Vol.2020-CG-180 No.14
Vol.2020-DCC-26 No.14
Vol.2020-CVIM-223 No.14
2020/11/5

商品画像のための画像クラスを考慮した前景抽出

山口　智史[1,a)]　金森　由博[1,b)]　遠藤　結城[1,c)]　三谷　純[1,d)]

概要：広告制作時の商品画像の切り抜きは，手作業で行われている場合が多い．これを自動化する研究が盛んであり，現在は教師あり学習ベースの手法が主流である．しかし切り抜き用の正解データを大量に集めるのは容易ではない．そこで本研究では，少数の切り抜き用正解データを用いた，半教師あり学習ベースの手法を提案する．提案手法では，既存の画像切り抜き用ニューラルネットワークに，被写体の商品クラスを分類するネットワークを追加し，商品特有の特徴を抽出する．容易に収集可能なクラス分類の正解データを用いることで，切り抜きの正解データ不足を補い，正確な切り抜き処理の実現を目指す．

Class-Aware Image Matting for Product Images

1. はじめに

画像中の物体の切り抜きは，広告制作などで広く利用されている．例えばホームセンターやスーパーマーケットの広告には，様々な種類の商品が掲載されているが，要求品質の高さから未だに主に手作業で切り抜きが行われている．この切り抜き作業を自動化する研究が盛んであり，現在は教師あり学習ベースの手法が主流である．しかしこれらの手法において，商品ごとに大量の正解データを集めることは容易ではなく，商品ごとの色や形状のバリエーションに対応できずに詳細部分を綺麗に切り抜けない恐れがある．

そこで本研究では，商品ごとの色や形状を考慮した半教師あり学習ベースの画像の切り抜き手法を提案する．収集が困難な切り抜き用の正解画像を補うために，収集が容易な，商品ラベルのみが付いた正解画像を追加して学習に用いる．提案するネットワークは，既存の切り抜きネットワークに商品特有の色や形状を考慮するための特徴抽出を行うネットワークを追加したものである．入力画像中の被写体を商品クラスに分類するネットワークを学習し，そのネットワークから得られる特徴マップが商品特有の特徴を保持していると考え，切り抜きに利用する．また本手法の汎化性能を検証するため，インターネット上からダウンロードした画像や，実際にスマートフォンのカメラで撮影した画像に対して本手法を適用した．

[1] 筑波大学
University of Tsukuba, Tennoudai 1-1-1, Tsukuba, Ibaraki, 305–8573, Japan
[a)] yama3104786@gmail.com
[b)] kanamori@cs.tsukuba.ac.jp
[c)] endo@cs.tsukuba.ac.jp
[d)] mitani@cs.tsukuba.ac.jp

2. 関連研究

画像中の物体切り抜きは前景抽出とも呼ばれるタスクであり，入力画像中の前景物を切り出すことを目標としている．各画素 p に対し，入力画像，前景領域，背景領域の画素値をそれぞれ $\mathbf{I}_p$, $\mathbf{F}_p$, $\mathbf{B}_p$ とし，各画素 p での透過度を表すアルファ値を $\alpha_p \in [0,1]$ とすると，これらは Matting 方程式と呼ばれる次式を満たす．

$$\mathbf{I}_p = \alpha_p \mathbf{F}_p + (1-\alpha_p)\mathbf{B}_p \tag{1}$$

ここで p は入力画像の画素である．通常，前景抽出タスクの出力は，各画素にアルファ値を持つグレースケール画像 (アルファマット) である．

こうした画像中の物体の切り抜き手法はすでに多くの手法が提案されている．これらの手法は，非学習ベースの手法 [1], [2], [3] と学習ベースの手法 [4], [5], [6], [7], [8] に分けられる．

2.1 非学習ベースの手法

従来の切り抜き手法として文献 [1], [2], [3] などがある．切り抜きのタスクは Matting 方程式を解くという，制約の数に比べて未知数が多い過小制約問題である．制約の少なさを補うため，これらの手法ではユーザ入力を必要としている．ユーザ入力とは，前景領域と背景領域を表すスクリ

ブルや Trimap である．Trimap とは，入力画像に対して前景領域と未知領域と背景領域がピクセル毎に指定された画像である．ユーザ入力で指定された領域の色情報をもとに，切り抜き処理が行われる．本研究ではユーザ入力なしで RGB 画像 1 枚のみを入力とすることを目標としている．

2.2 学習ベースの手法

切り抜きタスクを扱う手法の中で, 近年主流であるのが学習ベースの手法であり, その中でも畳み込みニューラルネットワーク (CNN) ベースの手法の研究 [4], [5], [6], [7], [8] が盛んに行われている．Xu らは CNN を含むネットワークを構築することで，Matting 方程式を解くことなく画像の前景のアルファマットを直接求めた [4]．Chen らは，Xu らの手法をベースに，Matting 処理にそれまで必要だった Trimap を入力に必要としないネットワークを提案した [5]．さらに，入力と学習どちらにも Trimap を必要としないネットワークが Zhang らによって提案された [8]．商品画像の切り抜きが行われる場面では，入力に Trimap を必要としないネットワークが望ましいが，学習時の Trimap 作成は容易であると考えたため，本研究では Chen らのネットワークをベースとした．

上述のいずれの手法も，被写体の物体クラスごとに共通の特徴は陽に考慮しておらず，本研究のようにそれを陽に考慮することで性能を向上させられる可能性がある．

3. 提案手法

本研究では，商品画像クラスごとに特有な特徴を考慮した切り抜き処理を行うため, 切り抜き処理のネットワークにクラス分類ネットワークを導入し，全体として end-to-end で前景領域を推定する．切り抜きを行う際にクラス分類ネットワークの特徴マップを参照することで，あるクラスに特有な特徴を考慮した切り抜き処理を目指す．

3.1 ネットワーク構造

クラス分類ネットワークには He らのネットワーク (ResNet18) [9] を使用し，切り抜き処理のネットワークは Chen らのネットワーク [5] をベースとした．クラス分類ネットワークの特徴マップは Trimap を推定する T-Net と Matting 処理を行う M-Net の両方に渡される．

図 1 に本研究で使用したネットワークの構成を示す．入力は 3 チャンネルの RGB 画像，出力は 1 チャンネルでグレースケールのアルファマットである．中間出力として 3 チャンネルの Trimap が生成される．

クラス分類ネットワークから T-Net と M-Net に渡される特徴マップは，4 (チャンネル)×512×512 である．この特徴マップは，ResNet18 の入力層から数えて 2, 3, 4, 5 層目から得られる特徴マップそれぞれに対して畳み込みおよびアップサンプリング処理をして，それらをチャンネル方向に連結することで得られる．T-Net の入力チャンネル数は，特徴マップと入力画像をチャンネル方向に連結して 7 チャンネル，M-Net の入力チャンネル数は，特徴マップと入力画像と Trimap をチャンネル方向に連結して 10 チャンネルとなる．

3.2 損失関数

本研究での損失関数は，切り抜き処理に関する損失関数とクラス分類に関する損失関数の 2 種類から構成される．切り抜き処理については，Chen らの損失関数 [5] に加えて，切り抜き処理に有効と考えられる損失関数を 3 つ使用した．クラス分類については，クロスエントロピー誤差関数 $\mathcal{L}_{class}$ を使用した．

切り抜き処理に関する損失関数について述べる．式 (2) に示す Chen らの損失関数は，Trimap についてのロスとアルファロス，合成ロスの 3 つの和で構成される．Trimap についてのロスは，クロスエントロピー誤差関数 $\mathcal{L}_t$ である．アルファロスは推定されたアルファマット $\tilde{\alpha}$ と正解のアルファマット α の L1 ロスである．合成ロスは推定された前景領域と背景画像が合成された画像 $\tilde{c}$ と正解の前景領域と背景画像が合成された画像 c についての L1 ロスである．

$$\mathcal{L}_{chen} = \mathcal{L}_t + 0.5\left\|\tilde{\alpha} - \alpha\right\|_1 + 0.5\left\|\tilde{c} - c\right\|_1 \tag{2}$$

Chen らの損失関数に加えて，式 (3,4) に示す重み付き TV ロス [10]，式 (5) に示す Gradient Loss [6], [7], [11], [12]，式 (6) に示すガボールロス [6] を使用した．

$$\mathcal{L}_{wTV} = \sum_{u,v\in N(p)} w\left(\mathbf{I}(u), \mathbf{I}(v)\right)\left\|\mathbf{I}(u) - \mathbf{I}(v)\right\|_1, \tag{3}$$

$$w\left(\mathbf{x}, \mathbf{y}\right) = \exp\left(-\frac{\left\|\mathbf{x} - \mathbf{y}\right\|_1}{\sigma}\right) \tag{4}$$

ここで $N(p)$ は，画素 p に隣接する画素の集合を表す．

$$\mathcal{L}_{grad} = \left\|\nabla\tilde{\alpha} - \nabla\alpha\right\|_2^2 \tag{5}$$

ここで式 (5) において，∇ は画像の勾配を求めるための演算子である．

$$\mathcal{L}_{gb} = \sum_{\phi\in\Phi}\left\|\phi(\tilde{\alpha}) - \phi(\alpha)\right\|_2^2 \tag{6}$$

ここで式 (6) において $\phi(\cdot)$ はガボールフィルタによる畳み込み処理を表し，Φ はパラメータの異なるガボールフィルタの集合である．パラメータは Li ら [6] と同様のものを使用した．

以上をまとめると，本実験で使用する損失関数 $\mathcal{L}$ は式 (7) となる．

$$\mathcal{L} = w_{chen}\mathcal{L}_{chen} + w_{wTV}\mathcal{L}_{wTV} + w_{grad}\mathcal{L}_{grad} + w_{gb}\mathcal{L}_{gb} + w_{class}\mathcal{L}_{class} \tag{7}$$

ここで w_* はそれぞれ対応するロスの重みである．

Vol.2020-CG-180 No.14
Vol.2020-DCC-26 No.14
Vol.2020-CVIM-223 No.14
2020/11/5

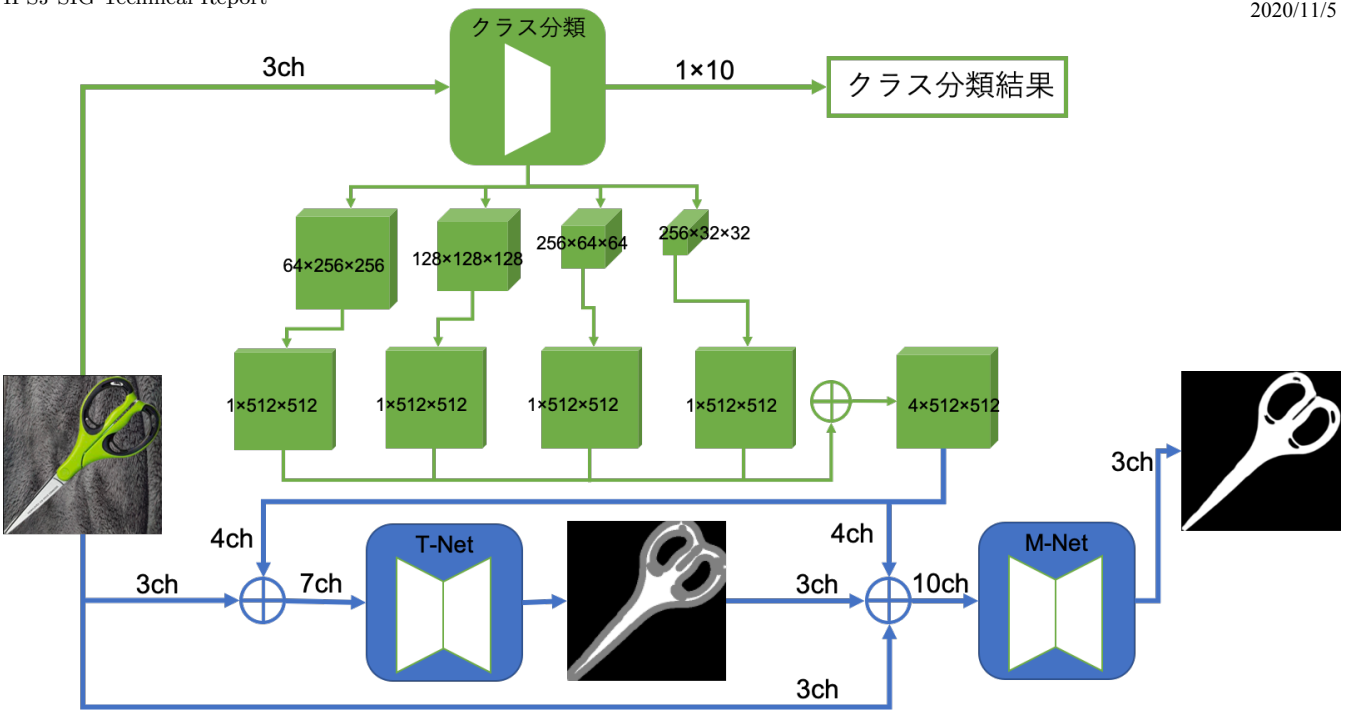

図 1: 我々のネットワークの構成．入力は 3 チャンネルの RGB 画像，出力は 1 チャンネルのアルファマットである．中間出力として 3 チャンネルの Trimap が生成される．クラス分類ネットワークによって入力画像のクラス情報を含んだ特徴マップが抽出され，特徴マップは Trimap とアルファマットを推定する際に用いられる．⊕ はチャンネル方向の連結を表す．

4. データセット作成

この節では，データセットの作成方法と内容について説明する．本研究では 1 節で述べた通り，商品の意味ラベルが付与された切り抜きのためのデータセットを新たに作成した．前景用の商品画像は Amazon からダウンロードした．ダウンロードした商品画像から背景が単純な (ほとんど白に近い) 画像を選別し，remove.bg という商用切り抜きサービスによって前景領域のアルファマットを抽出した．抽出した前景領域と，別に用意した背景用の画像を合成して入力画像とした．背景用の画像は，商品が撮影される環境を想定して室内の床や壁をスマートフォンのカメラで撮影した．商品の種類は全部で 10 種類 (自転車，椅子，やかん，マグカップ，フライパン，ペン，ペンチ，はさみ，トング，傘) とし，ラベルは手動で付与した．背景用に用意した画像は 172 枚である．データセットは，入力画像，正解の Trimap，正解のアルファマット，商品クラスの意味ラベルを 1 組として構成される．学習用に 1,971 組，テスト用に 494 組を用意し，画像の解像度はすべて 512×512 画素とした．

5. 実験

5.1 実験環境

提案手法の実装には Python 言語と PyTorch ライブラリを使用し，NVIDIA GeForce GTX 1080 Ti を用いてモデルを学習した．最適化には Adam を使用し，学習率は 0.0001，β_1 は 0.9，β_2 を 0.999 とした．学習は 800 エポック行なった．いずれの学習も 2 日以内に終了した．また学習時のみ，データ拡張としてクロップ，回転，上下左右反転，背景画像の差し替えをランダムに行なっている．

表 1: 本研究で用いるデータセットの分割パターン．

分割割合	切り抜き用正解データ	クラスラベルのみ正解データ
10:90	197 組	1,774 組
5:95	99 組	1,872 組

推定結果の評価には，絶対差和 (SAD)，平均二乗誤差 (MSE)，Gradient error，Connectivity error の 4 つを用いた．Gradient error，Connectivity error は文献 [13] で提案された切り抜きタスク用の評価指標である．

5.2 データセット分割

本研究では，半教師あり学習の効果を検証するため，少量の切り抜き用正解データと大量のクラスラベルのみの正解データを用いた学習を行う．学習に使用する切り抜き用正解データの数が精度にどう影響するか調査するため，データセットとして作成した 1,971 組の正解データを表 1 に示す割合で分割する．ここで，少量の切り抜き用正解データにはクラスラベルの情報は含まれない．

5.3 実験結果

本研究のモデルによる推定結果とベースライン手法による推定結果を比較した．本研究でのベースライン手法は，

表 2: 本研究で作成した独自データセットにおける各手法の定量的比較．太字はそれぞれの指標で最良の結果を示す．

分割割合	手法	SAD↓	MSE↓	Grad↓	Conn↓	Acc↑
10:90	Ours (no class)	**9.61**	**17.57**	**6.83**	**4.67**	-
	Ours	20.49	55.63	9.67	6.63	0.78
5:95	Ours (no class)	**13.01**	**30.54**	**8.61**	**5.60**	-
	Ours	21.21	59.51	10.64	6.29	0.72

提案手法からクラス分類ネットワークを除いたものとし，“Ours (no class)” と表記する．提案手法とベースライン手法のいずれの手法もデータセットを 10:90 および 5:95 に分割した実験を行う．ベースライン手法では切り抜き用正解データのみを学習に用いる．定量比較結果を表 2 に示す．表の Acc とはクラス分類の正解率であり，Grad と Conn はそれぞれ Gradient error，Connectivity error を表している．学習に使用した切り抜き用正解データが多い分割割合である 10:90 の方が精度が良い．提案手法は，全ての評価指標でベースライン手法を上回れなかった．これは，クラス分類ネットワークから抽出した特徴量が，切り抜きネットワークにとってノイズとして作用してしまったためだと考えられる．次に，定性比較結果を図 2 に示す．ほとんどの場合で，データセットの分割割合にかかわらずクラス情報ありの提案手法の方が悪い結果となった．

5.4 実写画像への適用

提案手法を，インターネットからダウンロード，もしくはスマートフォンで撮影した画像に対して適用した結果を図 3 に示す．およそ 70 枚をテストしたが，1 行目のはさみの例のように比較的成功したと言える結果は数枚程度であった．2〜4 行目の様に，切り抜き用正解データの数が少ない 5:95 の推定結果の方が物体の概形を捉えている結果が全体の 3, 4 割ほど見られた．5 行目の様に 10:90 の方が良い推定結果となったのは数枚程度で，残りの結果は 6 行目の様な悪い結果であった．

6. 今後の課題

本研究では，画像中の商品クラスを考慮した切り抜き手法を提案した．商品クラス情報の考慮のため，クラス分類ネットワークを導入した．残念ながら提案手法は，現時点ではクラス分類を含まないベースライン手法の性能よりも劣る結果となった．今後の課題として，提案手法の精度向上のためにネットワーク構造の改善，学習方法の工夫が必要である．

参考文献

[1] Anat Levin, Dani Lischinski, and Yair Weiss. A closed form solution to natural image matting. In *CVPR 2006*, page 6168, 2006.

[2] Q. Chen, D. Li, and C. Tang. KNN matting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(9):2175–2188, 2013.

[3] K. He, J. Sun, and X. Tang. Fast matting using large kernel matting laplacian matrices. In *CVPR 2010*, pages 2165–2172, 2010.

[4] Ning Xu, Brian L. Price, Scott Cohen, and Thomas S. Huang. Deep image matting. In *CVPR 2017*, pages 311–320, 2017.

[5] Quan Chen, Tiezheng Ge, Yanyu Xu, Zhiqiang Zhang, Xinxin Yang, and Kun Gai. Semantic human matting. In *Proceedings of the 26th ACM International Conference on Multimedia*, pages 618–626, 2018.

[6] Yaoyi Li and Hongtao Lu. Natural image matting via guided contextual attention. *AAAI*, abs/2001.04069, 2020.

[7] Yaoyi Li, Jianfu Zhang, Weijie Zhao, and Hongtao Lu. Inductive guided filter: Real-time deep image matting with weakly annotated masks on mobile devices. *CoRR*, abs/1905.06747, 2019.

[8] Yunke Zhang, Lixue Gong, Lubin Fan, Peiran Ren, Qixing Huang, Hujun Bao, and Weiwei Xu. A late fusion cnn for digital matting. *CVPR 2019*, pages 7461–7470, 2019.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CVPR 2016*, pages 770–778, 2015.

[10] Yuki Endo, Yoshihiro Kanamori, and Shigeru Kuriyama. Animating landscape: Self-supervised learning of decoupled motion and appearance for single-image video synthesis. *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH Asia 2019)*, 38(6):175:1–175:19, 2019.

[11] Alex Levinshtein, Cheng Chang, Edmund Phung, Irina Kezele, Wenzhangzhi Guo, and Parham Aarabi. Real-time deep hair matting on mobile devices. *2018 15th Conference on Computer and Robot Vision (CRV)*, pages 1–7, 2017.

[12] Xiao Dong Chen, Donglian Qi, and Jianxin Shen. Boundary-aware network for fast and high-accuracy portrait segmentation. *ArXiv*, abs/1901.03814, 2019.

[13] Christoph Rhemann, Carsten Rother, Jue Wang, Margrit Gelautz, Pushmeet Kohli, and Pamela Rott. A perceptually motivated online benchmark for image matting. *CVPR 2009*, pages 1826–1833, 2009.

Vol.2020-CG-180 No.14
Vol.2020-DCC-26 No.14
Vol.2020-CVIM-223 No.14
2020/11/5

図 2: クラス情報あり，なしの定性的比較．左から入力画像，10:90 のクラス情報なし，10:90 のクラス情報あり，5:95 のクラス情報なし，5:95 のクラス情報あり，正解画像．

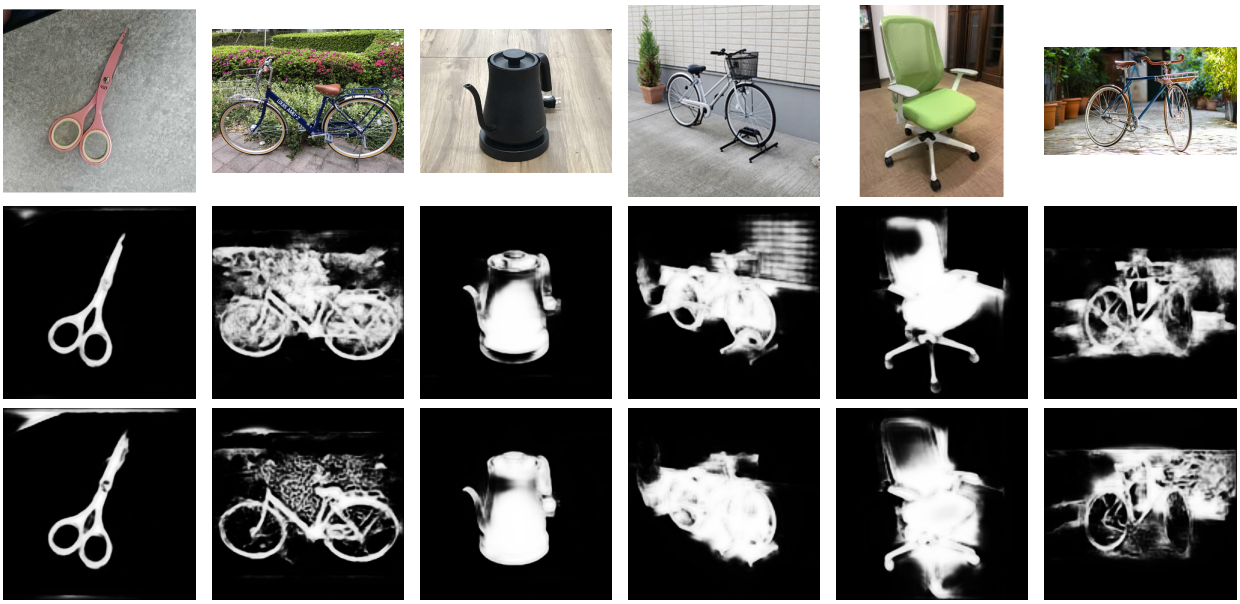

図 3: 実写画像に対する推定結果．列方向に対応する画像が並んでいる．1 行目が入力画像，2 行目が 10:90 の分割割合で学習したネットワークによる推定結果，3 列目は同じく 5:95 のネットワークによる推定結果である．