論　文

物体クラスラベルと微量の前景マスクによる前景・姿勢・種類の同時学習を行うディープニューラルネットワーク

米田　駿介[†a)]　　入江　豪[††]　　西山　正志[†]　　岩井　儀雄[†]

Deep neural network for simultaneous learning of segmentation, pose estimation, and object recognition using annotation of object class labels and a micro amount of foreground masks

Shunsuke YONEDA[†a)], Go IRIE[††], Masashi NISHIYAMA[†], and Yoshio IWAI[†]

あらまし　物流現場の人手不足解消を目的とした物体ピッキングシステムの実現のため，物体画像に対して前景推定，姿勢推定，クラス推定の 3 つのタスクについて，精度よく推論する手法が必要となる．近年では各タスクをディープニューラルネットワークで学習する様々な手法が提案されている．各タスクの推定精度を高めるには，訓練サンプルに大量の教師信号を付与することが望ましいが，人手による教師信号の付与は非常に手間がかかる．特に前景推定タスクを学習するための前景マスクと，姿勢推定タスクを学習するための姿勢パラメータとでは，それらの付与コストが非常に大きい．そこで本論文では，物体クラスラベルと微量の枚数の前景マスクのみを用いることで，物体画像に対して前景推定，姿勢推定，クラス推定の 3 つのタスクを同時に学習し，高精度に推論する新たなディープニューラルネットワークについて述べる．評価実験を行った結果，教師信号である前景マスクを微量の枚数だけ加えるという前提条件を満たす場合，既存手法と比べて，提案手法は全てのタスクを同時に学習でき，かつ高精度に推論できることを確認した．

キーワード　前景推定，姿勢推定，物体認識，教師信号，付与コスト削減

1. はじめに

近年，物流の人手不足を解消するため，倉庫でのピッキング作業を自動化する技術が強く求められている．自動化技術の 1 つとして，カメラとロボットアームとを用いた物体ピッキングシステム [1] が開発されている．このシステムでは，コンピュータビジョン技術を用いて対象物体の前景マスク，姿勢パラメータ，物体クラスラベルを自動的に推定し，ロボットアームを制御して対象物体のピッキングを行う．前景マスク，姿勢パラメータ，物体クラスラベルが正確に推定されていなければ，ロボットアームは対象物体をピッキングすることができなくなる．したがって前景推定，姿勢推定，クラス推定の 3 つのタスクを精度よく推論する手法を検討する必要がある．

前景推定，姿勢推定，クラス推定の 3 つのタスクをそれぞれ異なるディープニューラルネットワークで学習させる手法 [2]〜[6] がこれまでに多く提案されている．各タスクをネットワークに学習させるには，訓練サンプルに対してタスクごとに異なる教師信号を付与する必要がある．具体的には，前景推定タスクでは，画素ごとにラベル付けされた前景マスクが必要となる．姿勢推定タスクでは，物体の回転や並進などの度合いを数値で表した姿勢パラメータが必要となる．また，クラス推定タスクでは，画像中の物体がどの種類であるかを示す物体クラスラベルが必要となる．ネットワークの推定精度を高めるには，学習に用いる訓練サンプルに教師信号を大量に付与することが望ましい．しかし，人手による教師信号の付与は非常に手間がかかる．特に，前景マスクと姿勢パラメータとの付与コストは非常に大きい．

† 鳥取大学大学院工学研究科
Graduate School of Engineering, Tottori University
†† 東京理科大学
Faculty of Engineering, Tokyo University of Science
a) E-mail: yoneda@tottori-u.ac.jp
DOI:10.14923/transinfj.???????????

姿勢パラメータの付与コスト削減を目的として，姿勢推定タスクとクラス推定タスクとを同じネットワークで学習する弱教師あり手法である Joint Object recognition and Pose estimation (JOP) [7] が提案されている．JOP では，支配的側面と呼ばれるアイデアに基づいた見え損失を用いることで，姿勢パラメータを付与することなくネットワークを学習する．また，前景マスクの付与コスト削減を目的として，前景推定タスクとクラス推定タスクとを同じネットワークで学習する弱教師あり手法である Self-Supervised Equivariant Attention Mechanism (SEAM) [8], Pixel-to-Prototype Contrast (PPC) [9], Transformer-based Weakly-supervised learning framework (WeakTr) [10] が提案されている．これらの既存手法では，クラス推定タスクの学習過程で生成される Class activation maps (CAM) を用いることで，前景マスクを付与することなくネットワークを学習する．前景マスクと姿勢パラメータとの付与コストを削減しつつ前景推定，姿勢推定，クラス推定の 3 つのタスクを推論する単純なアイデアとして，姿勢推定タスクを推論できる JOP と，前景推定タスクを推論できる SEAM，PPC もしくは WeakTr とを組み合わせることが考えられる．しかしこの場合，前景推定タスクと姿勢推定タスクとを別々のネットワークで学習するため，3 つのタスクを同時に学習できず，互いのネットワークの学習内容が独立になり，推論時に推定精度が得られない課題が発生する．

そこで本論文では，教師信号である前景マスクを微量の枚数だけ加えるという前提条件の下で，前景マスクと姿勢パラメータとの付与コストを削減しつつ，前景推定，姿勢推定，クラス推定の 3 つのタスクを同時に学習し，かつ精度よく推論する弱教師あり手法について述べる．提案手法では，既存手法である JOP のネットワークをベースとして，前景推定タスクのためのネットワーク層と形状損失とを新たに導入している．形状損失は，前景マスクと姿勢パラメータと物体クラスラベルとの間に成り立つ一貫性というアイデアに基づき，前景推定タスクを学習するために用いる．形状損失の計算には事前に付与された前景マスクを用いるが，その枚数は学習全体で用いる画像枚数の 0.15% 程度であるため，前景マスクの付与コストを大幅に削減することができる．また JOP と同様に見え損失を用いて姿勢推定タスクを学習するため，姿勢パラメータの付与コストをなくすことができる．提案手法の精度を評価するため，倉庫で扱われる物体データセットを用いて評価実験を実施した．実験結果から，教師信号である前景マスクを微量の枚数だけ加えるという前提条件を満たす場合，前景推定，姿勢推定，クラス推定の 3 つのタスクにおいて，提案手法は既存の弱教師あり手法と比べて，全てのタスクを同時に学習でき，かつ高い精度で推論できることを確認した．

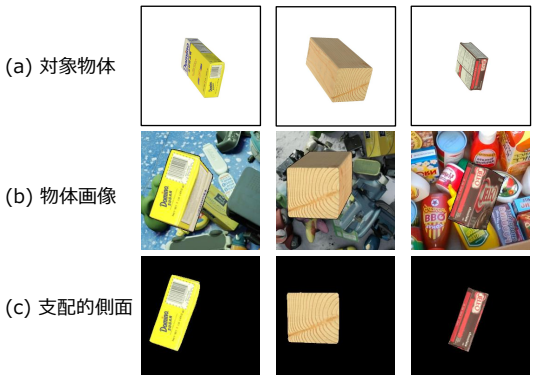


図 1　支配的側面の例.
Fig. 1　Examples of the dominant plane.

2. 既存手法

2.1 姿勢推定タスクとクラス推定タスクの同時学習

提案手法のベースとなる JOP [7] は，1. で述べたように，物体クラスラベルのみを教師信号として用いて，姿勢推定タスクとクラス推定タスクとを同時に学習する弱教師あり手法である．JOP について説明する前に，JOP において重要なアイデアである支配的側面について紹介する．

商品箱や書籍のように倉庫で扱われる物体は少数の側面で構成される．例えば一般的な商品箱は立方体であることが多く，図 1 (a) に示すように 6 個の側面で構成される．この対象物体を多様な物体が配置された状況下で撮影すると，(b) に示すような物体画像が得られる．対象物体を構成する側面のうち，画像中で最も大きな面積を占めるものが支配的側面と呼ばれる．図中 (c) に示すように，物体画像 1 枚につき，支配的側面 1 個が自ずと決まる．

支配的側面を利用して，JOP ではトリプレット学習 [11] と Spatial Transfer Networks (STN) [12] とを組み合わせたネットワークを学習させる．トリプレット学習とは，クラス推定タスクで用いるアンカー，そのアンカーと同クラスのポジティブ，そのアンカーと別

クラスのネガティブを入力として用いる学習手法である．JOP では物体画像 3 枚の組をアンカー，ポジティブ，ネガティブとして用いる．トリプレット学習のために，各画像に含まれる支配的側面の物体クラスラベルを教師信号として付与する．なお JOP では，各画像の背景は事前に取り除かれていることを仮定している．また学習全体において，アンカー用の画像は支配的側面 1 個につき 1 枚ずつ用意することを仮定している．アンカー用の画像枚数は物体側面の個数と等しいため，学習全体での画像枚数に対してアンカー用の画像枚数は微量となる．次に，STN について述べる．STN は，画像間での物体の位置関係を表す射影変換行列を，姿勢パラメータとして推定するネットワークである．JOP では，ポジティブとアンカーとで支配的側面の見え方が一致するように射影変換行列を推定し，その行列を用いてポジティブを変換する．

JOP ではネットワークの学習のため，見え損失と種類損失とをそれぞれ最小化する．なお，見え損失は JOP で提案された損失関数である．射影変換されたポジティブとアンカーとのそれぞれに含まれる同クラスの支配的側面の見え方の差から計算される．見え損失を用いることで，姿勢パラメータを教師信号として用いることなく姿勢推定タスクをネットワークに学習させることができる．種類損失は，畳み込み層とプーリング層とを用いて抽出された各画像の特徴ベクトルから計算される．特徴ベクトル空間上において，同クラスであるアンカーとポジティブとの距離が近いほど，また別クラスであるアンカーとネガティブとの距離が遠いほど小さな値をとる．種類損失を用いることで，クラス推定タスクをネットワークに学習させることができる．

既存手法の JOP では，学習に用いる画像の背景が事前に取り除かれていることを仮定しているため，そのままでは前景推定タスクを学習できない．そこで提案手法では，前景マスクの付与コスト削減を目的として，前景推定タスクとクラス推定タスクとを同時に学習する弱教師あり手法を JOP に導入することを考える．

2.2 前景推定タスクとクラス推定タスクの同時学習

物体クラスラベルのみを教師信号として用いて前景推定タスクとクラス推定タスクとを同時に学習する弱教師あり手法として，1. で述べたように，SEAM [8], PPC [9], WeakTr [10] が存在する．これらの既存手法では前景推定タスクの学習において，クラス推定タスクの学習過程で生成される CAM を利用する．CAM は，クラス推定タスクにおいてネットワークが注目する物体特徴を表現した重み行列である．CAM を擬似的な前景マスクとして用いることで，SEAM は物体クラスラベルのみを教師信号として用いて前景推定タスクを学習する．PPC では CAM に加えて AffinityNet [13] と前景推定ネットワークとを組み合わせることで前景推定タスクを学習する．WeakTr では Vision Transformer [14] ベースのエンコーダで CAM を生成し，それを用いて Vision Transformer ベースのデコーダネットワークによって前景推定タスクを学習する．

前景推定，姿勢推定，クラス推定の 3 つのタスクで推論する単純な方法として SEAM，PPC もしくは WeakTr と，2. 1 で述べた JOP [7] とを組み合わせることが考えられる．しかし 1. で述べたように，前景推定タスクと姿勢推定タスクとの学習を別々のネットワークで行うため，3 つのタスクを同時に学習できず，推論時に推定精度が得られないことが想定される．そこで提案手法では，前景マスクと姿勢パラメータと物体クラスラベルとの関係性に基づいた新たな損失関数を導入することで，3 つのタスクを同時に学習可能なネットワークを設計する．これにより，既存の弱教師あり手法 [7]〜[10] と比べて，提案手法は全てのタスクを高精度に推論することが期待できる．

3. 提 案 手 法

3.1 概　　要

本論文では，2. 1 で述べた JOP [7] のネットワークをベースとして，前景推定タスクのためのネットワーク層と損失関数とを新たに導入することで，前景推定，姿勢推定，クラス推定の 3 つのタスクの同時学習が可能な弱教師ありネットワークを設計する．提案手法では支配的側面のアイデアを発展させ，前景マスクと姿勢パラメータと物体クラスラベルとの一貫性というアイデアに着目する．この一貫性のアイデアに基づき，前景推定タスクのための損失関数である形状損失を新たに設計する．

一貫性のアイデアを説明するため，支配的側面の見え方について再考する．図 1 (c) に示すように，少数の側面で構成される物体の画像を撮影すると，画像 1 枚につき支配的側面 1 個が観測される．画像中の物体の見え方は，支配的側面の前景マスクと姿勢パラメータと物体クラスラベルとによって自ずと決まる．この場合，支配的側面の前景マスクと姿勢パラメータと物体クラスラベルとの間には一貫性が成り立つ．ここで

の一貫性とは，例えば入力画像における支配的側面の見え方と物体クラスラベルとが与えられば，その前景マスクと姿勢パラメータとは自ずと決まることを意味する．

上記した一貫性のアイデアを踏まえると，JOP は姿勢パラメータと物体クラスラベルとの一貫性を利用してネットワークを学習する手法と見なすことができる．JOP では，支配的側面の見え方と物体クラスラベルとのペアを学習に用いるため，姿勢パラメータと物体クラスラベルとの一貫性から姿勢パラメータは自ずと決まる．支配的側面の見え方と物体クラスラベルとを用いる見え損失によって，JOP では姿勢パラメータを与えることなく姿勢推定タスクをネットワークに学習させることができる．提案手法ではこの考え方を発展させ，前景マスクと物体クラスラベルとの一貫性を利用した損失関数を新たに設計することで，前景推定タスクをネットワークに学習させる．

具体的には，JOP のネットワークに前景推定層と形状損失とを新たに導入することで，姿勢推定タスクとクラス推定タスクとに加えて前景推定タスクを同時学習するネットワークを設計する．JOP のネットワークを拡張しているため，提案手法においても JOP と同様にアンカー，ポジティブ，ネガティブの 3 組の画像を入力とする．トリプレット学習のために，各画像に含まれる支配的側面の物体クラスラベルを教師信号として付与する．提案手法は JOP とは異なり，ポジティブとネガティブの背景を事前に取り除く必要がないという利点がある．学習全体において，アンカー用の画像は支配的側面 1 種類につき 1 枚ずつ用意することを仮定している．更にネットワークを安定に学習するために，アンカー用の画像に対して支配的側面の前景マスクを付与する．アンカー用の画像枚数は物体側面の個数と等しいため，学習全体での画像枚数に対して必要な前景マスク枚数は微量となる．これらの工夫によって，物体クラスラベルと微量の枚数の前景マスクとを用いて形状損失を計算し，前景推定タスクをネットワークに学習させることができる．

本論文で新たに導入する形状損失は，前景推定層で推定されたポジティブの前景マスクと，アンカーに付与された前景マスクとの形状の差から計算される．ポジティブの前景マスクの形状が，同クラスであるアンカーの前景マスクの形状に近いほど小さい値を返す．形状損失の値が十分に小さい場合，前景推定層によってポジティブから正確な前景マスクが推定されていることを意味する．

3.2 学習時のネットワーク

図 2 に提案手法の学習時におけるネットワーク構造を示す．学習時には，アンカー $\boldsymbol{I}^a$，ポジティブ $\boldsymbol{I}^p$，ネガティブ $\boldsymbol{I}^n$ の 3 組の画像を入力として用いる．アンカー $\boldsymbol{I}^a$ とポジティブ $\boldsymbol{I}^p$ とには同じ支配的側面が含まれており，それぞれの姿勢は異なるものと仮定する．もしポジティブ $\boldsymbol{I}^p$ の中に存在する物体が大きく回転し，画像中で見ている支配的側面が変化した場合，その見えている支配的側面に対応するアンカー $\boldsymbol{I}^a$ を用いると仮定する．更にアンカー $\boldsymbol{I}^a$ には支配的側面のみが含まれ，ポジティブ $\boldsymbol{I}^p$ とネガティブ $\boldsymbol{I}^n$ には他の側面や背景が含まれるものと仮定する．これらの仮定により，訓練サンプルの画像収集時に一手間が発生する．具体的には，1 つの物体に属するそれぞれの側面について，アンカー $\boldsymbol{I}^a$ とポジティブ $\boldsymbol{I}^p$ とを撮影することになる．アンカー $\boldsymbol{I}^a$ を収集する際，側面に正対するようにカメラを配置し撮影する．立方体の場合，6 個の側面が 1 つの物体に属するため，アンカー $\boldsymbol{I}^a$ の撮影は 6 回となる．ポジティブ $\boldsymbol{I}^p$ を収集する際，ある側面が見える範囲でカメラを自由に移動させながら動画を撮影し，その動画の各時刻における画像を用いる．これを全ての側面に対して行い，ポジティブ $\boldsymbol{I}^p$ を収集する．

学習の流れとして，まず前景推定層 r() の Nested-Unet [2] を用いて，ポジティブ $\boldsymbol{I}^p$ から，支配的側面のみを含むポジティブ側面領域 $\boldsymbol{I}^p_r$ を推定する．ポジティブ側面領域 $\boldsymbol{I}^p_r$ の推定を式 (1) に示す．

$$\boldsymbol{I}^p_r = \mathrm{r}(\boldsymbol{I}^p; \hat{\boldsymbol{I}}^p_m) \tag{1}$$

前景推定層 r() では，ポジティブ前景マスク $\hat{\boldsymbol{I}}^p_m$ を内部的に推定し，ポジティブ $\boldsymbol{I}^p$ とポジティブ前景マスク $\hat{\boldsymbol{I}}^p_m$ とのアダマール積を計算することで，ポジティブ側面領域 $\boldsymbol{I}^p_r$ を推定する．同様に，前景推定層 r() でネガティブ前景マスク $\hat{\boldsymbol{I}}^n_m$ を内部的に推定し，ネガティブ $\boldsymbol{I}^n$ とネガティブ前景マスク $\hat{\boldsymbol{I}}^n_m$ とのアダマール積を計算することで，ネガティブ側面領域 $\boldsymbol{I}^n_r$ を推定する．

ポジティブ側面領域 $\boldsymbol{I}^p_r$ の推定後，STN s() を用いて，ポジティブ側面領域 $\boldsymbol{I}^p_r$ を射影変換したポジティブ変換領域 $\boldsymbol{I}^p_s$ を推定する．ポジティブ変換領域 $\boldsymbol{I}^p_s$ の推定を式 (2) に示す．

$$\boldsymbol{I}^p_s = \mathrm{s}(\boldsymbol{I}^p_r; \hat{\mathbf{H}}^p) \tag{2}$$

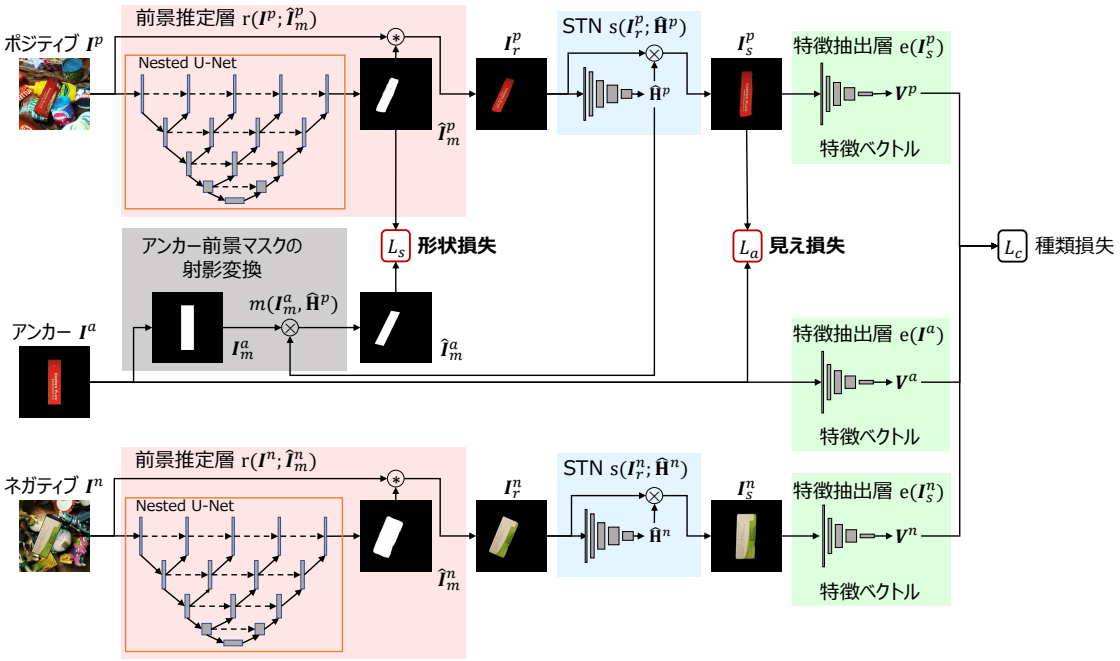


図 2　提案手法の学習時のネットワーク構造．形状損失 L_s を用いることで，ポジティブの前景マスク $\hat{I}^p_m$ の形状が疑似正解前景マスク $\hat{I}^a_m$ と一致するようにネットワークを学習する．見え損失 L_a を用いることで，射影変換後のポジティブ I^p_s の見え方がアンカー I^a と一致するようにネットワークを学習する．

Fig. 2　Overview of our network in the training process. By using the the shape loss L_s, the network is trained so that the shape of the positive foreground mask $\hat{I}^p_m$ matches the pseudo ground truth foreground mask $\hat{I}^a_m$. By using the appearance loss L_a, the network is trained so that the appearance of the transformed positive I^p_s matches the anchor I^a.

STN s() では，内部的に推定した射影変換行列 $\hat{\mathbf{H}}^p$ を用いて，ポジティブ側面領域 I^p_r を射影変換することで，ポジティブ変換領域 I^p_s を推定する．同様に，STN s() で内部的に推定した射影変換行列 $\hat{\mathbf{H}}^n$ を用いて，ネガティブ側面領域 I^n_r を射影変換することで，ネガティブ変換領域 I^n_s を推定する．

ポジティブ変換領域 I^p_s の推定後，特徴抽出層 e() として畳み込み層とプーリング層とを用いて，ポジティブ変換領域 I^p_s からポジティブ特徴ベクトル V^p を抽出する．特徴ベクトル V^p の抽出を式 (3) に示す．

$$V^p = \mathrm{e}(I^p_s) \tag{3}$$

特徴抽出層 e() では，ポジティブ変換領域 I^p_s からポジティブ特徴ベクトル V^p を抽出する．同様に，特徴抽出層 e() を用いてネガティブ変換領域 I^n_s からネガティブ特徴ベクトル V^n を抽出する．アンカー I^a においては，ポジティブ I^p やネガティブ I^n と異なり，特徴抽出層 e() を用いてアンカー特徴ベクトル V^a の抽出のみを行う．

最終的に，ネットワーク全体の損失関数 L を計算する．損失関数 L を式 (4) に示す．

$$L = L_s + \lambda_a L_a + \lambda_c L_c \tag{4}$$

ここで L_s は形状損失，L_a は見え損失，L_c は種類損失，λ_a は見え損失の重みとして与えるハイパーパラメータ，λ_c は種類損失の重みとして与えるハイパーパラメータである．

形状損失 L_s，見え損失 L_a，種類損失 L_c の計算について順に説明する．形状損失 L_s は，前景推定層 r() で内部的に推定したポジティブ前景マスク $\hat{I}^p_m$ の形状が，教師信号として与えられているアンカー前景マスク I^a_m の形状と一致するように前景推定層を学習させる．形状損失 L_s の計算の前処理として，アンカー前景マスク I^a_m に対して，ポジティブから推定された射影変換行列 $\hat{\mathbf{H}}^p$ を用いて射影変換を行う．これにより，ポジティブ前景マスク $\hat{I}^p_m$ に姿勢を近づけた疑似正解

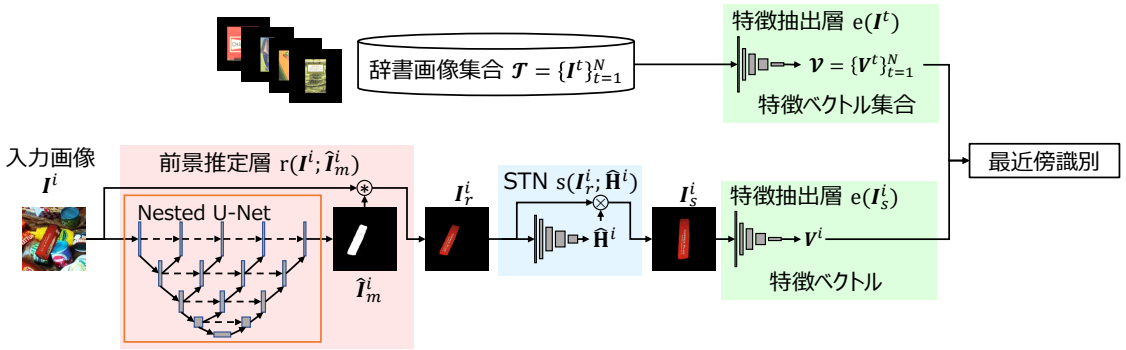


図 3　提案手法の推論時のネットワーク構造．3.2 で学習された前景推定層 r() と STN s() を用いて，入力画像 $\boldsymbol{I}^i$ の前景マスクと姿勢パラメータとを推定する．物体クラスラベルの推定は，特徴抽出層 e() を用いて辞書画像集合 $\mathcal{T} = \{\boldsymbol{I}^t\}_{t=1}^N$ から抽出した特徴ベクトル集合 $\mathcal{V} = \{\boldsymbol{V}^t\}_{t=1}^N$ と，入力変換領域 $\boldsymbol{I}_s^i$ から抽出した特徴ベクトル $\boldsymbol{V}^i$ との最近傍識別によって行う．

Fig. 3　Overview of our network in the inference process. The foreground mask and pose parameter of the input image $\boldsymbol{I}^i$ are estimated by the segmentation layer r() and STN s() trained in section 3.2. The object recognition task is performed by nearest neighbor identification between a feature vector set $\mathcal{V} = \{\boldsymbol{V}^t\}_{t=1}^N$ extracted from a target image set $\mathcal{T} = \{\boldsymbol{I}^t\}_{t=1}^N$ and a feature vector $\boldsymbol{V}^i$ extracted from the input image $\boldsymbol{I}^i$ by feature extraction layer e().

前景マスク $\hat{\boldsymbol{I}}_m^a$ を生成する．疑似正解前景マスク $\hat{\boldsymbol{I}}_m^a$ の生成を式 (5) に示す．

$$\hat{\boldsymbol{I}}_m^a = \mathrm{m}(\boldsymbol{I}_m^a, \hat{\mathbf{H}}^p) \tag{5}$$

ここで $\mathrm{m}(\boldsymbol{I}_m^a, \hat{\mathbf{H}}^p)$ は，ポジティブから推定された射影変換行列 $\hat{\mathbf{H}}^p$ を用いてアンカー前景マスク $\boldsymbol{I}_m^a$ を射影変換し，疑似正解前景マスク $\hat{\boldsymbol{I}}_m^a$ を生成する関数である．疑似正解前景マスク $\hat{\boldsymbol{I}}_m^a$ を生成した後，ポジティブ前景マスク $\hat{\boldsymbol{I}}_m^p$ を用いて形状損失 L_s を計算する．形状損失 L_s の計算を式 (6) に示す．

$$L_s = ||\hat{\boldsymbol{I}}_m^p - \hat{\boldsymbol{I}}_m^a||_2^2 \tag{6}$$

見え損失 L_a は，ポジティブ変換領域 $\boldsymbol{I}_s^p$ の見え方がアンカー $\boldsymbol{I}^a$ と一致するように STN を学習させる．見え損失 L_a の計算を式 (7) に示す．

$$L_a = ||\boldsymbol{I}_s^p - \boldsymbol{I}^a||_1 \tag{7}$$

種類損失 L_c は，同クラスのアンカー特徴ベクトル $\boldsymbol{V}^a$ とポジティブ特徴ベクトル $\boldsymbol{V}^p$ との距離が近く，別クラスのアンカー特徴ベクトル $\boldsymbol{V}^a$ とネガティブ特徴ベクトル $\boldsymbol{V}^n$ との距離が遠ざかるように特徴抽出層 e() を学習させる．種類損失 L_c の計算を式 (8) に示す．

$$L_c = \max(||\boldsymbol{V}^p - \boldsymbol{V}^a||_2^2 - ||\boldsymbol{V}^n - \boldsymbol{V}^a||_2^2 + \epsilon,\ 0) \tag{8}$$

ここで ϵ はマージンとして与えるハイパーパラメータである．

3.3　推論時のネットワーク

図 3 に既存手法の推論時のネットワークを示す．推論時のネットワークでは，3.2 で学習された前景推定層 r()，STN s()，特徴抽出層 e() を用いて，入力画像 $\boldsymbol{I}^i$ に対して前景マスク，姿勢パラメータ，物体クラスラベルの推定を行う．クラス推定タスクを行うために，提案手法では辞書画像集合 $\mathcal{T} = \{\boldsymbol{I}^t\}_{t=1}^N$ を利用する．辞書画像 $\boldsymbol{I}^t$ は支配的側面のみを含む画像であり，支配的側面の物体クラスラベルを持つ．N は支配的側面の個数である．推論の流れとして，まず前景推定層 r() を用いて入力画像 $\boldsymbol{I}^i$ から入力前景マスク $\hat{\boldsymbol{I}}_m^i$ を内部的に推定し，入力画像 $\boldsymbol{I}^i$ と入力前景マスク $\hat{\boldsymbol{I}}_m^i$ のアダマール積を計算することで，入力側面領域 $\boldsymbol{I}_r^i$ を推定する．次に STN s() で内部的に推定した射影変換行列 $\hat{\mathbf{H}}^i$ を用いて入力側面領域 $\boldsymbol{I}_r^i$ を射影変換することで，入力変換領域 $\boldsymbol{I}_s^i$ を推定する．そして特徴抽出層 e() を用いて辞書画像集合 $\mathcal{T}$ から抽出した特徴ベクトル集合 $\mathcal{V} = \{\boldsymbol{V}^t\}_{t=1}^N$ と，入力変換領域 $\boldsymbol{I}_s^i$ から抽出した特徴ベクトル $\boldsymbol{V}^i$ との最近傍識別を行う．特徴ベクトル $\boldsymbol{V}^i$ と最も類似した特徴ベクトルを持つ辞書画像の物体クラスラベルを入力画像 $\boldsymbol{I}^i$ に割り当てることで，物体クラスラベルを推定する．

図 4　YCB データセット，APC データセット，ARC データセットから選択した直方体または平面と見なす物体モデル．赤枠内の物体は直方体，オレンジ枠内の物体は平面である．

Fig. 4　Target object models from the YCB dataset, APC dataset, and ARC dataset used in our experiments. The color frame indicates the type of object shape (red: cuboid, orange: plane).

図 5　実験に使用した乱雑背景画像の例．

Fig. 5　Examples of the clutter background condition used in our experiments.

4. 実　　験

4.1 データセット

本論文では，3 次元物体モデルの公開データセットである YCB [15]，APC [16]，ARC [17] から，直方体または平面と見なすことができる物体モデルを用いて，学習時の訓練サンプルと推論時の入力画像とを生成した．直方体物体は 13 種類，平面物体は 5 種類で合計 18 種類の物体モデルを使用した．それらの物体モデルを図 4 に示す．直方体 1 つにつき側面が 6 種類，平面物体 1 つにつき裏表で側面が 2 個存在するため，支配的側面の個数は合計で $N = 13 \cdot 6 + 5 \cdot 2 = 88$ 個となった．本論文では支配的側面 1 個を 1 つのクラスとして扱うため，物体クラスラベルの種類数は支配的側面の個数と同じく 88 種類とした．物体画像の背景として，乱雑背景画像を用いた実験を行った．乱雑背景画像として，3 次元物体モデルの公開データセットである Household Objects for Pose Estimation (HOPE) [18] と HomebrewedDB (HB) [19] とで提供されている画像を使用した．HOPE 背景条件では，図 5 (a) に示すように HOPE データセットに含まれる 28 種類の物体モデルがランダムに配置された提供画像から 6 枚を使用した．HB 背景条件では，図 5 (b) に示すように HB データセットに含まれる 33 種類の物体モデルがランダムに配置された提供画像から 8 枚を使用した．なお，HOPE データセットと HB データセットとは，YCB データセットと APC2016 データセットと ARC2017 データセットとのいずれとも異なる種類の物体モデルを含む．

学習時の訓練サンプルはアンカー $\boldsymbol{I}^a$，そのアンカーと同クラスで姿勢が異なるポジティブ $\boldsymbol{I}^p$，そのアンカーと別クラスのネガティブ $\boldsymbol{I}^n$ の 3 枚からなる画像組を 30000 組収集した．アンカー $\boldsymbol{I}^a$ の撮影では，黒色背景上に物体モデルを配置し，光軸が物体モデルの支配的側面の重心を通るようにカメラを配置した．アンカー $\boldsymbol{I}^a$ の枚数は，実験に用いる支配的側面の個数と等しいため，訓練サンプル数に関わらず 88 枚とした．ポジティブ $\boldsymbol{I}^p$ の撮影では，乱雑背景上にアンカー $\boldsymbol{I}^a$ と同じ支配的側面を含むように物体モデルを配置し，物体モデルに対して回転と平行移動と拡大縮小とを行った．回転角度の範囲は [-30, 30] 度，平行移動の範囲は [-150, 150] 画素，拡大縮小の範囲は [0.8, 1.2] 倍とした．これらのパラメータは画像の撮影ごとにランダムに決定した．ネガティブ $\boldsymbol{I}^n$ の撮影では，乱雑背景上にアンカー $\boldsymbol{I}^a$ と異なる支配的側面を含むように物体モデルを配置し，ポジティブ $\boldsymbol{I}^p$ と同様に物体モデルに対して回転と平行移動と拡大縮小とを行った．なお，提案手法の学習に用いる画像組 1 つにつきポジティブ $\boldsymbol{I}^p$ が 1 枚とネガティブ $\boldsymbol{I}^n$ が 1 枚とが含まれるため，画像組を 30000 組とアンカー $\boldsymbol{I}^a$ を 88 枚とを用いる場合の学習全体の画像枚数は $30000 \cdot 2 + 88 = 60088$ 枚となる．教師信号として用いる前景マスクはアンカー $\boldsymbol{I}^a$ にのみ付与するため，学習全体の画像枚数に対する前景マスク枚数の割合は 88/60088 となり 0.15% となる．推論のために，3000 枚の入力画像 $\boldsymbol{I}^i$ と辞書画像集合 $\mathcal{T} = \{\boldsymbol{I}^t\}_{t=1}^{N}$ を収集した．入力画像 $\boldsymbol{I}^i$ の撮影では，ポジティブ $\boldsymbol{I}^p$ と同様の手順で物体モデルを配置した．入力画像 $\boldsymbol{I}^i$ の撮影において，回転角度の範囲は [-30, 30] 度，平行移動の範囲は [-150, 150] 画素，拡大縮小の範囲は [0.8, 1.2] 倍とした．これらのパラメータは画像の撮影ごとにランダムに決定した．辞

表 1　提案手法と既存手法との精度比較．(a) から (e) は完全教師あり手法，(f) から (j) は弱教師あり手法である．n/a はそのタスクを学習しないことを意味する．

Table 1　Performance of our network and existing networks. (a)-(e) are the fully supervised networks and (f)-(j) are the weakly supervised networks. n/a means that the corresponding task is not trained.

	前景推定精度 ↑	教師数 (前景)	姿勢推定誤差 ↓	教師数 (姿勢)	クラス推定精度 ↑
(a) Nested-Unet [2]	$\mathbf{0.99 \pm 0.01}$	大量	n/a	n/a	n/a
(b) DeepLabV3 [3]	$\mathbf{0.99 \pm 0.01}$	大量	n/a	n/a	n/a
(c) HomographyNet [4]	n/a	n/a	0.20 ± 0.01	大量	n/a
(d) GeM [5]	n/a	n/a	n/a	n/a	0.96 ± 0.01
(e) SwinT [6]	n/a	n/a	n/a	n/a	$\mathbf{0.99 \pm 0.01}$
(f) SEAM [8]	0.20 ± 0.06	不要	n/a	n/a	0.98 ± 0.02
(g) PPC [9]	0.71 ± 0.04	不要	n/a	n/a	0.98 ± 0.01
(h) WeakTr [10]	0.50 ± 0.04	不要	n/a	n/a	$\mathbf{0.99 \pm 0.01}$
(i) JOP [7]	n/a	n/a	435.28 ± 112.93	不要	0.02 ± 0.01
(j) Ours	0.90 ± 0.01	微量	$\mathbf{0.12 \pm 0.01}$	不要	$\mathbf{0.99 \pm 0.01}$

書画像集合 $\mathcal{T} = \{\boldsymbol{I}^t\}_{t=1}^N$ のサンプル数 N は，3.3 で述べたように物体側面の個数と等しく，$N = 88$ とした．辞書画像 $\boldsymbol{I}^t$ の撮影では，アンカー $\boldsymbol{I}^a$ と同様の手順で物体モデルを配置した．なお学習時と推論時とのいずれにおいても，画像の大きさは 256×256 画素とした．

4.2 実験条件

前景推定層 r() には 5 層の Nested-Unet [2] を用いた．STN s() の層数は 5 層とした．特徴抽出層 e() には 6 個の畳み込み層とプーリング層を用いた．事前に行った予備実験の結果に基づき，式 (4) のハイパーパラメータは $\lambda_a = 10, \lambda_c = 1$，式 (8) のマージンは $\epsilon = 0.2$ とした．損失関数の最適化アルゴリズムとして Stochastic Gradient Descent (SGD) を用いた．SGD の学習率は 0.01，モーメンタムは 0.9 とした．学習時のエポック数は 100 とした．

提案手法の精度評価のため，以下の 3 種類の評価指標を用いて複数の既存手法と精度比較を行った．前景推定精度として，前景推定ネットワークで推定した前景マスクと正解前景マスクとの Intersection over Union (IoU) の平均を用いた．クラス推定精度として，最近傍識別による 1 位正解率を用いた．姿勢推定誤差として，STN [12] で推定した射影変換行列と正解射影変換行列との差のフロベニウスノルムの平均を用いた．なお前景推定精度と姿勢推定誤差との計算は，入力画像 $\boldsymbol{I}^i$ の物体クラスラベルを正しく推定できた場合のみ行うこととした．

4.3 前景推定精度の評価

前景推定タスクにおける提案手法の精度評価のため，複数の既存手法との間で前景推定精度を比較した．完全教師あり前景推定手法として Nested-Unet [2], DeepLabV3 [3] を使用した．弱教師あり前景推定手法として SEAM [8]，PPC [9]，WeakTr [10] を使用した．なお，これらの既存手法のハイパーパラメータは公開されているデフォルト値を使用した．結果を表 1 (a), (b), (f), (g), (i), (j) に示す．Nested-Unet と DeepLabV3 との前景推定精度と比べると，提案手法の前景推定精度は低かった．Nested-Unet と DeepLabV3 とは大量の前景マスクを教師信号として用いているのに対して，提案手法は微量の枚数の前景マスクを教師信号として用いていることが理由であると考えられる．一方，SEAM と PPC と WeakTr との前景推定精度と比べると，提案手法の前景推定精度は大きく上回っていた．2.2 で述べたように，SEAM と PPC と WeakTr とではクラス推定タスクの学習過程で生成される CAM を利用して前景推定タスクを学習する．しかし提案手法では，物体クラスラベルと微量の枚数の前景マスクとに加えて，STN で推定された射影変換行列を用いる形状損失によって前景推定タスクを学習する．SEAM と PPC と WeakTr とでは姿勢推定タスクを学習しないが，提案手法は姿勢推定タスクも同時に学習するため，提案手法の前景推定精度が向上したと考えられる．以上のことから，教師信号である前景マスクを微量の枚数だけ加えるという前提条件を満たす場合であれば，提案手法は既存の弱教師あり前景推定手法よりも高精度に前景推定タスクを推論できる．

4.4 姿勢推定誤差の評価

姿勢推定タスクにおける提案手法の精度評価のため，複数の既存手法との間で前景推定誤差を比較した．完全教師あり姿勢推定手法として，HomographyNet [4] を使用した．弱教師あり姿勢推定手法として，JOP [7]

を使用した．なお，これらの既存手法のハイパーパラメータは公開されているデフォルト値を使用した．結果を表 1 (c)，(h)，(j) に示す．HomographyNet や JOP の姿勢推定誤差と比べると，提案手法の姿勢推定誤差は最も低かった．2.1 で述べたように，JOP では背景が事前に取り除かれているという仮定を置いているため，背景が複雑なデータセットでは，姿勢推定誤差が極端に大きくなった．HomographyNet では，姿勢推定タスクのみを学習し，前景推定タスクやクラス推定タスクを学習していない．一方，提案手法では，前景推定層で推定された支配的側面のみを含む画像を用いて姿勢推定タスクを学習しているため，その姿勢推定誤差が小さくなったと考えられる．

4.5 クラス推定精度の評価

クラス推定タスクにおける提案手法の精度評価のため，複数の既存手法との間でクラス推定精度を比較した．クラス推定タスクのみを学習する既存手法として，GeM [5]，SwinTransformer (SwinT) [6] を使用した．前景推定タスクとクラス推定タスクとを学習する既存手法として，SEAM [8]，PPC [9]，WeakTr [10] を使用した．姿勢推定タスクとクラス推定タスクとを学習する既存手法として，JOP [7] を使用した．クラス推定タスクにおいては，提案手法を含む全手法が物体クラスラベルを教師信号として用いて学習した．なお，これらの既存手法のハイパーパラメータは公開されているデフォルト値を使用した．結果を表 1 (d) から (j) に示す．提案手法のクラス推定精度は，SwinT と WeakTr とのクラス推定精度と並んで最も高かった．この結果より，提案手法のクラス推定精度は十分であると考えられる．

4.6 前景推定タスクと姿勢推定タスクとの可視化

提案手法による前景推定タスクと姿勢推定タスクとの可視化結果を図 6 に示す．(a) の入力画像 $\boldsymbol{I}^i$ に対して，前景推定層で推定された入力前景マスク $\hat{\boldsymbol{I}}^i_m$ とのアダマール積を計算することで，(b) の入力側面領域 $\boldsymbol{I}^i_r$ が得られた．(b) の入力側面領域 $\boldsymbol{I}^i_r$ に対して，STN で推定された射影変換行列 $\hat{\mathbf{H}}^i$ を用いて射影変換を行うことで，(c) の入力変換領域 $\boldsymbol{I}^i_s$ が得られた．(c) の入力変換領域 $\boldsymbol{I}^i_s$ と (d) の辞書画像 $\boldsymbol{I}^t$ との間で，それぞれに含まれる支配的側面の見え方が似ていることから，前景推定タスクと姿勢推定タスクとにおいて，提案手法により精度よく推論されたと考えられる．

図 6　提案手法による前景推定タスクと姿勢推定タスクの可視化．

Fig. 6　Qualitative results of segmentation and pose estimation task using our network.

4.7 完全教師あり前景推定と弱教師あり姿勢推定兼クラス推定とを組み合わせた場合の性能評価

提案手法において，微量の枚数の前景マスクを教師信号として用いることの有効性を確認するため，既存手法との比較実験を行った．具体的には，前景マスクを教師信号とする完全教師あり前景推定手法である DeepLabV3 [3] と，物体クラスラベルを教師信号とする弱教師あり姿勢推定兼クラス推定手法である JOP [7] との組み合わせ（以下，DeepLabV3+JOP と表記）について，提案手法と同じ枚数の前景マスクを教師信号として用いた場合の性能を評価した．DeepLabV3 の学習では，提案手法と同じ枚数の前景マスクを教師信号としてファインチューニングを行った．具体的には，4.1 で述べた学習全体の画像枚数 60088 枚のうち 0.15% に相当する提案手法のアンカー $\boldsymbol{I}^a$ の 88 枚に付与された前景マスクを，DeepLabV3 の教師信号として用いた．また，60088 枚の学習画像全てに前景マスクを付与した上で，DeepLabV3 の学習に用いる前景マスクの枚数を 688 枚，6088 枚，60088 枚と増加させた場合についても性能を評価した．JOP では学習画像から背景領域が事前に取り除かれている必要があるため，DeepLabV3 で推定された前景マスクを用いて，学習画像から物体の支配的側面の領域を決定し，背景領域を取り除いた．なお，JOP の学習に用いる物体クラスラベルの個数を 60088 個と常に固定し，4.1 で述べた学習全体の画像枚数 60088 枚の全てを JOP の学習に使

表 2　提案手法と DeepLabV3 [3]+JOP [7] の精度比較.
Table 2　Performance of our network and DeepLabV3 [3]+JOP [7].

	前景マスク枚数	物体クラスラベル個数	前景推定精度 ↑	姿勢推定誤差 ↓	クラス推定精度 ↑
(a) DeepLabV3+JOP	88	60088	0.09 ± 0.01	2.58 ± 0.03	0.66 ± 0.01
(b) DeepLabV3+JOP	688	60088	0.91 ± 0.01	0.19 ± 0.01	0.91 ± 0.01
(c) DeepLabV3+JOP	6088	60088	0.98 ± 0.01	0.17 ± 0.01	0.94 ± 0.01
(d) DeepLabV3+JOP	60088	60088	**0.99 ± 0.01**	0.17 ± 0.01	0.94 ± 0.03
(e) Ours	88	60088	0.90 ± 0.01	**0.12 ± 0.01**	**0.99 ± 0.01**

表 3　提案手法と PPC [9]+JOP [7] の精度比較.
Table 3　Performance of our network and PPC [9]+JOP [7].

	前景推定精度 ↑	姿勢推定誤差 ↓	クラス推定精度 ↑
(a) PPC+JOP	0.71 ± 0.04	0.40 ± 0.01	0.81 ± 0.01
(b) Ours	**0.90 ± 0.01**	**0.12 ± 0.01**	**0.99 ± 0.01**

用した．上記以外の実験条件は 4.2 と同じとした．

既存手法の組み合わせである DeepLabV3+JOP の結果を表 2 に示す．まず教師信号である前景マスクを 88 枚用いてファインチューニングした DeepLabV3 と組み合わせることで，JOP の精度がどの程度改善されたかを確認した．表 1 (i) の JOP の結果と，表 2 (a) の DeepLabV3+JOP の結果とを比較すると，DeepLabV3+JOP では前景推定を実行可能としつつ，姿勢推定誤差とクラス推定精度とを大幅に改善したことがわかる．次に表 2 (e) の提案手法の性能と，表 2 (a) の前景マスクの枚数が 88 枚の場合の DeepLabV3+JOP の性能とを比較した．その結果，提案手法と同じ枚数の前景マスクを教師信号とした DeepLabV3 ＋ JOP では前景推定精度，姿勢推定誤差，クラス推定精度が全て低かった．同じ枚数だけの微量の前景マスクを教師信号として用いる場合，DeepLabV3+JOP よりも提案手法は，前景推定，姿勢推定，クラス推定を高精度に行えることがわかる．

更に表 2 (e) の提案手法の性能と，表 2 (b), (c), (d) の教師信号である前景マスクの枚数が 688 枚，6088 枚，60088 枚の場合の DeepLabV3+JOP の性能とを比較した．その結果，提案手法と比較して，前景マスク枚数が増加するにつれて，DeepLabV3 ＋ JOP では前景推定精度が改善されていった．多量の枚数の前景マスクを教師信号として用いることが可能であれば，DeepLabV3+JOP は前景推定を高精度に行えることがわかる．しかし，前景マスク枚数が増加したとしても，DeepLabV3+JOP の姿勢推定精度とクラス推定精度とにおいて，提案手法ほどの改善は見られなかった．DeepLabV3+JOP の学習に用いる前景マスク枚数が増加した場合と比べて，提案手法は前景推定の性能では劣るものの，姿勢推定とクラス推定との性能では上回ることがわかる．

4.8　弱教師あり前景推定兼クラス推定と弱教師あり姿勢推定兼クラス推定とを組み合わせた場合の性能評価

提案手法の有効性を確認するために，物体クラスラベルを教師信号とする弱教師あり前景推定兼クラス推定手法である PPC [9] と，物体クラスラベルを教師信号とする弱教師あり姿勢推定兼クラス推定手法である JOP [7] との組み合わせ（以下，PPC+JOP と表記）について，提案手法と性能を比較した．PPC の学習では，公開されている学習済みモデル (注2) に対して，クラス推定の教師信号である物体クラスラベルを，提案手法と同じ個数だけ用いてファインチューニングを行った．JOP では学習画像から背景領域が事前に取り除かれている必要があるため，PPC で推定された前景マスクを用いて，学習画像から物体の支配的側面の領域を決定し，背景領域を取り除いた．PPC+JOP の性能評価では，JOP で推定された物体クラスラベルを用いてクラス推定精度を算出した．これら以外の実験条件は 4.7 と同じとした．

既存手法の組み合わせである PPC ＋ JOP の結果を表 3 に示す．まず PPC と組み合わせることによって，JOP の精度がどの程度改善されたかを確認した．表 1 (i) の JOP の結果と，表 3 (a) の PPC+JOP の結果とを比較すると，PPC+JOP では前景推定を実行可能としつつ，姿勢推定誤差とクラス推定精度とを大幅に改善したたことがわかる．次に表 3 (b) の提案手法の性能と，表 3 (a) の PPC+JOP の性能とを比較する．その結果，PPC+JOP と比較して，提案手法は，前景推

(注1)：https://github.com/qubvel/segmentation_models

定精度，姿勢推定誤差，クラス推定精度おw全て改善していた．教師信号である前景マスクを微量枚数だけ加えるという前提条件を満たすのであれば，提案手法は，既存手法を組み合わせた PPC+JOP と比べて，前景推定，姿勢推定，クラス推定の性能を改善できると言える．

5. ま と め

本論文では，教師信号である前景マスクと姿勢パラメータとの付与コストを削減するために，物体クラスラベルと微量の枚数の前景マスクのみとを教師信号として用いて，前景推定，姿勢推定，クラス推定の 3 つのタスクを同時に学習し，高精度に推論できる弱教師あり手法について述べた．既存手法である JOP [7] のネットワークをベースとして，前景推定層と形状損失とを導入することで，物体クラスラベルと微量の枚数の前景マスクのみとを用いて学習可能なネットワークを設計した．実験結果から，教師信号である前景マスクを微量の枚数だけ加えるという前提条件を満たす場合，提案手法は教師信号の付与コストを削減しつつ全てのタスクを同時に学習し，既存の弱教師あり手法よりも各タスクを高精度に推論できることを確認した．

今後の課題として，提案手法の前景推定タスクと姿勢推定タスクとの精度を向上させることで，完全教師あり手法の精度に近づけることを目指す．また，提案手法では直方体や平面といった物体のみを扱うことを想定していたが，今後は曲面を持つ物体も扱える手法を開発していく予定である．

謝辞　本研究の実験及び分析において，白幡十佳知氏には多大なご協力を頂きました．ここに感謝の意を表します．

文　　献

[1] N. Correll, K.E. Bekris, D. Berenson, O. Brock, A. Causo, K. Hauser, K. Okada, A. Rodriguez, J.M. Romano, and P.R. Wurman, “Analysis and observations from the first Amazon Picking Challenge,” IEEE Transactions on Automation Science and Engineering, vol.15, no.1, pp.172–188, 2018.

[2] Z. Zhou, M.R. Siddiquee, N. Tajbakhsh, and J. Liang, “Unet++: A nested u-net architecture for medical image segmentation,” Proceedings of the Deep learning in medical image analysis and multimodal learning for clinical decision support (DLMIA), pp.3–11, 2018.

[3] L. Chen, G.Papandreou, I. Kokkinos, K. Murphy, and A.L. Yuille, “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs,” IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol.40, no.04, pp.834–848, 2018.

[4] D. DeTone, T. Malisiewicz, and A. Rabinovich, “Deep image homography estimation,” arxiv, pp.1–6, 2016.

[5] F. Radenović, G. Tolias, and O. Chum, “Fine-tuning cnn image retrieval with no human annotation,” IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol.41, no.7, pp.1655–1668, 2019.

[6] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp.10012–10022, 2021.

[7] K. Ueno, G. Irie, M. Nishiyama, and Y. Iwai, “Weakly supervised triplet learning of canonical plane transformation for joint object recognition and pose estimation,” Proceedings of the IEEE International Conference on Image Processing (ICIP), pp.2476–2480, 2019.

[8] Y. Wang, J. Zhang, M. Kan, S. Shan, and X. Chen, “Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp.12275–12284, 2020.

[9] Y. Du, Z. Fu, Q. Liu, and Y. Wang, “Weakly supervised semantic segmentation by pixel-to-prototype contrast,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp.4320–4329, 2022.

[10] L. Zhu, Y. Li, J. Fang, Y. Liu, H. Xin, W. Liu, and X. Wang, “WeakTr: Exploring plain vision transformer for weakly-supervised semantic segmentation,” arxiv, pp.1–20, 2023.

[11] J. Wang, Y. Song, T. Leung, C. Rosenberg, Wang, J. Philbin, B. Chen, and Y. Wu, “Learning fine-grained image similarity with deep ranking,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1386–1393, 2014.

[12] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, “Spatial transformer networks,” Proceedings of the Advances in Neural Information Processing Systems (NIPS), pp.2017–2025, 2015.

[13] J. Ahn and S. Kwak, “Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.4981–4990, 2018.

[14] A. Kolesnikov, A. Dosovitskiy, D. Weissenborn, G. Heigold, J. Uszkoreit, L. Beyer, M. Minderer, M. Dehghani, N. Houlsby, S. Gelly, T. Unterthiner, and X. Zhai, “An image is worth 16x16 words: Transformers for image recognition at scale,” Proceedings of the International Conference on Learning Representations (ICLR), pp.1–21, 2021.

[15] B. Calli, A. Singh, A. Walsman, S. Srinivasa, P. Abbeel, and A.M. Dollar, “The YCB object and model set: Towards common benchmarks for manipulation research,” Proceedings of the International Conference on Advanced Robotics (ICAR), pp.510–517, 2015.

[16] C. Rennie, R. Shome, K.E. Bekris, and A.F.D. Souza, “A dataset for improved RGBD-based object detection and pose estimation for warehouse pick-and-place,” IEEE Robotics and Automation

（注2）：https://github.com/usr922/wseg

Letters (RA-L), vol.1, pp.1179–1185, 2016.

[17] R.Araki, T.Yamashita, and H.Fujiyoshi, “ARC2017 RGB-D dataset for object detection and segmentation,” Proceedings of the Late Breaking Results Poster on International Conference on Robotics and Automation (ICRA), pp.1–1, 2018.

[18] S. Tyree, J. Tremblay, T. To, J. Cheng, T. Mosier, J. Smith, and S. Birchfield, “6-dof pose estimation of household objects for robotic manipulation: An accessible dataset and benchmark,” Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp.13081–13088, 2022.

[19] R. Kaskman, S. Zakharov, I. Shugurov, and S. Ilic, “Homebreweddb: Rgb-d dataset for 6d pose estimation of 3d objects,” Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, pp.1–10, 2019.

（xxxx 年 xx 月 xx 日受付）

米田　駿介　（学生員）

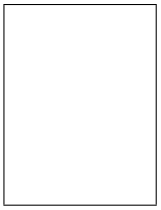
2022 年 鳥取大学大学院持続性社会創生科学研究科博士前期課程修了．現在鳥取大学大学院工学研究科博士後期課程に在籍．弱教師あり学習を用いた物体画像認識技術に関する研究に従事．

入江　　豪　（正員）

2004 年慶應義塾大学理工学部卒業，2006 同大大学院理工学研究科修士課程修了，2011 年東京大学大学院情報理工学系研究科博士課程修了．2006 年～2022 年日本電信電話（株）研究員，2012 年～2013 年米コロンビア大学客員研究員等を経て，現在東京理科大学工学部情報工学科准教授．パターン認識，機械学習，メディア理解の研究に従事．博士（情報理工学）．

西山　正志　（正員：シニア会員）

2000 年 岡山大学工学部情報工学科卒業. 2002 年 同大大学院博士前期課程了. 同年株式会社東芝入社. 同社研究開発センターを経て, 現在鳥取大学大学院工学研究科教授. 2011 年 東京大学大学院学際情報学府にて博士（学際情報学）を取得. カメラを用いた人物認識を始めとするパターン認識およびインタラクションの研究に従事. 山下記念研究賞や画像センシングシンポジウム優秀学術賞など受賞. 電子情報通信学会, 情報処理学会各会員.



岩井　儀雄　（正員）

1992 年 (平成 4 年) 大阪大学基礎工学部情報工学科卒業. 1997 年 (平成 9 年) 大阪大学大学院基礎工学研究科博士課程後期修了. 同年同大学院助手. 2003 年 (平成 15 年) 同大学院助教授.2004 年 (平成 16 年)5 月～2005 年 (平成 17 年)3 月英国ケンブリッジ大学客員研究員. 2007 年 (平成 19 年) 同大学院准教授.2011 年 (平成 23 年）鳥取大学大学院工学研究科教授. コンピュータビジョン, パターン認識の研究に従事. 博士（工学）

Abstract A method that accurately and simultaneously perform segmentation, pose estimation, and object recognition task is necessary to realize an automatic object picking system to reducing labor shortages for warehouse solutions. Recently, deep neural networks have been proposed for segmentation, pose estimation, and object recognition task. To improve the accuracy of each tasks, the network requires large datasets with annotation, but annotation requires expensive manual labor. In particular, the annotation cost of the foreground mask for segmentation task and the pose parameter for the pose estimation task is costly. We propose a novel deep neural network that performs segmentation, pose estimation, and object recognition task using annotation of class labels a micro amount of foreground masks. Experimental results show that our network can perform segmentation, pose estimation, and object recognition tasks with higher accuracy than existing methods, if the assumption that a micro amount of foreground masks are available as annotation.

Key words weakly supervised segmentation, weakly supervised pose estimation, object recognition, class label annotation