MaxiMask：教師あり機械学習に基づく天文画像中の偽天体検出法

Maxime Paillassa[1]

訳：西　澤　　淳[2]

〈[1, 2] 名古屋大学素粒子宇宙起源研究所　〒464–8602 愛知県名古屋市千種区不老町〉

〈[2] 岐阜聖徳学園大学 DX 推進センター　〒501–6194 岐阜県岐阜市柳津高桑西 1–1〉

e-mail: [1]maxime.paillassa.c3@f.mail.nagoya-u.ac.jp, [2]atsushi.nisizawa@gifu.shotoku.ac.jp

本稿では，教師あり機械学習を応用した，MaxiMask[1] という畳み込みニューラルネットワーク（Convolutional Neural Network, 以降 CNN と略記）による天文画像の中の偽天体を見つけ出すソフトウェアについて解説する．この MaxiMask の教師データを作成するために，複数の可視・近赤外のカメラによる画像やシミュレーションによる画像を使用した．ひとたび学習すると，MaxiMask は実際の天文画像の中から複数の異なる種類の偽天体を見つけ出すことができる．MaxiMask は GitHub や Python の pip パッケージを通して利用可能である *1.

1. 導　　入

1.1 天文画像における偽検出

可視・近赤外のイメージング観測を用いた天文・天体物理学の研究は，多分にして画像から検出された天体のカタログを使って研究が行われている．したがって，そのカタログの中にある天体ではないもの（偽天体）をなるべくなくすことが重要である．とはいえ，画像の中には多種多様な偽天体が紛れ込んで天文画像を“汚染”しているので，これを実現するのは困難を極める．この偽天体の検出は様々な要因に起因するが，大きく三つのカテゴリに分類できる．

・望遠鏡の特性に由来する光学的要因
　―回折スパイク：副鏡や検出器を支えるアームによって発生する，明るい星のまわりの十字パターン．
　―フリンジパターン：光学系と電子制御系の間の干渉による広がった縞模様
・（CCD）検出器特性に由来する電気的要因
　―損傷（デッド／ホット）ピクセル：異常に低い（または高い）値を持つピクセル．CCD 読み出しは列ごとに行うため，これらは列ごとに並んでいるが，時には小さい塊を形成したり，点状のピクセルに影響する場合もある．
　―飽和（サチュレーション）：ピクセルがその最大値に達する場合に発生する．溢れた電荷は近隣のピクセルに漏れ出して，CCD 列にわたってオーバーフローを起こす．
　―残像効果：前の撮像でサチュレーションを起こした星などを写していた場合，次の撮像にもその信号が残像として残ってしまうことがある．

*1 https://github.com/mpaillassa/MaxiMask

・自然現象による外的要因
—宇宙線：カメラを通過する高エネルギー粒子によって写し出される明るいパターン．カメラへの入射角によって，点像やワームのような形状になる．
—光跡：明るい流れるような軌跡で，露光中に視野を横切る流星，人工衛星，飛行機などによって生じる．
—星団・星雲：これらは実際には天体からの信号であるが，広視野サーベイの場合にはこれらの天体検出と測定は非常に複雑であるため，偽天体として扱うことにする．

図1にこれらの偽天体を例示している．このようにしてみると，多くの偽天体があると思うかもしれないが，これらは単にMaxiMaskで取り扱うことができる偽天体の限定的なリストに過ぎない．他にもCCD間のクロストークや，明るい星によってもたらされるゴーストやハロー，望遠鏡鏡筒内部による反射や屈折などの光も影響する．

1.2 偽天体検出に挑む従来の方法

偽天体検出は時と場合により千差万別様々な形で現れてくるうえに，装置の特性にも強く依存するため，多くの現存手法は以下のような四つの戦略を取らざるを得ない：

1）一種類の偽検出の同定に焦点を絞る．例えば，LACosmic [2]は，ラプラシアンエッジ検出法に基づいて宇宙線を検出する方法である．他の方法では，線や光跡のパターンを同定することに焦点を当てた方法がある[3, 4].
2）HSC [5]やDES [6]など特定の装置の知識がある場合は，特に電子デバイス系の偽天体検出は，異なる露出でも同じピクセルに出やすいことがわかっているので，取り扱いが容易である．
3）CFHTLSサーベイ[7]のように，目視によって明るい星の周りの領域における偽天体を同定することもある．これらは通常非常に複雑で，サチュレーション，回折，スパイク，ハローやゴーストが入り乱れており，それぞれの成分を同定するよりも，目視で除去した方が効率がよいことが多い．
4）同じ領域で何度も露出し，画像同士の引き算を

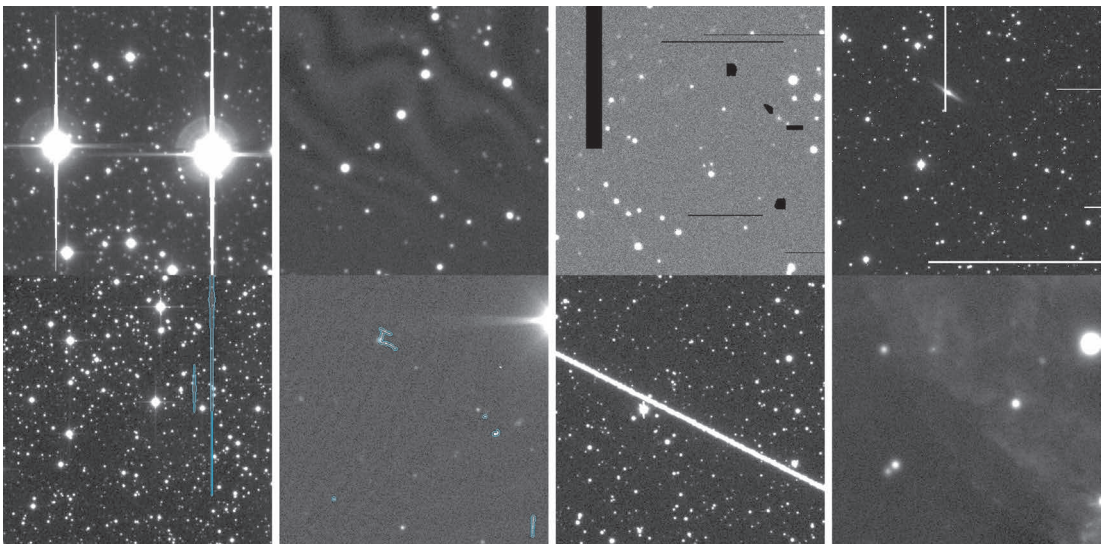

図1　偽天体検出の例．一列目は左から順に明るい星による回折スパイクとサチュレーションのにじみ跡，フリンジパターン，デッドピクセル，ホットピクセル．二列目は左から順に，残像効果，宇宙線，光跡，星団．図は[1]より引用．残光効果と宇宙線は青枠により効果を強調している．

した時に一つの画像にのみ現れたものを偽天体として取り除く試みもある．これは将来計画であるVera C. Rubin天文台のLegacy Survey of Space and Timeサーベイ[8]で策定されている戦略である．

これらの四つの戦略はどれもMaxiMaskのスコープに合致するもので，COSMIC-DANCE[9]やEuclid[10]のようなプロジェクトでも必要となる考え方である．ここで挙げた二つのプロジェクトはその科学目標はかなり違うものの（前者は星の初期質量関数を，後者は宇宙論的大規模構造を取り扱う），両者には様々な観測機器に由来する膨大なデータを取り扱う必要があるという共通点がある．したがってすべての偽天体を一度に見つけ出すような，統一された仕組みが必要になってくる．そのため，上記戦略1はほぼ諦めなければならないだろう．そして我々の手法は不均質なデータに対してもうまく機能する必要があるので，戦略2も捨てなければならない．また，大量のデータは目視ではチェックしきれないため，戦略3も役立たない．最後に，多くのサーベイでは同じフィールドを何度も繰り返し観測するとは限らないため，戦略4も使えるとは限らない．したがって，我々は今後のサーベイ観測において，偽天体検出を取り扱うために何か新しい方策を取らなければならないということは明白である．

1.3 我々の方法

前章で述べた理由もさることながら，コンピュータービジョン分野においては，近年目覚ましく機械学習が進展しており，我々はこのような先端技術の恩恵を受ける形でMaxiMaskの開発に着手した．MaxiMaskは，教師あり機械学習の手法に基づいて様々な種類の偽天体を一度に検出するプログラムであり，より正確には畳み込みニューラルネット（Convolutional Neural Network: CNN）を用いたアルゴリズムである．CNNは画像の中の特徴量がどの場所にあっても効率的にその特徴量を検出してくるので，画像解析の分野の中で特に詳しく調査されているモデルである[11]．CNNがコンピュータビジョン問題で最も優れた成績を収めたのが，画像分類の問題である[12]．画像分類は，その画像全体が何を示すものなのかを定義する一つのラベルを与えるような問題である．後に画像分類CNNは画像ピクセル一つひとつにラベルを与える，画像セグメンテーション[13]へと拡張される．MaxiMaskはまさにこの画像セグメンテーションの技術であり，画像のピクセルレベルで偽天体を同定する．画像セグメンテーションの結果を図2に示している．

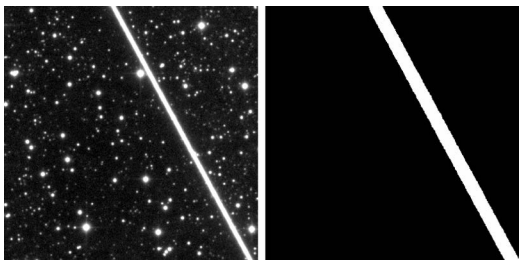

図2　インプット画像（左図）と光跡の二値マスク（右図）．画像分類システムは，その画像全体が光跡かどうかのラベル付けを予言するのに対して，セマンティックセグメンテーションは各ピクセルが光跡かどうかを判定し，二値のマスクとして表現する．

2. MaxiMask

2.1 教師あり機械学習

機械学習による方法は，従来の方法論からの完全なるパラダイムシフトをもたらした．アルゴリズムをデザインするのではなく，入力と出力データセットからどのようにモデルを学習させるかということをデザインする．学習プロセスを図3に示しているが，入力と出力のデータセットの組に対して，何度も最適化過程を繰り返すことで学習が進む．各学習は以下のような段階を経て進んでいく．

- モデルパラメータを使って，入力から出力の予測を立てる．
- 出力と予測の差を予測誤差とし，これを損失関数を通して定量化する．

・最急降下法などのあらかじめ定義された最適化の方法を用いることで，予測誤差に基づいてモデルパラメータを最適化する．他の最適化の手法などに興味のある読者は[15]を参照されたい．

このような教師あり学習の枠組みの中では，入力／出力のデータセットの組を用意する必要がある．実際には，そのような画像は偽天体を含む入力画像と，その偽天体の種類が何であるかをラベル付けした答えとしての出力の組が必要である（後者は例えば図2で示した光跡の二値マスク画像などである）．

2.2 MaxiMask教師データ

MaxiMaskのような教師あり学習のモデルを学習させるときに最も気にしなければならないのは，学習に用いていない新しいデータにも正しく適用できるかどうかという観点である．これを汎化性能という．MaxiMaskの汎化性能を最大限に担保するために，教師データを作成する際にはなるべく多くの実データを使うことに注意を払った．ここで最も難しい問題は，入力画像に写っているものがどの種類の偽天体であるのかという正解を知る必要があるということである．この問題の解決策として，偽天体が写っていない実データに，偽天体の画像を入れ込むということを行う．人工的にどこに何を入れたのかはわかっているため，正解のマスクマップは容易に作成することができる．

したがって次にやるべきことは，実データ（ここではCOSMIC-DANCEサーベイ[9]）の中から最もクリーンな（偽天体の少ない）画像を同定することである．我々は最終的に三つの異なる観測装置，DECam [16]，MegaCam [17]，HSC [18]で得られた画像を用いて，偽天体のない純粋な画像のベースを作ることにした．各観測機器の画像を用意する際には，その機器に特化して用意されているソフトウェアパイプラインを使うことで，可能な限りクリーンな画像を得ることができる．

偽天体のない画像を取得したら，それらに偽天体を追加すると同時に，対応する正解のマスクを作成する．ただし，注意しなければならないのは，偽天体が含まれていないと考えている“実際の”画像には，実際には偽天体が潜んでいるということである．例えば，サチュレーションや回折スパイクなどである．したがって偽天体を画像に足す前に，そのように画像にもともと含まれているような偽天体をすべて同定し，それらを正解マスクとして定義しておく必要がある．サチュレーションピクセルは検出器の特性として事前に知ることができるし，回折スパイクはここでは詳細には触れないが複雑で経験的な方法で同定することが可能である（興味のある読者はMaxiMaskの文献[2]を参照されたい）．ここで，天体の中でも最も明るいものは個別のクラスとして同定しておく．これによって経験的に分類がよりうまくいくことがわかっているからだ．これらの準備が終わると，人工的な偽天体を画像に入れていく．ここでも同じく汎化性能を担保するために，偽天体の画像もなるべく本物の画像を使うようにする．挿入した偽天体は以下のとおりである．

・宇宙線:「ダーク画像」という較正用のデータを用いる．ダーク画像には電子回路起因の信号と宇宙線のみが信号として捉えられる．
・デッド・ホットピクセル：シミュレーションを用いる
・残像効果：ハッブル宇宙望遠鏡（HST）の広視野カメラWFC3 [19]用に開発されたモデルと，残像効果を引き起こす偽の明るい星を作り出すSkyMaker [20]を用いたシミュレーションを用いる．
・フリンジ：実際のフリンジマップを使う
・星雲：ハーシェル望遠鏡[21]のSPIRE検出器[22]から得た実データを使う
・光跡：SkyMaker [20]を用いて様々な星のシミュレーションを行う
・オーバースキャン：シミュレーションを用いる．オーバースキャンとはCCDの端にある空（から）のピクセルである．実際のイメージ

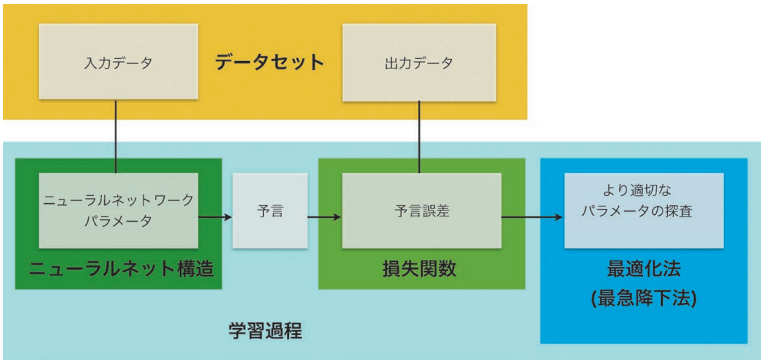

図3　教師あり学習の概略図．図は[14]から翻訳．

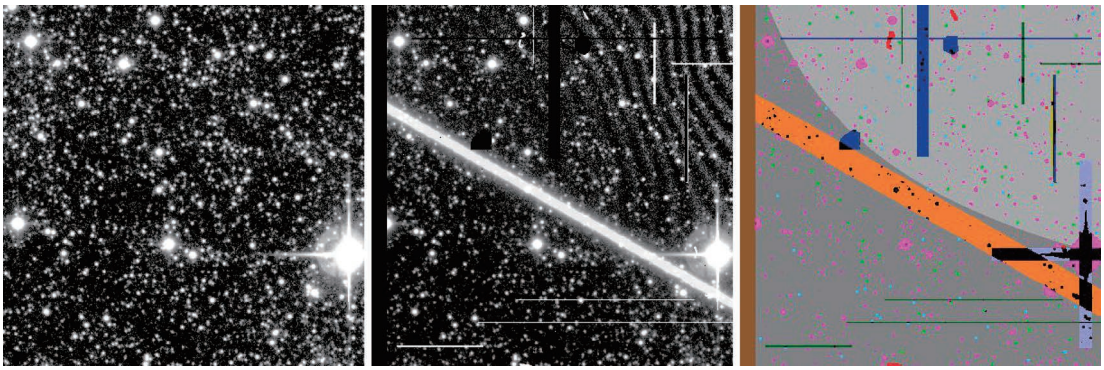

図4　MegaCamの偽天体を含まない画像（左），偽天体を人為的に加えた画像（中），正解のマスク画像（右）．マスク画像は実際には各偽天体は正解の二値マスクを持つ．図のように色付きの正解マスクマップを作るには，各偽天体の種類ごとに色を割り当てる．赤：宇宙線，深緑：列状ホットピクセル，群青：列状デッドピクセル，緑：点状のホットピクセル，シアン：点状のデッドピクセル，黄色：残光効果，オレンジ：軌跡，灰色：フリンジ，紫：サチュレーション，薄紫：回折スパイク，茶：オーバースキャン，マジェンタ：明るい光源，暗灰：背景．黒いピクセルは複数の偽天体がオーバーラップしているピクセルを示している．可視化のために点状のデッドピクセルなどは1ピクセルではなく3×3の大きさを持った四角で表現してある．

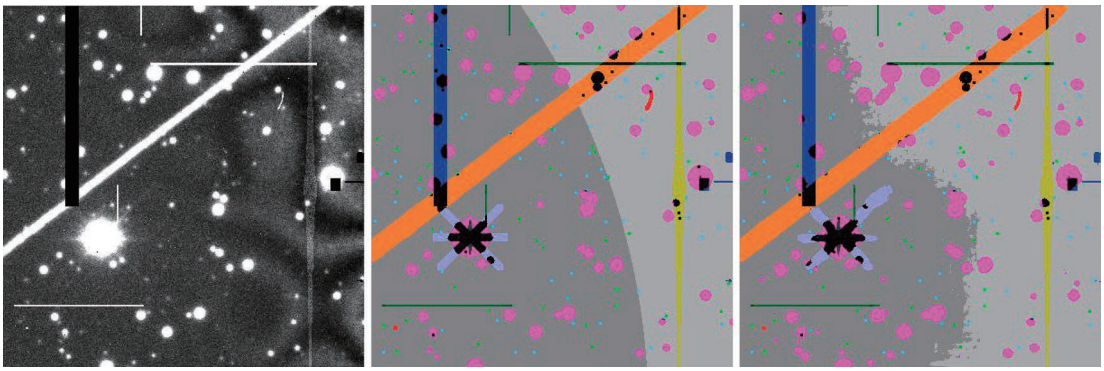

図5　入力テスト画像（左），正解のマスク（中），MaxiMaskが予測したマスク画像（右）．

データを取り扱うために，オーバースキャン領域でのモデル予測も考慮する．

図4に偽天体のないクリーンな画像と，学習に使われる偽天体を人工的に埋め込んだ画像を例示する．

2.3 学習とテスト

MaxiMaskを学習させるのに，図4で示したような画像を50,000サンプル用意した．MaxiMaskの学習は30エポック行う．すなわち，50,000枚の画像を30回機械に与えて学習をさせる．図3で説明されている通り，各学習の段階でMaxiMaskは予測を行うが，その予測というのは各ピクセルの偽天体種別ごとの[0, 1]の範囲の連続値である確率で，アウトプット全体としてはそのような確率のマップになる．ここでは，アウトプットの確率マップと正解マップが等しい時に最小化するような損失関数を導入する．

学習が終わると，学習には使っていないテストデータセットを用いてMaxiMaskの予測が正しいかどうかの検証を行う．このようなテストデータセットを用いた検証は一般的に行われていることで，モデルが学習データにオーバーフィット（過学習）していないかを確かめる重要な指標となる．図5にテスト画像に対するモデル予測の例を示す．

ここで，我々は各テストデータに対して正解の偽天体マスクを知っているので，定量的に予測性能を評価することが可能である．より正確に言えば，MaxiMaskによって予測された確率の閾値を[0, 1]の範囲で徐々に変化させた時に各偽天体について以下の基準を満たすものの計数を調べる．

- 真陽性（True Positive, TP）：偽天体であると予言され，実際にも偽天体であるピクセル．
- 偽陽性（False Positive, FP）：偽天体であると予言されたが，実際には偽天体ではなかったピクセル．
- 真陰性（True Negative, TN）：偽天体ではないと予言された，実際にも偽天体でなかったピクセル．

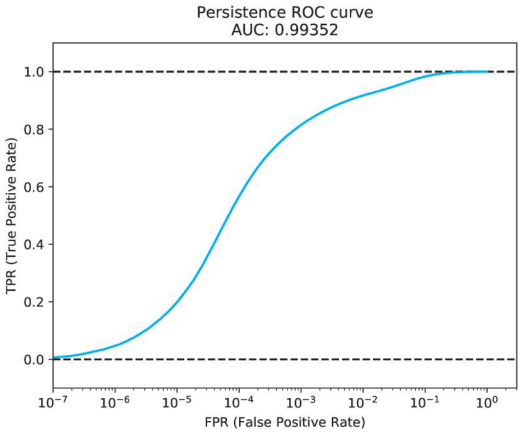

図6　テストデータを使って計算された残光効果に対するROC曲線を示す．各点は閾値を[0, 1]の範囲で連続的に変えた時の（FPR, TPR）の組みを表す．閾値を小さく取るとTPRは増加する一方で，FPRもまた増加してしまう．逆に閾値を大きく取ると，FPRを小さく抑えることができるが，TPRも小さくなってしまう．理想的な分類がなされた時は，左上（TPR＝1かつFPR＝0）に点が打たれる．図の出典[1]．

- 偽陰性（False Negative, FN）：偽天体ではないと予言されたが，実際には偽天体であったピクセル．

ここで性能評価としてよく使われるのがROC（受信者動作特性）曲線であるが，これはある閾値を定めた時の真陽性率（TPR）と偽陽性率（FPR）を連続的に比較したものである．

- TPR＝TP/（TP＋FN）と定義されるが，これは偽天体であるピクセルのうち，正しく偽天体であると分類されたものの割合を示す．
- FPR＝FP/（FP＋TN）と定義されるが，これは偽天体ではないピクセルのうち，偽天体であると間違って認識されてしまったものの割合を示す．

完全に正しい分類が行われたとすると，TPR＝1かつFPR＝0となる．残光効果に対するROC曲線の例を図6に示す．

2.4 実データへの応用

最後にMaxiMaskの実際の画像データへの応用について紹介する．図7はDECam, 図8はHST

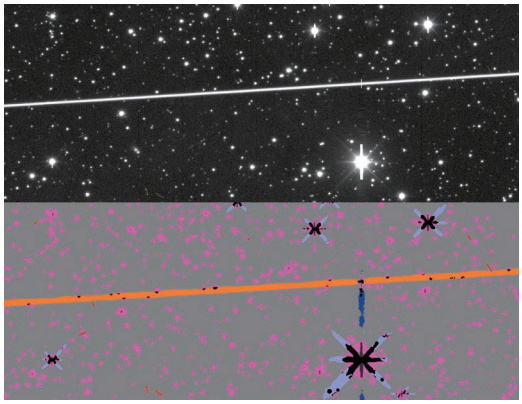

図7　MaxiMaskをDECamの実画像データに適用した例．実際に光跡，宇宙線，星の回折スパイクが検出されている．

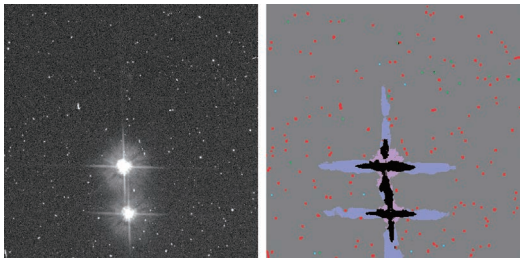

図8　MaxiMaskをHST搭載のACSカメラによる実画像データに応用した事例．このデータは宇宙望遠鏡のデータであるが，一方で学習はすべて地上望遠鏡の画像を用いて行われている．入力画像（左），MaxiMaskの予測マップ（右）．

のACSカメラの画像に対する応用を示している．ACSの画像データはMaxiMaskの学習には一切使われていなかったことに注意していただきたい．この結果から，MaxiMaskは学習に使わなかったデータに対してもうまく機能しており，高い凡化性能を示していることがわかる．

3. 終わりに

天文画像の中から偽天体を同定するのに，教師あり機械学習の方法がうまく使えることを見てきた．可能な限り実際の天文画像を用いた教師データを工夫して作成することで，畳み込みニューラルネットワークであるMaxiMaskはテストデータでも実際の画像でもうまく偽天体を検出できる．MaxiMaskは，GitHubやPythonのpipパッケージから利用可能である．

今後の展望としては，1.1節で述べたようないまだに同定されていない偽天体の検出をできるようにしたり，MaxiMaskを拡張して偽天体を含む画像を補正する機能を実装することである．少なくとも今回紹介した図8のハッブル宇宙望遠鏡の画像データに対してはよい汎化性能を示していたが，学習用データとして，より多様な画像（たとえば，系外銀河の画像や，宇宙望遠鏡による画像など）を用いることで，より高い汎化性能を追求することができるだろう．

参考文献

[1] Paillassa, M., et al., 2020, A&A, 634, A48
[2] van Dokkum, et al., 2012, L. A. Cosmic: Laplacian Cosmic Ray Identification, ascl:1207.005
[3] Bektešević, D., & Vinković, D., 2017, MNRAS, 471, 2626
[4] Nir, G., et al., 2018, AJ, 156, 229
[5] Bosch, J., et al., 2018, PASJ, 70, S5
[6] Morganson, E., et al., 2018, PASP, 130, 074501
[7] Heymans, C., et al., 2012, MNRAS, 427, 146
[8] Bosch, J., et al., 2019, ASP Conf. Ser., 523, 521
[9] Bouy, H., et al., 2013, A&A, 554, A101
[10] Racca, G. D., et al., 2016, SPIE Conf. Ser., 9904, 99040O
[11] LeCun, Y., et al., 1995, The handbook of brain theory and neural networks, 3361
[12] Krizhevsky, A., et al., 2012, Advances in neural information processing systems, 25
[13] Badrinarayanan, V., et al., 2017, IEEE Trans. Pattern Anal. Mach. Intell., 39, 2481
[14] Paillassa, M., 2020, Ph.D. Dissertation, Universite de Bordeaux, Talence, France
[15] Ruder, S., 2016, arXiv preprint arXiv:1609.04747
[16] Flaugher, B. L., et al., 2010, SPIE Conf. Ser., 7735, 77350D
[17] Boulade, O., et al., 2003, SPIE Conf. Ser., 4841, 72
[18] Miyazaki, S., et al., 2018, PASJ, 70, S1
[19] Long, K. S., et al., 2015, Persistence in the WFC3 IR Detector: an Improved Model Incorporating the Effects of Exposure Time, Space Telescope WFC Instrument Science Report
[20] Bertin, E., 2009, Mem. Soc. Astron. Italiana, 80, 422
[21] Pilbratt, G. L., et al., 2010, A&A, 518, L1
[22] Griffin, M. J., et al., 2010, A&A, 518, L3

MaxiMask: A Supervised Machine Learning Based Method to Identify Contaminants in Astronoimcal Images

Maxime Paillassa, Atsushi J. Nishizawa (translation)

Kobayashi Maskawa Institute, Nagoya University, Furocho Chikusa-ku, Nagoya, Aichi 464–8602, Japan

Abstract: Leveraging supervised machine learning techniques, we design MaxiMask [1], a Convolutional Neural Network (CNN) that can identify contaminants in astronomical images. We use imaging data from several optical and near-infrared cameras and simulations to build training data samples for MaxiMask. Once trained, MaxiMask is able to identify various contaminants in real images. MaxiMask is available for use on GitHub and as a Python pip package.