Source detection through semantic segmentation with convolutional neural networks

Debating the potential of machine learning in astronomical surveys

Maxime Paillassa[1,3], Emmanuel Bertin[2,4] and Hervé Bouy[1]

[1]Laboratoire d'astrophysique de Bordeaux, Univ. Bordeaux, CNRS, B18N, allée Geoffrey Saint-Hilaire, 33615 Pessac, France
[2]Sorbonne Université, CNRS, UMR 7095, Institut d'Astrophysique de Paris, 98 bis bd Arago, 75014 Paris, France
[3]Division of Physics and Astrophysical Science, Graduate School of Science, Nagoya University, Furo-cho, Chikusa, Nagoya 464-8602, Japan
[4]Canada-France-Hawaii-Telescope, 65-1238 Mamalahoa Hwy Kamuela, Hawaii 96743, USA

Source detection

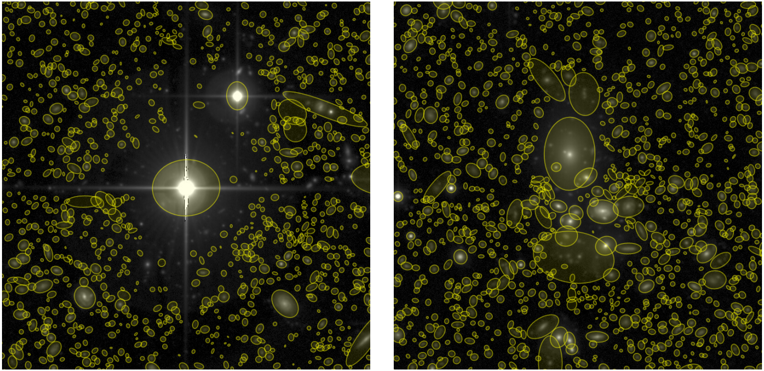


Figure 1: SExtractor (Bertin and Arnouts, 1996) detections in CFHTLS images (Cuillandre and Bertin, 2006).

- Source catalogs are at the basis of many Astrophysical studies.
- Large amounts of data require automatic source detection.
- Current automatic source detection techniques are limited.

Current source detection techniques

- In practice, source detection pipelines proceed in several steps:
  - Sky background subtraction.
  - Matched filter.
  - Peak search or thresholding.
  - Deblending procedures.

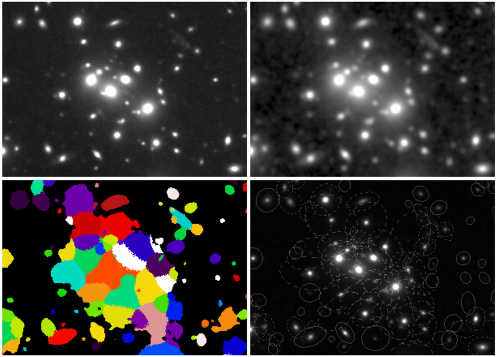


Figure 2: Example of SExtractor processing (taken from documentation).

Current algorithm limitations (1/2)

- Sources come in various scales and shapes.
- Sources can overlap, a phenomenon known as blending.

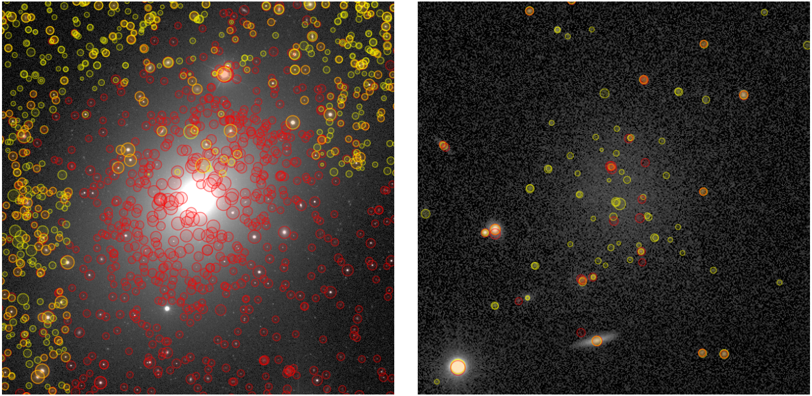


Figure 3: SDSS (yellow) and Pan-STARRS (red) catalogs.

Current algorithm limitations (2/2)

- Images can be contaminated by defects triggering false detections.
- Major source of noise in catalogs.

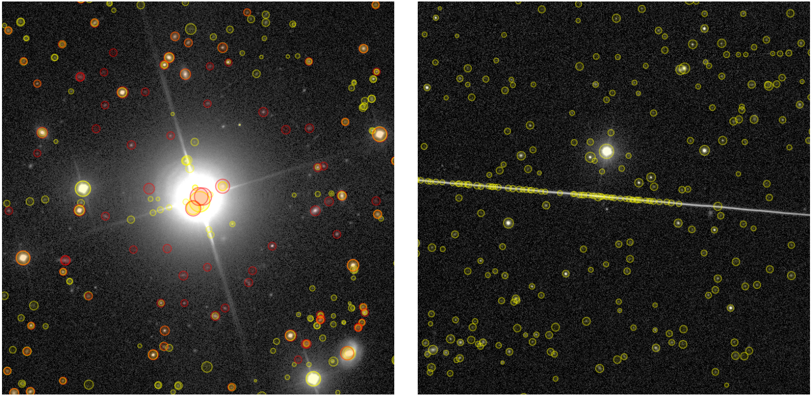


Figure 4: SDSS (yellow) and Pan-STARRS (red) detections.

Toward machine learning

- Extend current methods with supervised machine learning to:
    - Perform adaptative filtering and segmentation.
    - Train robust and versatile models.
    - Learn directly from pixels (with convolutional neural networks).

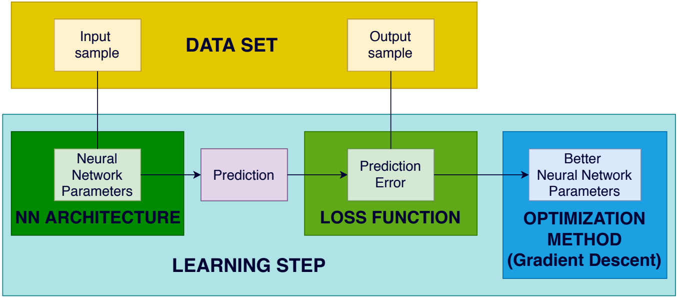


Figure 5: Schematic view of our supervised learning framework.

- Huge amounts of data available in astronomy.

Our approach: deep coloring

- Source detection needs to be instance-aware.
  $\rightarrow$ Makes difficult to solve detection and deblending simultaneously.

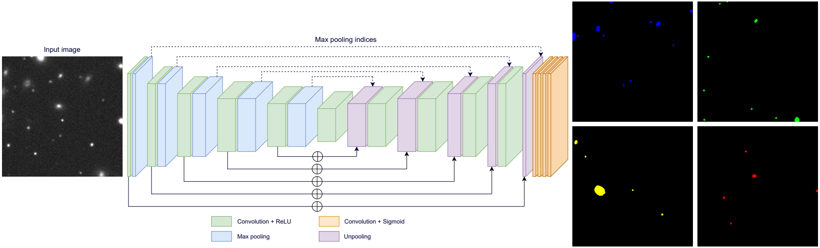


Figure 6: Deep coloring approach illustration, based on Kulikov et al., 2018.

- Rely on a semantic segmentation CNN, i.e. pixel labeling.
- The CNN can freely identify each source in output/color maps.
- Constraint so that close objects are identified in different colors.
- Need to know in which color is detected each object to compute loss!

The deep coloring approach

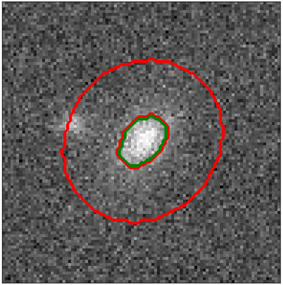


Figure 7: Footprint (green) and halo (red) of a source.

- Each source k has a footprint $M^{(k)}$ and a halo $M_{halo}^{(k)}$.
- Each source is dynamically affected a color at each training step:

$$c_k = \arg\max_{c \in C} \left(\frac{1}{|M^{(k)}|} \sum_{p \in M^{(k)}} \log(\hat{y}(c,p)) + \mu \frac{1}{|M_{halo}^{(k)}|} \sum_{p \in M_{halo}^{(k)}} \log(1 - \hat{y}(c,p)) \right)$$

Training data: stars (1/3)

- Rely on noise-free images of isolated sources.
  $\rightarrow$ Any ground truth information can be computed for each source.
  $\rightarrow$ Whole images can be built from scratch.

- SkyMaker (Bertin, 2006) for stars.

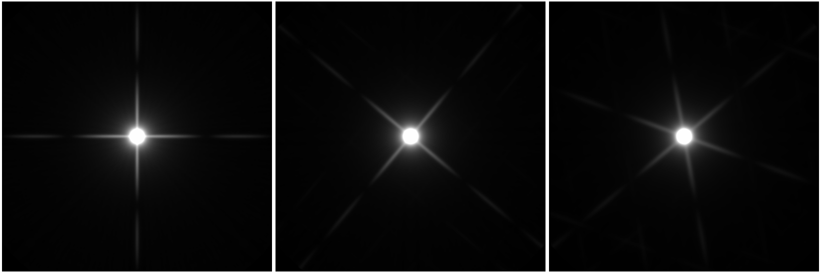


Figure 8: Examples of star profiles including different diffraction spike configurations.

Training data: galaxies (2/3)

- Rely on noise-free images of isolated sources.
  $\rightarrow$ Any ground truth information can be computed for each source.
  $\rightarrow$ Whole images can be built from scratch.

- Cosmological simulations for galaxies.

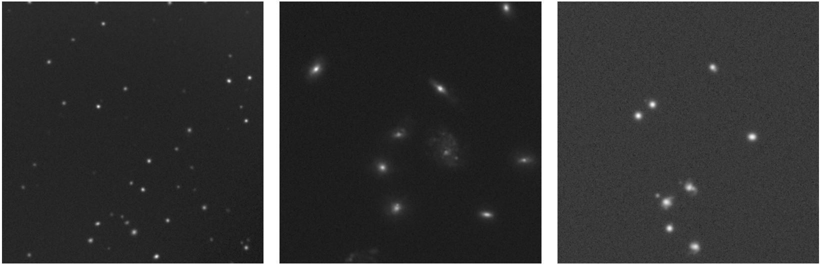


Figure 9: Example of galaxy images. From left to right: galaxies from Horizon-AGN (Dubois et al. 2014, rendered by C.Laigle, private communication), IllustrisTNG (Nelson et al. 2019), Vela (Simons et al. 2019).

Training data: images (3/3)

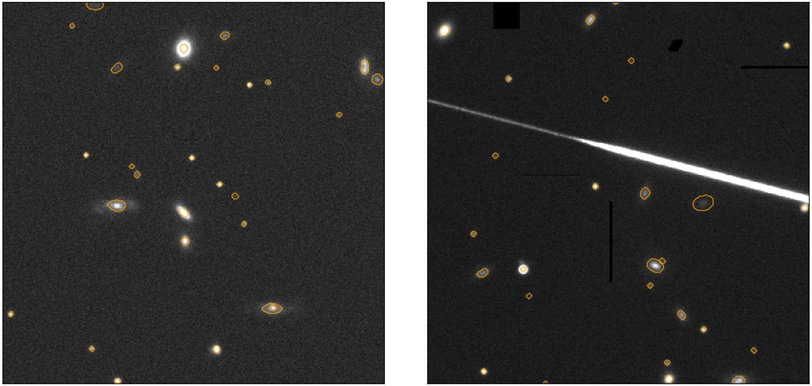


Figure 10: Examples of training images with ground truth footprint overlays.

- Contaminants are added in images such as cosmic rays, bad pixels, persistence effects, fringes, nebulosities, trails, saturation.
  $\rightarrow$ Rely on MaxiMask training data (Paillassa et al. 2020).

Qualitative comparison with SExtractor

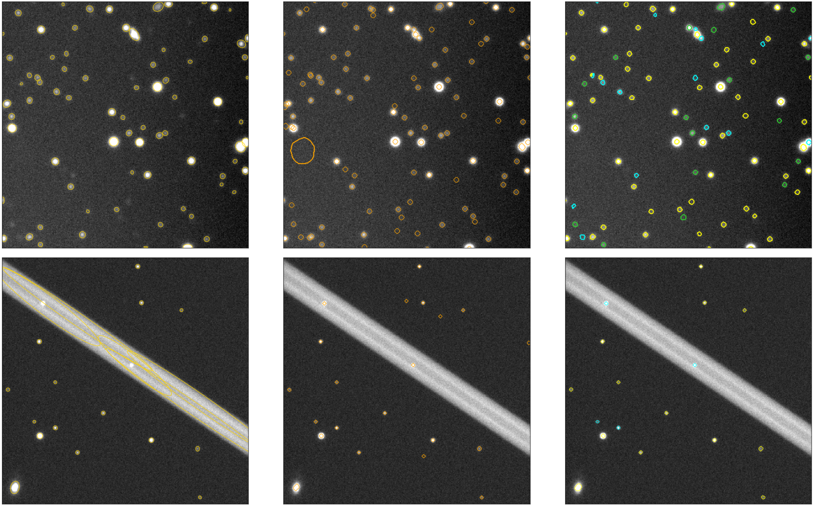


Figure 11: *Left:* SExtractor detections. *Middle:* input image. *Right:* CNN prediction.

Quantitative comparison with SExtractor

- Completeness and contamination at various detection thresholds:
  - CNN thresholds: every 0.02 probability in $[0, 1]$.
  - SEXTRACTOR thresholds: every 0.25 sky σ in $[0.25, 10]$.
  - Completeness: $\frac{\mathrm{TP}}{\mathrm{TP+FN}}$.
  - Contamination: $\frac{\mathrm{FP}}{\mathrm{TP+FP}}$.

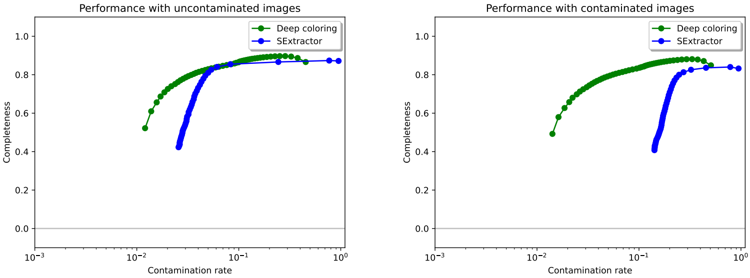


Figure 12: *Left:* performance in an uncontaminated regime. *Right:* performance in a contaminated regime.

Qualitative results on real data (1/2)

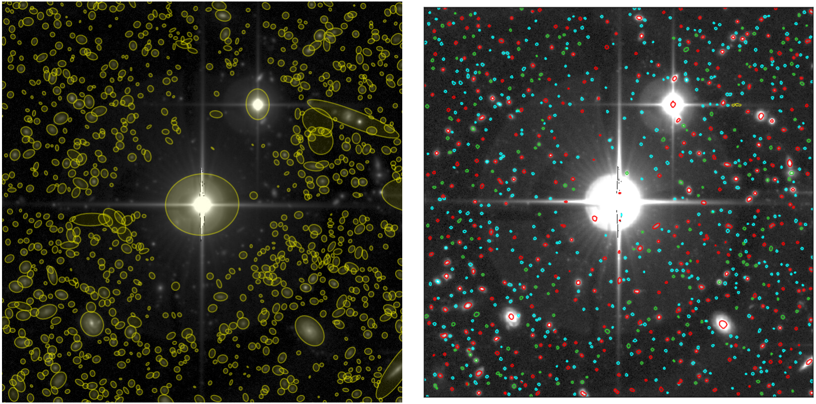
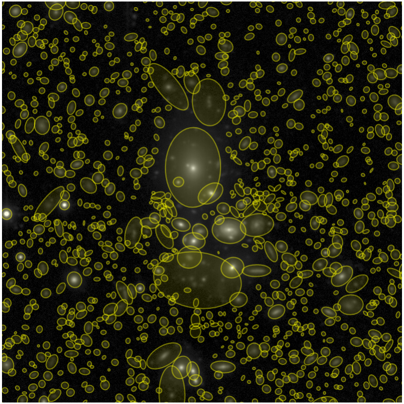


Figure 13: *Left:* SExtractor CFHTLS detections. *Right:* Deep coloring CNN.

Qualitative results on real data (2/2)



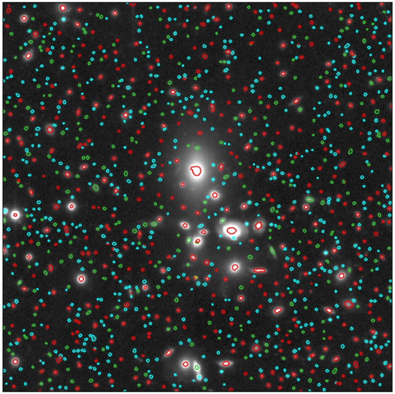


Figure 14: *Left:* SExtractor CFHTLS detections. *Right:* Deep coloring CNN.

The detector in practice

- Python with Nvidia Titan X and without multithreading:
  $\rightarrow$ Image pre-processing alone: $\approx$2.3 MPix/s.
  $\rightarrow$ CNN segmentation alone: $\approx$5 MPix/s.
  $\rightarrow$ Overall: $\approx$1.5 MPix/s: $\approx$11-12s for a 4k2x4k2 CCD.

- Integration in SourceXtractor++ (Bertin et al. 2020).
  $\rightarrow$ Already done thanks to SX++ modularity and ONNX (Open Neural Network Exchange).
  $\rightarrow$ Segmentation maps are naturally handled in SX++.
  $\rightarrow$ Possibility to optimize for various hardwares.
  $\rightarrow$ Facilitate use, benchmarks and comparisons.

Conclusions

- We have a generic and efficient source detection method with CNNs:
  - Deep coloring approach.
    $\rightarrow$ Enables to separate detections in different output maps.
  - Comprehensive and diverse data.
    $\rightarrow$ Able to detect various source morphologies.
    $\rightarrow$ Robust to the presence of contaminants.

- Internal testing is ongoing.
- Will be available soon !

Thank you for your attention.