルエット画像、中段下段が 3 次元形状推定結果の 3 次元モデルを様々な方向から表示した結果である。

（ａ）元画像とマスク領域

（ｂ）補完シルエット画像

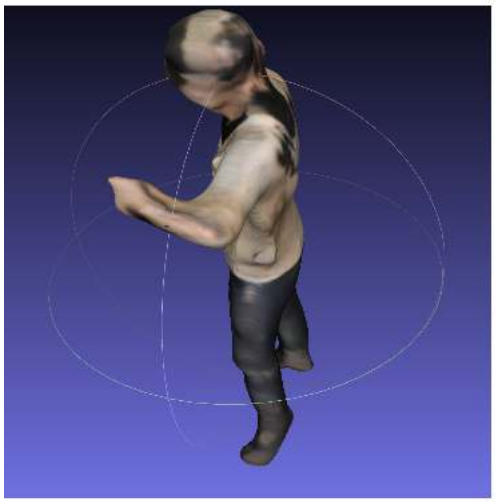


図 12　提案手法による３次元形状推定結果(画像 1)

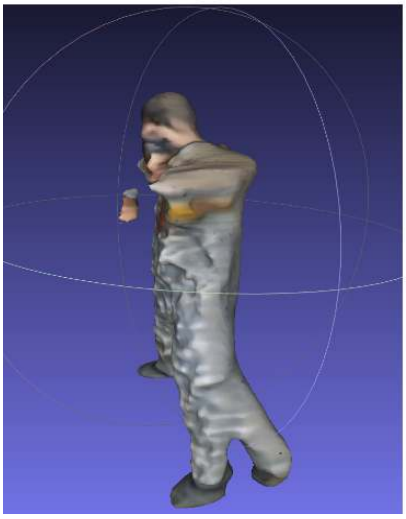
（ａ）元画像とマスク領域

（ｂ）補完シルエット画像

図 13 提案手法による３次元形状推定結果(画像 5)

（ａ）元画像とマスク領域

（ｂ）補完シルエット画像



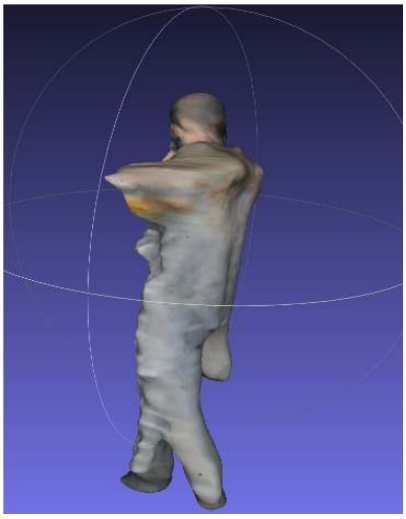


図 14 提案手法による３次元形状推定結果(画像 4)

図 12 では、肩の部分にマスク領域を設定している。補完シルエット画像では、肩の部分が適切に補完されている。補完元画像では、服のテクスチャではなく背景のテクスチャ(この場合、壁の白色)として補完されるため、推定結果でも肩の部分が白くなっているが、形状は適切に推定できている。その他の部分も含めて、違和感のない 3 次元形状推定が行えている。

図 13 では、左足の太ももの部分にマスク領域を設定している。マスク領域部分については、補完シルエット画像、3 次元形状推定結果のいずれについても違和感のない結果が得られている。一方、右手で支えている統も前景として抽出されているが、3 次元形状推定結果では、右足として 2 重に形状推定されている。本研究では、人物が持っているものも人体の一部分として前景抽出しているが、これにより、3 次元形状推定処理が影響を受ける場合があることが分かった。

図 14 では、人物が持っているもの(銃の先端部分)をマスク領域とした。補完シルエット画像では、この部分は背景とされた。3 次元形状推定結果でも、マスク領域に設定した銃の先端部分は削除され、人体の 3 次元形状推定に影響を与えなくなっている。このように、マスク領域を設定することで、人物が持っているものを削除して 3 次元形状推定を適切に行うという適用方法も有効と考えられる。

5.おわりに

本研究では、深層学習を用いた前景抽出、画像補完、3 次元形状推定という 3 つの技術を組み合わせることによって、遮蔽がある単一画像から、画像内の人体の 3 次元形状を推定できるか検証を行った。前景抽出に関しては、様々な背景の中でも比較的良好に人物の領域をシルエット画像として抽出できた。画像補完については、腕や足の途中部分が遮蔽されている場合は、シルエット画像の補完が適切に行えた。元画像の画像補完結果では、背景に近い色やテクスチャで補完されることがあったため、3 次元形状推定では、結果の 3 次元モデルのテクスチャに一部影響があったが、形状推定自体は違和感なく行えた。一方、腕先や足先の部分が遮蔽されている場合は、画像補完により先端部分まで補完することができなかった。

今後の課題として、シルエット画像の補完において、腕先や足先を適切に補完できる手法の検討があげられる。本研究では、画像補完に GMCNN を利用したが、他の画像補完手法により適切に画像補完が行えれば、より実用的な場面に提案手法を適用できる。また、本研究では各段階に深層学習を利用した一貫した処理を構築したが、実装上は、別々のネットワーク構造に分かれている。これらのネットワークを 1 つにまとめ、end-to-end で学習することで、全体の精度向上が図れるかも検討する必要がある。実験結果の評価では、生成された 3 次元モデルについて定性的な比較しか行えていない。正解となる 3 次元モデルと、その 3 次元モデルを使って生成した画像を提案手法で処理した結果を比較するなどして、定量的に提案手法の精度評価をすることも今後の課題である。また、本研究では人体を対象に 3 次元形状推定を行ったが、犬や猫など様々な対象を扱えるようにすることで、提案手法の活用範囲が広まっていくと考えられる。

謝辞

本研究を行うにあたり、多くの方々に協力をしていただきました。ご指導いただいた椋木雅之教授に感謝いたします。指導教員である椋木雅之教授には研究や論文作成に関して様々なご指導を頂きました。本当にありがとうございました。坂本教授、油田准教授には、副査を務めていただきました。お忙しい中、貴重な時間を割いて下さいましてありがとうございます。また、画像のデータをいただいた宮崎エアソフトガン競技、企画・運営協会の会長の平野さん、画像編集の許可をいただいた佐藤さん、中村さんには、大変感謝しております。椋木研究室の皆様には、研究を進めるにあたって、様々な助言を頂きました。特に、写真撮影などに協力いただいた兒玉君には大変お世話になりました。皆さんのおかげで楽しく、充実した研究生活を過ごすことができました。大変感謝しています。

参考文献

[1] Saito, S., Huang, Z., Natsume, R., Morishima, S., Li, H., & Kanazawa, A. PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization.”,Proc. International Conference on Computer Vision, pp. 2304-2314,2019

[2] Yi Wang, Xin Tao, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia, “Image Inpainting via Generative Multi-column Convolutional Neural Networks,”,Proc. Neural Information Processing Systems, 2018

[3] Qin, Xuebin, et al. "U^2-Net: Going deeper with nested U-structure for salient object detection." Pattern Recognition 106 :107404,2020

[4] O.Ronneberger, P.Fischer, T.Brox, “U-net: Convolutional networks for biomedical image segmentation”, Lecture Notes in Computer Science, Vol. 9351, pp. 234-241, 2015

[5]U^2-Netの学習済みモデル,https://drive.google.com/file/d/1-Yg0cxgrNhHP-016FPdp902BR-kSsA4P/view?usp=sharing

[6] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, Aaron Courville, "Improved Training of Wasserstein GANs",arXiv:1704.00028, 2017

[7]GMCMM の学習モデル,https://drive.google.com/file/d/1aakVS0CPML_Qg-PuXGE1Xaql96hNEKOU/view?usp=sharing

[8]RenderPeople https://renderpeople.com/free-3d-people/