Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **527,** 1534–1562 (2024) https://doi.org/10.1093/mnras/stad3143
Advance Access publication 2023 October 13

deep PACO: combining statistical models with deep learning for exoplanet detection and characterization in direct imaging at high contrast

Olivier Flasseur,[1,2,3★†] Théo Bodrito,[2] Julien Mairal,[4] Jean Ponce,[2,5] Maud Langlois[3] and Anne-Marie Lagrange[1,6]

[1]*Laboratoire d'Études Spatiales et d'Instrumentation en Astrophysique, Observatoire de Paris, Université PSL, Sorbonne Université, Université Paris Diderot, France*
[2]*Département d'Informatique de l'École Normale Supérieure (ENS-PSL, CNRS, Inria), France*
[3]*Centre de Recherche Astrophysique de Lyon, CNRS, Univ. de Lyon, Université Claude Bernard Lyon 1, ENS de Lyon, France*
[4]*Grenoble INP, Inria, CNRS, Université Grenoble Alpes, LJK, F-38000 Grenoble, France*
[5]*Courant Institute of Mathematical Sciences, Center for Data Science, New York University, USA*
[6]*Institut de Planétologie et d'Astrophysique de Grenoble, Université Grenoble Alpes, France*

Accepted 2023 October 11. Received 2023 October 11; in original form 2023 March 7

ABSTRACT

Direct imaging is an active research topic in astronomy for the detection and the characterization of young substellar objects. The very high contrast between the host star and its companions makes the observations particularly challenging. In this context, post-processing methods combining several images recorded with the pupil tracking mode of telescope are needed. In previous works, we have presented a data-driven algorithm, PACO, capturing locally the spatial correlations of the data with a multivariate Gaussian model. PACO delivers better detection sensitivity and confidence than the standard post-processing methods of the field. However, there is room for improvement due to the approximate fidelity of the PACO statistical model to the time evolving observations. In this paper, we propose to combine the statistical model of PACO with supervised deep learning. The data are first pre-processed with the PACO framework to improve the stationarity and the contrast. A convolutional neural network (CNN) is then trained in a supervised fashion to detect the residual signature of synthetic sources. Finally, the trained network delivers a detection map. The photometry of detected sources is estimated by a second CNN. We apply the proposed approach to several data sets from the VLT/SPHERE instrument. Our results show that its detection stage performs significantly better than baseline methods (cADI and PCA), and leads to a contrast improvement up to half a magnitude compared to PACO. The characterization stage of the proposed method performs on average on par with or better than the comparative algorithms (PCA and PACO) for angular separation above 0.5 arcsec.

Key words: methods: data analysis – methods: numerical – methods: statistical – techniques: high angular resolution – techniques: image processing.

1 INTRODUCTION

High-contrast imaging is an observational method used to study the close environment of stars (Traub & Oppenheimer 2010; Bowler 2016; Pueyo 2018). It is particularly adapted to detect young, massive, and hot exoplanets (see e.g. Chauvin et al. 2004, 2005; Schneider et al. 2011; Nielsen et al. 2019), thus complementing well indirect exoplanet detection methods, such as transit photometry or Doppler spectroscopy (Santos 2008). Direct imaging offers other appealing characteristics like the detection of candidate companions from a few hours of observations and the ability to characterize them in terms of age, surface gravity, effective temperature, and composition (Allard et al. 2003, 2007), or to predict their evolution (Burrows et al. 1997; Chabrier et al. 2000). Despite these promises, only a few dozens exoplanets have been unveiled and characterized since the emergence of direct imaging in the early 2000s (Marois et al. 2008; Lagrange et al. 2009; Nielsen et al. 2012; Macintosh et al. 2015; Chauvin et al. 2017; Keppler et al. 2018). This is mainly due to (i) the relatively low occurrence of giant exoplanets, (ii) the very high contrast between the host star and the exoplanets (typically, higher than 10^5 in the infrared), and (iii) the required high angular resolution (typically, better than a tenth of arcseconds).

In this context, cutting-edge ground-based facilities like VLT/SPHERE (Beuzit et al. 2019), Gemini/GPI (Macintosh et al. 2014), Keck/NIRC2 (Castellá et al. 2016), Magellan/MagAO (Morzinski et al. 2014) or SUBARU/SCExAO (Jovanovic et al. 2015) are equipped with an (extreme) adaptive optics system and a coronagraph to attenuate as much as possible the glare of the star. Currently, the non-blocked residual starlight contamination and its temporal fluctuations remain the main limitation. It takes the form of spatially correlated *speckles* that resemble the expected signature of a point-like source (e.g. an exoplanet, a brown dwarf, and a background star). The observations are also impacted by additional

★ E-mail: olivier.flasseur@univ-lyon1.fr
† O. Flasseur is currently with CRAL (5), and was previously with LESIA (1) and Inria (2) where most of this work was performed.

sources of noise (i.e. thermal background flux, detector readout, and photon noise). Together, speckles and noise form a spatially and non-stationary *nuisance component* that corrupts the signals of the sought objects. Off-axis objects can either take the form of point-like sources or that of spatially extended features like circumstellar discs. In this paper, we focus on the detection of point-like sources, leaving the problem of disc reconstruction for future work.

In order to unmix the sought objects from the nuisance component, high-contrast observations are performed with dedicated strategies like angular differential imaging (ADI; Marois et al. (2006)), that we consider in this paper. ADI consists in tracking the observed target over time, with the telescope derotator tuned to keep the telescope pupil stable while the field of view rotates around the host star. Consequently, in the resulting 3D data sets (2D + time), the objects of interest follow an apparent motion along a deterministic circular trajectory centred on the star while the telescope pupil remains static. With ADI, speckles due to residual starlight aberrations are *quasi-static*, that is, they are strongly correlated across exposures. ADI also allows to extract the astrometry and the photometry of detected sources. These estimates can be used to characterize the physical properties of the detected sources by fitting orbital and atmospheric models (Vigan et al. 2010; Cheetham et al. 2019; Mesa et al. 2019). In this paper, we address two tasks: (i) the detection of point-like sources, and (ii) the estimation of their photometry.

The last cornerstone of high-contrast imaging is data *reduction*, that is, the combination of the recorded images by dedicated post-processing algorithms. This critical step brings the additional gain in contrast (between one and three orders of magnitudes for existing methods) needed to detect the faint signals coming from thermal self-emission of giant exoplanets. The classical principle is to estimate a reference image (so-called *on-axis PSF*, point spread function) of the nuisance component, that can be subtracted from the data in order to unveil the sought objects. A simple practical implementation of this general strategy consists in subtracting the temporal mean or median of the data set from each frame of the ADI stack. The residual images are then co-aligned to the true-North so that the signals of the sought objects are superimposed and can be combined by temporal stacking. This is the principle of the cADI method designed to process the first direct imaging observations (Marois et al. 2006; Lagrange et al. 2009). In the last decade, several more advanced methods have been developed, see, for example, Pueyo (2018) for a review. In particular, TLOCI (Marois et al. 2013, 2014) or its variants such as LOCI (Lafrenière et al. 2007), ALOCI (Currie et al. 2012a, b), MLOCI (Wahhaj et al. 2015)), and KLIP/PCA (Amara & Quanz 2012; Soummer, Pueyo & Larkin 2012) are currently implemented in most reduction pipelines (Amara & Quanz 2012; Gonzalez et al. 2017; Galicher et al. 2018). LOCI-based and PCA-based algorithms are considered as standards to process high-contrast observations. With the (A, M, T)-LOCI algorithm, the on-axis PSF is estimated by combining images selected in a library. The combined images are selected and weighted to minimize the residual noise and maximize the throughput of point-like sources simultaneously. The PCA algorithm performs a principal component analysis of the data, and a low-rank estimate of the on-axis PSF is formed by keeping the first principal components of the decomposition. In the same vein, the LLSG algorithm (Gonzalez et al. 2016) decomposes the data set into low-rank, sparse, and Gaussian components. Because few sources are expected in the field of view, their signatures are mainly recovered in the sparse component. The RSM algorithm (Dahlqvist, Cantalloube & Absil 2020; Dahlqvist, Louppe & Absil 2021a; Dahlqvist, Cantalloube & Absil 2021b) combines residual images obtained with different post-processing algorithms (e.g. cADI and PCA) to leverage their specific benefits and mitigate their respective drawbacks. Since all of these algorithms are based on the estimation and subtraction of an off-axis PSF, they are facing a common pitfall: they fail in deriving a statistically grounded detection map, especially at short angular separations. As a consequence, the identification of candidate sources partly relies on visual inspection of the detection map.

To circumvent this issue, several works have considered alternative ways to reduce the data in order to produce more quantitative outputs. The derivation of a custom signal-to-noise ratio (S/N) through a *t*-test empirically corrected for the varying number of samples as a function of the angular separation (Mawet et al. 2014) is a pioneering work in this direction. Jensen-Clem et al. (2017) recommend the adoption of metrics combining the achievable contrast with the fraction of detected sources. Instead of normalizing the combined residual images by the empirical standard deviation of the noise (roughly) approximated on annuli, Pairet et al. (2019) propose to build a detection map directly from the set of residual images by comparing the variance of samples located on the expected trajectory of putative sources. Other methods reformulate the detection task as an inverse problem. Among them, ANDROMEDA (Mugnier et al. 2009; Cantalloube et al. 2015) and FMMF (Ruffio et al. 2017) build a model of the residual off-axis signal after subtraction of the estimated on-axis PSF. The PACO algorithm (Flasseur et al. 2018a, b, c, 2020a, b) builds a more consistent statistical model, self-calibrated on the data that accounts for the spatial correlations of the nuisance component at the scale of small image patches of a few tens of pixels, see Section 2.1.1.

Given the success of data-driven approaches in solving various high-level imaging tasks (e.g. detection, segmentation, and classification) in very diverse fields (e.g. photography, microscopy, biomedical imaging, and remote sensing), machine learning, and deep learning approaches have also been investigated by the direct imaging community. Fergus et al. (2014) report one of the first steps in this direction with a discriminative approach exploiting the specific structure of high-contrast data. Based on support vector machines, the underlying model is trained from two-classes samples generated by resorting to massive injections of fake companions. Gonzalez, Absil & Van Droogenbroeck (2018) formalize the detection problem as a binary classification task and propose a fully supervised deep learning approach also trained in a supervised fashion. They use collections of patches pre-processed by PCA for different numbers of principal components as input of a random forest or of a convolutional neural network (CNN) that decides in favour of the presence or on the absence of a point-like source in each patch. While demonstrating powerful detection capabilities, this algorithm showed to be prone to a high level of false alarms in some cases (Cantalloube et al. 2020). Besides, the tuning of hyper-parameters remains a critical point making the operating point difficult to reach. Generative adversarial networks (GANs, Goodfellow et al. 2014) have been used to produce multiple realizations of *pure* nuisance component as an alternative way to generate a large basis of labelled samples used to train a deep discriminative model (Yip et al. 2019). Recent works (Samland et al. 2021; Gebhard et al. 2022) recast the unmixing problem as a regularized regression task. The underlying linear model is in charge of explaining the evolution of the nuisance component (and possibly, the evolution of the source signals) in a time-series extracted at a given pixel location from temporal series of reference selected to be signal free and causally independent from the putative source signals.

Intensive testing of PACO, both on public (Cantalloube et al. 2020) as well as on consortia data challenges, and more recently on a subsample of about 75 observations from the SPHERE's SHINE survey (Chomez et al. 2023) shows that PACO is one of the algorithms

of choice to process high-contrast observations. In particular, the latter work (Chomez et al. 2023) demonstrates the expected gain in terms of achievable contrast (up to 10^{-7} at a few arcseconds), and in terms of the underlying exoplanet population (with a mass up to $5\,M_{Jup}$ at 5 au for stars at about 60 pc away). Thanks to its unique data-driven modelling of the nuisance component, accounting for its non-stationary spatial correlations, PACO is especially well suited to process observations in which the typical spatial extent of speckles lies in a patch of a few tens of pixels. However, the statistical model of PACO is approximate in case of spatial correlations spread over a patch of a few tens of pixels (e.g. for background-limited observations and/or in case of unstable observing conditions). This motivates the path we follow in this work: we propose a new detection algorithm, named deep PACO, that combines the statistical model of PACO with a supervised deep learning framework. The statistical model of PACO is used to improve the stationarity and the contrast of the data in a pre-processing step, while deep learning is in charge of correcting for the (putative) approximate fidelity of the statistical model of PACO to the reality of the observations. To do so, the data are centred and whitened locally using the PACO framework, and a CNN is trained in a supervised fashion to detect the residual signature of synthetic sources from pre-processed science data. The network is trained from scratch using full frame samples generated with a custom data augmentation strategy allowing to build a large training set from a single ADI data set. Finally, the underlying discriminative model is applied to the pre-processed observations and delivers a detection map. Detected sources are then photometrically characterized by a second deep neural network, also trained from scratch using patch samples generated with a dedicated data augmentation step. On this latter part, while the recent astronomy literature reports several works for photometry estimation through deep learning models (see e.g. Boucaud et al. 2020; Cabayol et al. 2021; Huertas-Company & Lanusse 2023) for galaxy's photometry or redshift estimation), to the best of your knowledge this is the first time that deep learning is employed to estimate the flux of detected sources in direct imaging at high contrast.

This paper[1] is organized as follows. Section 2 presents the main ingredients of the detection part of the proposed algorithm. Section 3 focuses on the characterization stage of the proposed method. Section 4 evaluates the detection and characterization performance on several high-contrast observations from the InfraRed Dual-band Imager and Spectrograph (IRDIS) imager (Dohlen et al. 2008) of the VLT/SPHERE instrument (Beuzit et al. 2019). Finally, Section 5 draws the paper conclusions and gives future research prospects.

Throughout the text, the reader can refer to Table 1, Figs 1 and 2, and Table B1, summarizing respectively the main notations, the processing pipeline of the proposed approach, and the main setting of the different parameters.

Table 1. Summary of the main notations.

Notation	Range	Definition
		▶ Constants
N	$\mathbb{N}$	Number of pixels in a frame
M	$\mathbb{N}$	Number of pixels in a detection map
T	$\mathbb{N}$	Number of temporal frames
K	$\mathbb{N}$	Number of pixels in a patch (pre-processing)
J	$\mathbb{N}$	Number of pixels in a patch (characterization)
Q	$\mathbb{N}$	Number of samples involved in shrinkage
P	$\mathbb{N}$	Total number of training sources
S	$\mathbb{N}$	Total number of training sets
		▶ Indices
n	$[\![1; N]\!]$	Pixel index
t	$[\![1; T]\!]$	Temporal index
p	$[\![1; P]\!]$	Source index
s	$[\![1; S]\!]$	Training set index
ϕ	$\mathbb{R}_+^2$	2D (subpixel) angular location of a source
		▶ Data quantities
$\boldsymbol{r}$	$\mathbb{R}^{N\times T}$	Observations
$\boldsymbol{f}$	$\mathbb{R}^{N\times T}$	Nuisance component
$\boldsymbol{h}$	$\mathbb{R}^{N}$	Off-axis PSF
$\widetilde{\boldsymbol{r}}$	$\mathbb{R}^{N\times T}$	Pre-processed observations without injections
$\bar{\boldsymbol{r}}$	$\mathbb{R}^{N\times T}$	Observations with injections
$\check{r}$	$\mathbb{R}^{N\times T}$	Pre-processed observations with injections
$\check{\boldsymbol{r}}$	$\mathbb{R}^{N\times T}$	Input (images) of the CNN (detection)
$\check{\boldsymbol{p}}$	$\mathbb{R}^{J}$	Input (patch) of the CNN (characterization)
		▶ Operators
$\mathbf{E}$	$\mathbb{R}^{N\times\cdot} \mapsto \mathbb{R}^{K\times\cdot}$	Patch extractor (pre-processing)
$\mathbf{W}$	$\mathbb{R}^{K\times\cdot} \mapsto \mathbb{R}^{K\times\cdot}$	Centring and whitening (pre-processing)
$\mathbf{D}$	$\mathbb{R}^{\cdot\times T} \mapsto \mathbb{R}^{\cdot\times T}$	Frame derotator by parallactic angles
		▶ Losses and metrics
$\mathcal{L}_{\text{detect.}}$	$[0; 1]^{2\times M} \mapsto \mathbb{R}_+$	Dice2 score (detection loss)
$\mathcal{L}_{\text{carac.}}$	$\mathbb{R}_+^2 \mapsto \mathbb{R}_+$	Absolute relative error (characterization loss)
TPR	$[0; 1]$	True positive rate (detection metric)
FDR	$[0; 1]$	False discovery rate (detection metric)
F1R	$[0; 1]$	Harmonic mean of TPR and FDR (detection metric)
		▶ Estimated quantities
$\widehat{\boldsymbol{y}}$	$[0; 1]^{M}$	Detection map
$\widehat{\alpha}$	$\mathbb{R}_+$	Photometry (in source to star contrast)
$\widehat{\boldsymbol{m}}$	$\mathbb{R}^{N}$	Temporal mean
$\widehat{\rho}$	$[0; 1]$	Shrinkage factor
$\widehat{\mathbf{S}}$	$\mathbb{R}^{K\times K}$	Empirical spatial covariance
$\widehat{\mathbf{C}}$	$\mathbb{R}^{K\times K}$	Shrunk spatial covariance
$\widehat{\mathbf{L}}$	$\mathbb{R}^{K\times K}$	Cholesky's factorization of $\widehat{\mathbf{C}}$

2 SOURCE DETECTION ALGORITHM

2.1 Statistical model of the non-stationary patch covariances

Section 2.1.1 recalls for completeness the main ingredients of the statistical model embedded in the PACO algorithm for ADI observations (Flasseur et al. 2018b, 2020a). Section 2.1.2 describes how to use this model to attenuate the strong and spatially non-stationary correlations of the data as well as to improve the contrast. Resulting residuals from this pre-processing step are centred and whitened observations from which our detection and characterization models are built by supervised deep learning, see Section 2.2. While obtained through the statistical model of PACO, these custom pre-processed observations are not directly produced by the PACO algorithm.

Ablation studies have shown that this pre-processing step is of primary importance: the training step fails to converge in its absence due to the high non-stationarity of the nuisance component and the large fluctuations it displays near the star, see Fig. 1. This effect is due to the fact that standard deep learning architectures assume some degree of invariance, in particular that the data are normalized in a specific range and are corrupted by a stationary noise, see also Section 4.2.1.

[1] A preliminary version of this work was presented in the form of a conference contribution in Flasseur et al. (2022). The present manuscript contains a significant amount of additional methodological developments, technical improvements, and experiments.

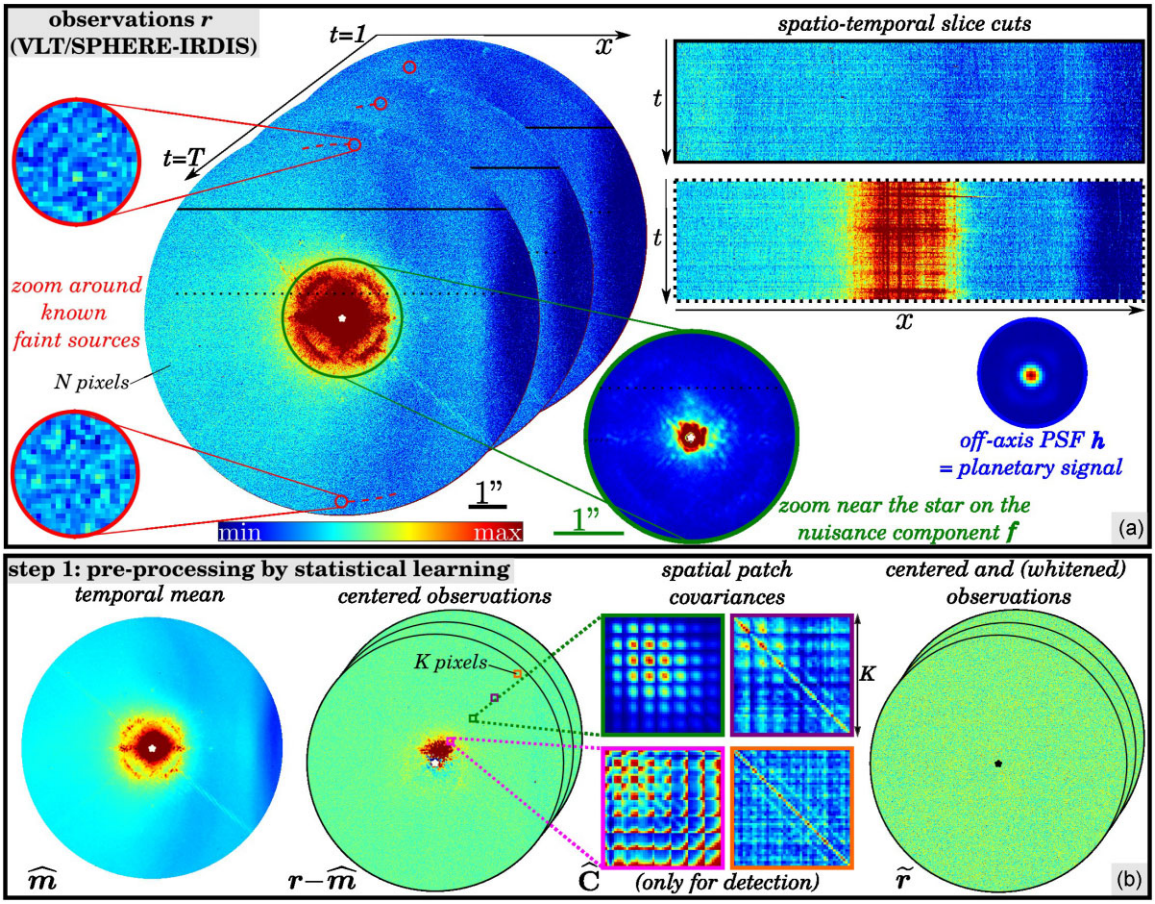

Figure 1. (a) Typical observations $\boldsymbol{r}$ from the VLT/SPHERE-IRDIS instrument conducted in pupil tracking mode (i.e. with the ADI technique). Zooms around two known background faint sources are displayed in the red circles, a zoom on the nuisance component $\boldsymbol{f}$ near the star is highlighted in the green circle, and a view of the sought exoplanetary signal, taking the form of the off-axis PSF $\boldsymbol{h}$, is shown in the blue circle. Two spatio-temporal slice cuts along the black solid and dashed lines are shown on the right, as an illustration of the spatial non-stationarity of the correlations of the nuisance. (b) Illustration of the main operations performed during step 1 of the proposed approach, namely the pre-processing of the observations by statistical learning. The estimation of the mean $\widehat{m}$ and of the covariance matrices $\widehat{\mathbf{C}}$ are based on the PACO model of the nuisance component. Examples of estimated spatial covariance matrices are displayed in squares for four regions of interest pickled at about 0.5 (pink), 1.0 (green), 1.5 (purple), and 2.0 (orange) arcsec. It results from this pre-processing step centred and whitened observations from which our detection and characterization models are built by supervised deep learning, see Fig. 2. Data set: HIP 72192 (2015 May 10), see Section 4 for the recording logs.

2.1.1 *Statistical model of the nuisance component*

A data set $\boldsymbol{r}$ in $\mathbb{R}^{N\times T}$ recorded with the ADI technique is formed by N-pixel images captured at different times t in $[\![1;T]\!]$. The direct model for the observed intensity is:

$$\boldsymbol{r} = \boldsymbol{f} + \sum_{p=1}^{P} \alpha_p \, \boldsymbol{h}(\phi_p), \tag{1}$$

where $\boldsymbol{f}$ in $\mathbb{R}^{N\times T}$ is the nuisance component, and $\boldsymbol{h}\left(\phi_p\right)$ in $\mathbb{R}^{N\times T}$ stands for the contribution of a point source $p \in [\![1;P]\!]$ with a contrast α_p in $\mathbb{R}_+$ that is assumed constant during the few hours of the total observations. The contribution of a source p takes the form of the off-axis PSF centered at location $\mathcal{F}_t(\phi_p)$ in the tth image, where ϕ_p is its initial location on an image at a reference time t_{ref} (e.g. $t_{\text{ref}} = t_1$). The function $\mathcal{F}_t$ is a geometric transform (typically in ADI, a circular translation with respect to the star located at the centre of the images) modelling the apparent motion of the field of view between the observation configurations at time t_{ref} and time t. The mapping $\mathcal{F}_t$ is deterministic since it depends solely on the measured parallactic angles. Given that very few sources are expected in the field of view, we assume that the measured intensity is the superimposition of the nuisance component and at most one unresolved point-like source p at each pixel location n, that is, multiple sources do not overlap.

In previous works on the PACO algorithm (Flasseur et al. 2018a, b, c), we have proposed to describe the random fluctuations of the nuisance component $\boldsymbol{f}$ by a statistical model whose parameters are estimated in a data-driven fashion. We recall hereafter the main ingredients of this statistical model.

Given the spatial non-stationarity of the nuisance component, the model is built locally at the scale of a patch with an area of a few tens of pixels. It models the distribution of T patches[2] $\boldsymbol{f}_n = \{\mathbf{E}_{n,t}\,\boldsymbol{f}\}_{t=1:T}$ in $\mathbb{R}^{K\times T}$ extracted around pixel n ($\mathbf{E}_{n,t}$ denotes the K-pixel patch extraction operator at location n and time t) with a multivariate Gaussian $\mathcal{N}(\boldsymbol{m}_n, \mathbf{C}_n)$. The covariance matrix $\mathbf{C}_n$ is non-diagonal, that is, it accounts for the local correlations of $\boldsymbol{f}$. The sample estimators $\{\widehat{\boldsymbol{m}}_n; \widehat{\mathbf{S}}_n\}$ of the local mean and covariances coming from

[2]For the convenience, we use in the equations a vectorized version of 2D patches.

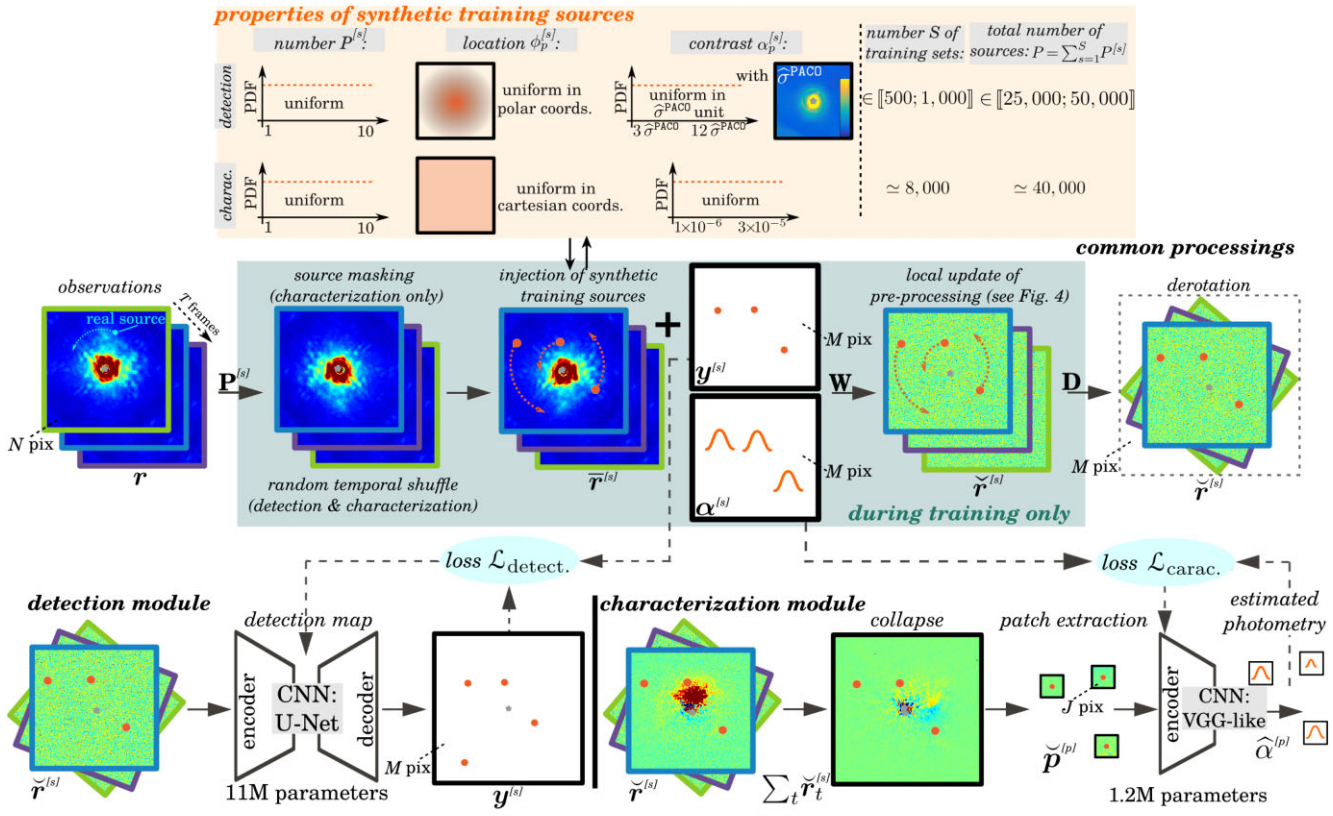

Figure 2. Schematic representation of the main operations performed during the detection and characterization steps of the proposed algorithm by supervised deep learning. The first line displays a view of the main parameters defining synthetic sources injected into the pre-processed observations (see Fig. 1) at training time. The second line shows common operations performed for both the detection and the characterization steps. The left (respectively, the right) part of the third line is for operations applied solely during the detection (respectively, the characterization) step. Throughout this paper, synthetic training sources injected to build our models are highlighted in orange, while (possibly unknown) real and synthetic sources that we aim to detect and to characterize at inference time are displayed in light blue in the schematic representations. Data set: HIP 72192 (2015 May 10), see Section 4 for the recording logs.

the maximum likelihood are the following:

$$\begin{cases} \widehat{\boldsymbol{m}}_n = \frac{1}{T}\sum_{t=1}^{T} \mathbf{E}_{n,t}\,\boldsymbol{r} \;\in \mathbb{R}^K\,, \\ \widehat{\mathbf{S}}_n = \frac{1}{T}\sum_{t=1}^{T} (\mathbf{E}_{n,t}\,\boldsymbol{r} - \widehat{\boldsymbol{m}}_n)(\mathbf{E}_{n,t}\,\boldsymbol{r} - \widehat{\boldsymbol{m}}_n)^\top \in \mathbb{R}^{K\times K}\,. \end{cases} \tag{2}$$

Since the number of samples available, that is, the number T of temporal frames, is typically equivalent or smaller than the number K of pixels in a patch,[3] the sample covariance $\widehat{\mathbf{S}}_n$ is very noisy and may even be rank deficient. A form of regularization must be enforced to stabilize the estimate and allow the inversion of the covariance matrix involved in the data whitening step (see Section 2.1.2). We use a *shrinkage* estimator (Ledoit & Wolf 2004; Chen et al. 2010) formed by a convex combination between the low bias/high variance estimator $\widehat{\mathbf{S}}_n$ and a high bias/low variance estimator $\widehat{\mathbf{F}}_n$:

$$\widehat{\mathbf{C}}_n = (1-\widehat{\rho}_n)\,\widehat{\mathbf{S}}_n + \widehat{\rho}_n\,\widehat{\mathbf{F}}_n, \tag{3}$$

[3]The number of pixels in a patch is determined in a data-driven fashion, as described in Flasseur et al. (2018b) for the PACO algorithm. It corresponds to twice the FWHM of the measured off-axis PSF at the given wavelength. In practice, this empirical rule typically leads to K in $[\![7^2; 12^2]\!]$ pixels for the VLT/SPHERE instrument operating at a wavelength $\lambda \in [0.9; 2.2]$ μm.

where $\widehat{\mathbf{F}}_n$ is a diagonal matrix encoding the sample variances:

$$\left[\widehat{\mathbf{F}}_n\right]_{kk'} = \begin{cases} \left[\widehat{\mathbf{S}}_n\right]_{kk'} & \text{if} \quad k = k' \\ 0 & \text{if} \quad k \neq k'\,. \end{cases} \tag{4}$$

The hyper-parameter $\widehat{\rho}_n$ plays a central role since it governs a bias-variance trade-off. In our previous works (Flasseur et al. 2018b, 2021), we have derived its closed-form expression, which is an extension of the results of Chen et al. (2010) in the case of a non-constant shrinkage matrix $\widehat{\mathbf{F}}_n$:

$$\widehat{\rho}_n = \frac{\mathrm{tr}\left(\widehat{\mathbf{S}}_n^2\right) + \mathrm{tr}^2\left(\widehat{\mathbf{S}}_n\right) - 2\sum_{k=1}^{K}\left[\widehat{\mathbf{S}}_n\right]_{kk}^2}{(Q+1)\left(\mathrm{tr}\left(\widehat{\mathbf{S}}_n^2\right) - \sum_{k=1}^{K}\left[\widehat{\mathbf{S}}_n\right]_{kk}^2\right)}, \tag{5}$$

where Q is the number of non-null patches involved in the computation of $\widehat{\mathbf{S}}_n$. Here, Q is equal to T everywhere.

In Appendix A1, we discuss a refinement of this statistical model to account for the temporal fluctuations of the observations. It leads to a slight improvement in terms of detection performance at the cost of an increase of the computational burden by a factor 10–30. For these reasons, it is not applied by default in the following. We recommend to use it, in a second step, to refine the analysis of ambiguous candidate detections found by the proposed method embedding a multivariate Gaussian model.

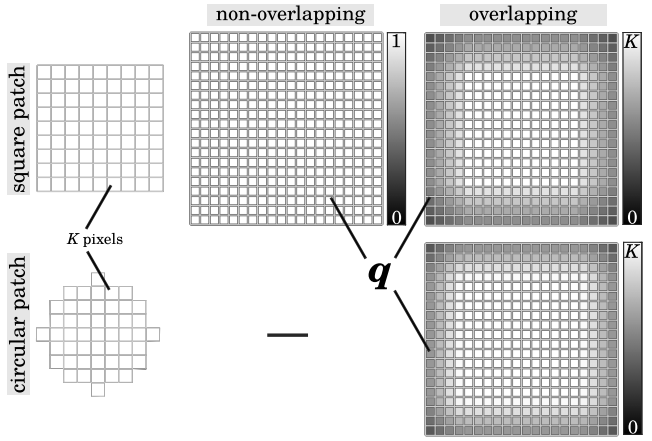

Figure 3. Schematic view of the quantity $\boldsymbol{q}$, representing the number of patches contributing to the computation of the pre-processed observations $\widetilde{\boldsymbol{r}}$ at each pixel of the field of view, as a function of the patch shape and of the tessellation of the field of view. Non-overlapping circular patches are not considered in this work since they do not allow a complete paving of the field of view. The spatial scale is not respected for the purpose of illustration.

2.1.2 *Centring and local whitening of the observations*

We consider a set of locations $\mathbb{P}$ where the statistics of the nuisance component should be computed. The cardinal of $\mathbb{P}$ depends solely on the patch shape and on the *patch stride*[4] used to cover the whole field of view. For a given patch stride, we first define the number of (centred and whitened) patches averaged at each location n' of the field of view:

$$q_{n'} = \sum_{n \in \mathbb{P}} \delta_{\mathbf{1}_n \mathbf{1}_{n'}^\top \neq 0}\,, \forall n' \in [\![1; N]\!], \tag{6}$$

where δ is the indicator function (i.e. $\delta_{x=y}$ is either equal to 1 if the condition $x = y$ is met, and equal to 0 otherwise), and $\mathbf{1}_n \in \mathbb{R}^K$ (respectively, $\mathbf{1}_{n'} \in \mathbb{R}^K$) is the vectorization of a one-valued patch centered at location n (respectively, n'). Fig. 3 gives a view of the quantity $\boldsymbol{q} \in \mathbb{N}^N$ in the general case, that is, for either square and circular patches as well as non-overlapping and overlapping patches. By default, we consider square patches of K pixels area. The pre-processed images $\widetilde{\boldsymbol{r}}$ in $\mathbb{R}^{N \times T}$, after centring and whitening, are obtained by:

$$\begin{aligned}\widetilde{r}_{n'} &= \frac{\left[\sum_{n \in \mathbb{P}} \mathbf{E}_n^\top \mathbf{W}_n \mathbf{E}_n \boldsymbol{r}\right]_{n'}}{q_{n'}}, \\ &= \frac{\left[\sum_{n \in \mathbb{P}} \mathbf{E}_n^\top \widehat{\mathbf{L}}_n^\top (\boldsymbol{r}_n - \widehat{\boldsymbol{m}}_n)\right]_{n'}}{q_{n'}}\,, \forall n' \in [\![1; N]\!],\end{aligned} \tag{7}$$

where $\mathbf{W}_n$ is an operator performing centering and whitening of the collection of patches $\boldsymbol{r}_n \mathbb{R}^{K \times T}$ at location n, such that $\widehat{\mathbf{L}}_n$ is the Cholesky's factorization of $\widehat{\mathbf{C}}_n^{-1}$ (i.e. $\widehat{\mathbf{L}}_n \widehat{\mathbf{L}}_n^\top = \widehat{\mathbf{C}}_n^{-1}$). In the specific case (considered by default in this paper) of non-overlapping square patches of K pixels, $\mathrm{card}(\mathbb{P}) = \lfloor N/K \rceil$, and equations (6) and (7) simplify as:

$$\begin{cases} q'_n = 1\,, \forall n' \in [\![1; N]\!]\,, \\ \widetilde{\boldsymbol{r}}_n = \mathbf{W}_n \boldsymbol{r}_n = \widehat{\mathbf{L}}_n^\top \left(\boldsymbol{r}_n - \widehat{\boldsymbol{m}}_n\right)\,, \forall n \in \mathbb{P}\,. \end{cases} \tag{8}$$

[4]We define the patch stride as the distance (in pixels) between the centres of two adjacent patches, both in the x- and y-directions.

In Appendix A2, we discuss a refinement of the intermediate quantity $\widetilde{\boldsymbol{r}}$ produced by the pre-processing step under the same statistical model as described in Section 2.1.1 or in Appendix A1. It can be noted that overlapping patches should be used with this refinement, and that the patch shape can be either squared or circular. This alternative approach leads to a better stability and robustness of the method for data sets recorded under bad observing conditions.[5] However, the computational burden of this variant is increased by a factor $2 \times K$, typically lying in $[\![100; 250]\!]$ for the VLT/SPHERE instrument. For these reasons, this variant is not applied by default in the following. We recommend to apply it, in a second step, when the training, validation, or inference results are clearly impacted by a significant number of false alarms, much higher than expected, which is an unambiguous sign of the limited fidelity to the observations of the pre-processing procedure used by default.

2.2 Exoplanet detection by supervised deep learning

We formalize the detection problem as a supervised pixelwise classification task: starting from a temporal series of pre-processed images including synthetic sources, the goal is to infer a detection map $\widehat{\boldsymbol{y}}$ in $[0; 1]^M$, where each pixel value represents a score between 0 and 1 such that a high (respectively, a low) score values the presence (respectively, the absence) of a source centred at that location. Interpreting this score as a true probability of presence of a source requires a control of the uncertainties with dedicated methods (see e.g. Gawlikowski et al. 2023; Hüllermeier & Waegeman 2021, for recent review papers) that is left for future work. For this reason, in the following, we refer to this score as a *pseudo-probability*. Besides, the M-pixel detection map is larger than a N-pixel single temporal frame of the pre-processed data. Due to the apparent rotation induced by ADI, it is theoretically possible to detect a source lying in the sensor field of view at a single date, see Figs 2(c) and 9. By derotating each individual temporal frame with the corresponding parallactic angle and combining the resulting transformed measurements, an area larger than N pixels can be scrutinized. Its resulting spatial extent depends solely on the total amount of parallactic rotation between the first and the last frame.

Section 2.2.1 details the construction process of the training set, Section 2.2.2 describes the selected model architecture, and Section 2.2.3 discusses the metrics we consider to evaluate the performance of the proposed method.

2.2.1 *Construction of the training basis*

In high-contrast imaging, obtaining real ground-truth data is a twofold challenge.

First, the overall number of positive samples is limited as relatively few point-like sources have been confirmed to date. Second, negative samples are hard to define since some undiscovered sources might be present in the observed data. To overcome these difficulties, we

[5]Since the notion of *observing conditions* is relative and can be characterized by several metrics (e.g. Strehl ratio, coherence time, air mass, etc.), we did not find quantitative values for these measures indicating the strict use of the variant version described in Appendix A2. This large effort would require the processing of hundreds of data sets, which is left for future work. Qualitatively, we observed that data sets impacted by a bright wind-driven halo, displaying the same apparent motion than the objects of interest, are more subject to lead to an increased false alarm rate without the pre-processing described in Appendix A2.

adopt the following training strategy: the training set consists of S pairs $\{\check{\boldsymbol{r}}^{[s]}; \boldsymbol{y}^{[s]}\}_{s=1:S}$ of samples resulting from the massive injection of synthetic point-like sources. In this framework, $\check{\boldsymbol{r}}^{[s]} \in \mathbb{R}^{N\times T}$ represents observations, with injected synthetic sources, that have been pre-processed. The quantity $\boldsymbol{y}^{[s]} \in [\![0, 1]\!]^M$ is the ground-truth map pointing the injection locations of any synthetic training source falling within the field of view at least in one temporal frame. The ground truth map is built for a given (and arbitrary) orientation of the field of view, for example, aligned with the true North. The implemented simulation process is quite realistic since the injected source signature corresponds to the off-axis PSF of the target star usually measured just before or just after the main sequence of observations by decentring the coronagraph.

Second, the nuisance component varies drastically from one observation to the other, as it is highly dependent on the observing conditions, the magnitude of the star, and the instrument settings. As a consequence, we follow an observation-dependent approach, and train a different model on each observation. It means that the model parameters (except algorithmic and optimization hyper-parameters, see Section 2.2.4) are optimized from scratch for each data set.

This setup implies the design of a custom data-augmentation strategy (i) to prevent overfitting of the model that is trained from a unique temporal series of images, and (ii) to account for our lack of knowledge about real sources – *unknown at training time but that we aim to detect at test time*. To circumvent these issues, we apply a random permutation of the T images forming the observations $\boldsymbol{r}$ for each new training sample $s \in [\![1; S]\!]$. This operation allows us (i) to create artificially different training sets, and (ii) to break the temporal consistency of (known and unknown) real sources. Synthetic sources are then injected in the temporally permuted data using the parallactic angles and the best fit of the off-axis PSF by a Gaussian and an Airy pattern. Besides, the off-axis PSF is assumed to be time-invariant. We have checked numerically that this assumption is reasonable for our classification task. Similarly, assuming a slightly different pattern for the off-axis PSF (e.g. measured *versus* fitted model) between the data generation process and the training step does not lead to a significant drop in the detection performance. At this intermediate stage, each training sample $\bar{\boldsymbol{r}}^{[s]}$ is obtained by:

$$\bar{\boldsymbol{r}}^{[s]} = \mathbf{P}^{[s]}\,\boldsymbol{r} + \sum_{p=1}^{P^{[s]}} \alpha_p^{[s]}\,\boldsymbol{h}\left(\phi_p^{[s]}\right), \tag{9}$$

where $\mathbf{P}$ is an operator performing the random temporal permutation of the images of $\boldsymbol{r}$ and $\boldsymbol{h}\left(\phi_p\right)$ in $\mathbb{R}^{N\times T}$ represents the spatio-temporal contribution of a synthetic source centered at location ϕ_p on a reference image at time t_{ref}, see Introduction. The number of sources $P^{[s]}$, their contrasts $\{\alpha_p^{[s]}\}_{p=1:P^{[s]}}$ and their initial locations $\{\phi_p^{[s]}\}_{p=1:P^{[s]}}$ are free parameters. In practice, the number $P^{[s]}$ of injected sources in each training sample is drawn uniformly in $[\![1; 10]\!]$. This setting represents a realistic scenario since we expect a few faint point-like sources in the field of view. We denote by P the total number of injected sources over the S training sets, that is, $P = \sum_{s=1}^{S} P^{[s]}$. The initial locations $\{\phi_p\}_{p=1:P}$ of the injected sources are drawn uniformly in polar coordinates (i.e. the centre of the field of view is more sampled than farther away). Note that we have compared this setting with a uniform sampling in Cartesian coordinates (i.e. uniform density over the field of view), and we found very similar detection performance for both settings. The selected one (i.e. uniform in polar coordinates) slightly speeds up the training procedure, likely because the pre-processed observations fluctuate slightly more near the star than farther away, thus requiring more training samples to discriminate a source from the nuisance. The range of injected flux is also a critical choice. For instance, if the lower bound is too low, class overlap can occur and the model is not able to discriminate between sources and the nuisance component leading to a high level of false alarms. In the opposite case, if the upper bound is too low, evident bright sources will not be detected since there are no similar examples in the training set. In practice, we set the injection range in an unsupervised fashion. We train our model on sources which are challenging to detect with other methods: the contrast $\{\alpha_p\}_{p=1:P}$ of the injected sources is drawn uniformly in $\left[3\widehat{\sigma}_{\phi_p}^{\texttt{PACO}}; 12\widehat{\sigma}_{\phi_p}^{\texttt{PACO}}\right]$ where $\widehat{\sigma}_{\phi_p}^{\texttt{PACO}}$ is the 1σ contrast reached by PACO at location ϕ_p. This setting covers both sources that are detectable above the standard 5σ detection confidence and sources whose detection confidence remains below the 5σ detection limit reached by PACO. In practice, we found that this setting is suitable to detect both faint sources and bright sources without generating large number of false alarms.

As the pre-processing is an expensive procedure and becomes the bottleneck during online data generation, we adopt a local update strategy to reduce its computational cost. Prior to the injection of synthetic sources, the whole data set is pre-processed, that is, centred and spatially whitened, see Fig. 1(b). We denote by $\widetilde{\boldsymbol{r}}$ the pre-computed cube. After each batch s of injections, the set $\mathbb{S}^{[s]}$ of locations impacted by the signal of the $P^{[s]}$ sources is determined. Outside $\mathbb{S}^{[s]}$, the pre-processed images are obtained from the temporal permutation of $\widetilde{\boldsymbol{r}}$. Inside $\mathbb{S}^{[s]}$, the statistics of the nuisance component are updated given the contamination of the $P^{[s]}$ injected sources, and the pre-processed images are updated with these refined statistics. At this intermediate stage, each training sample $\check{\boldsymbol{r}}^{[s]}$ is obtained by:

$$\check{\boldsymbol{r}}_n^{[s]} = \begin{cases} \mathbf{W}_n\,\bar{\boldsymbol{r}}_n^{[s]}, & \text{for } n \in \mathbb{S}^{[s]} \cap \mathbb{P}\,, \\ \mathbf{P}^{[s]}\,\widetilde{\boldsymbol{r}}_n, & \text{for } n \in \mathbb{P} - \mathbb{S}^{[s]} \cap \mathbb{P}\,. \end{cases} \tag{10}$$

This dual strategy, illustrated by Fig. 4, is applied to prevent any detection bias (i.e. an overestimation of the actual detection performance of the proposed algorithm) since we have shown in previous work on the PACO algorithm (Flasseur et al. 2018b) that the statistics of the nuisance component can suffer from a (slight) bias, in particular at short angular separations and/or when the total amount of parallactic rotation is low. This slight potential bias is due to the contamination of a source whose signal is partly encoded both in the mean and the spatial covariances of the nuisance component.

Finally, the intermediate images of each training sample are derotated with the opposite of the parallactic angles so that signal of the synthetic sources are spatially co-aligned along the temporal axis:

$$\check{\boldsymbol{r}}^{[s]} = \mathbf{D}\,\check{\boldsymbol{r}}^{[s]}, \tag{11}$$

where $\mathbf{D}$ is a derotation operator. This derotation step is mandatory to perform a semantic segmentation with the CNN we consider (see Section 2.2.2) given the limited spatial extent of its receptive field.

The binary ground-truth segmentation map $\boldsymbol{y}^{[s]}$ is obtained by setting to *1* circular areas centred at the locations $\{\phi_p\}_{p=1:P^{[s]}}$ of the $P^{[s]}$ injected sources. Other regions of $\boldsymbol{y}^{[s]}$ are set to 0. The radius of the circles is set to the full width at half-maximum (FWHM) of the off-axis PSF, which corresponds to the expected spatial extent of an exoplanetary signature in the data.

A schematic summary of the construction of the training set is given in Fig. 2.

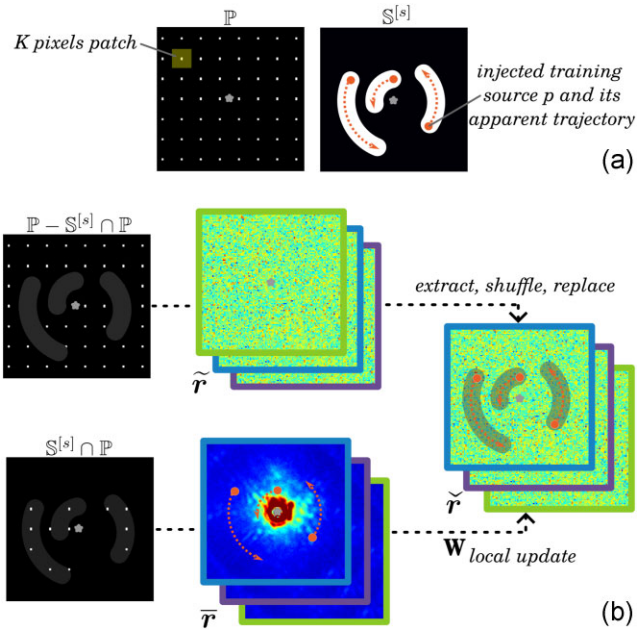

Figure 4. Schematic illustration of the local update of the pre-processing embedded in the training step of the proposed detection algorithm. (a) Illustration of the sets $\mathbb{P}$ and $\mathbb{S}^{[s]}$ for a given training set s with three injected training sources displayed in orange. (b) Illustration of the computation of the pre-processed observations $\check{r}$ in the presence of injected training sources from pre-computed (i.e. before injections) centred and whitened observations $\widetilde{r}$. The spatial scale is not respected for the purpose of illustration.

2.2.2 *Model and architecture*

Deep CNNs reach state-of-art performance in solving pixelwise classification tasks for various imaging fields including microscopy, astronomy, medical imaging, or remote sensing. A large variety of model architecture has been studied in the literature (e.g. auto-encoder, Badrinarayanan, Kendall & Cipolla 2017; VGG, Simonyan & Zisserman 2014; ResNet, He et al. 2016) and their performance often rely on an intricate trade-off between model complexity, the amount of data available for training, and their fidelity with the data used at test time. A common feature of some classical deep architectures is to encode the diversity of the training samples in a low-dimensional subspace by transforming the network input with a cascade of convolution and downsampling operations. Starting from this latent representation, the initial image size is retrieved by a decoder performing reverse transformations with a cascade of deconvolution and upsampling operations.

We also based our model on the above-mentioned category of architectures. We chose a U-Net (Ronneberger, Fischer & Brox 2015) with a ResNet18 (He et al. 2016) as encoder backbone ($\simeq$11 millions of free parameters), which is an architecture widely used for image segmentation. Its residual connections preserve of the input's spatial information along the cascade of convolution and downsampling operations thanks to a direct mapping of the output of each layer of the compression arm into the corresponding layer of the decompression arm. We use the architecture implemented in the SMP package.[6] The encoder and decoder parts are formed by four blocks, each one is composed by a series of convolution layers, batch normalization layers, rectified linear unit (ReLU) activation, and max pooling layers. The final layer of the network has a sigmoid activation function to produce a detection map $\widehat{\boldsymbol{y}} \in [0;1]^M$. The detailed description of the architecture, the number of parameters per layer, and the input/output shapes of each layer can be found at the above-mentioned link.

The network weights are trained from scratch with the samples generated with the procedure described in Section 2.2.1. Initial weights are drawn uniformly through a He-Kaiming distribution (He et al. 2015). The SMP package also provides pre-trained weights. Pre-training is performed with the ImageNet data set (RGB conventional images) either in a supervised, semi-supervised, or weakly supervised learning fashion (Yalniz et al. 2019). In case of pre-training, the weights of the first convolutional layer are replicated in order to match the T-depth of our inputs. We compared all of these strategies against a supervised learning from scratch with our custom training set (see Section 2.2.1). We found similar performance with all approaches, and opted for an end-to-end learning.

2.2.3 *Loss function and accuracy metrics*

The design of loss function used for optimizing the network weights at training time is driven by three criteria: (i) handling with the strong class imbalance (the number of background pixels being much larger than the number of pixels from the sources), (ii) being computationally efficient, and (iii) matching the astrophysical goals (i.e. having a measure close to a detection accuracy score). We compare losses classically used for semantic segmentation, such as the binary cross-entropy (BCE), ℓ_1 and ℓ_2 norms, mean square error and Hinge loss. We have also compared losses based on a similarity measure such as the Dice score (Milletari, Navab & Ahmadi 2016) and hybrid losses combining at least two individual loss measurements (e.g. BCE with Dice score). Our experiments have consistently shown better performance with Dice-based scores. We selected the Dice2 loss (the 2 means for *two* classes), first introduced for biomedical imaging segmentation with very unbalanced classes (Sudre et al. 2017; Wang et al. 2020). Given a training set of ground-truth and predicted detection maps $\{\boldsymbol{y}^{[s]}; \widehat{\boldsymbol{y}}^{[s]}\}$, the Dice2 score is defined by:

$$\mathcal{L}_{\text{detect.}}\left(\boldsymbol{y}^{[s]}, \widehat{\boldsymbol{y}}^{[s]}\right) = 1 - \underbrace{\frac{\sum_{m=1}^{M}(1-\boldsymbol{y}_m^{[s]})(1-\widehat{\boldsymbol{y}}_m^{[s]}+\epsilon)}{\sum_{m=1}^{M} 2-\boldsymbol{y}_m^{[s]}-\widehat{\boldsymbol{y}}_m^{[s]}+\epsilon}}_{\text{background error}} - \underbrace{\frac{\sum_{m=1}^{M}\boldsymbol{y}_m^{[s]}\widehat{\boldsymbol{y}}_m^{[s]}+\epsilon}{\sum_{m=1}^{M}\boldsymbol{y}_m^{[s]}+\widehat{\boldsymbol{y}}_m^{[s]}+\epsilon}}_{\text{source error}}, \tag{12}$$

where ϵ is a minimum-value smoothing and stability parameter added to avoid division by zero. Targeted loss property (i) is satisfied since errors in the source and background areas are penalized equally regardless the relative occurrence of these two classes in $\boldsymbol{y}^{[s]}$. Property (ii) is also satisfied, and we illustrate numerically in the two following paragraphs that property (iii) is also reached.

At validation time, we evaluate the ability of the model to detect point-like sources while simultaneously avoiding false alarms at best as possible. In other words, we aim to obtain a model obeying a precision-recall trade-off. For a predicted detection map $\widehat{\boldsymbol{y}}^{[s]}$ in

[6]The SMP package containing the network architecture used in this paper is available at https://github.com/qubvel/segmentation_models.pytorch.

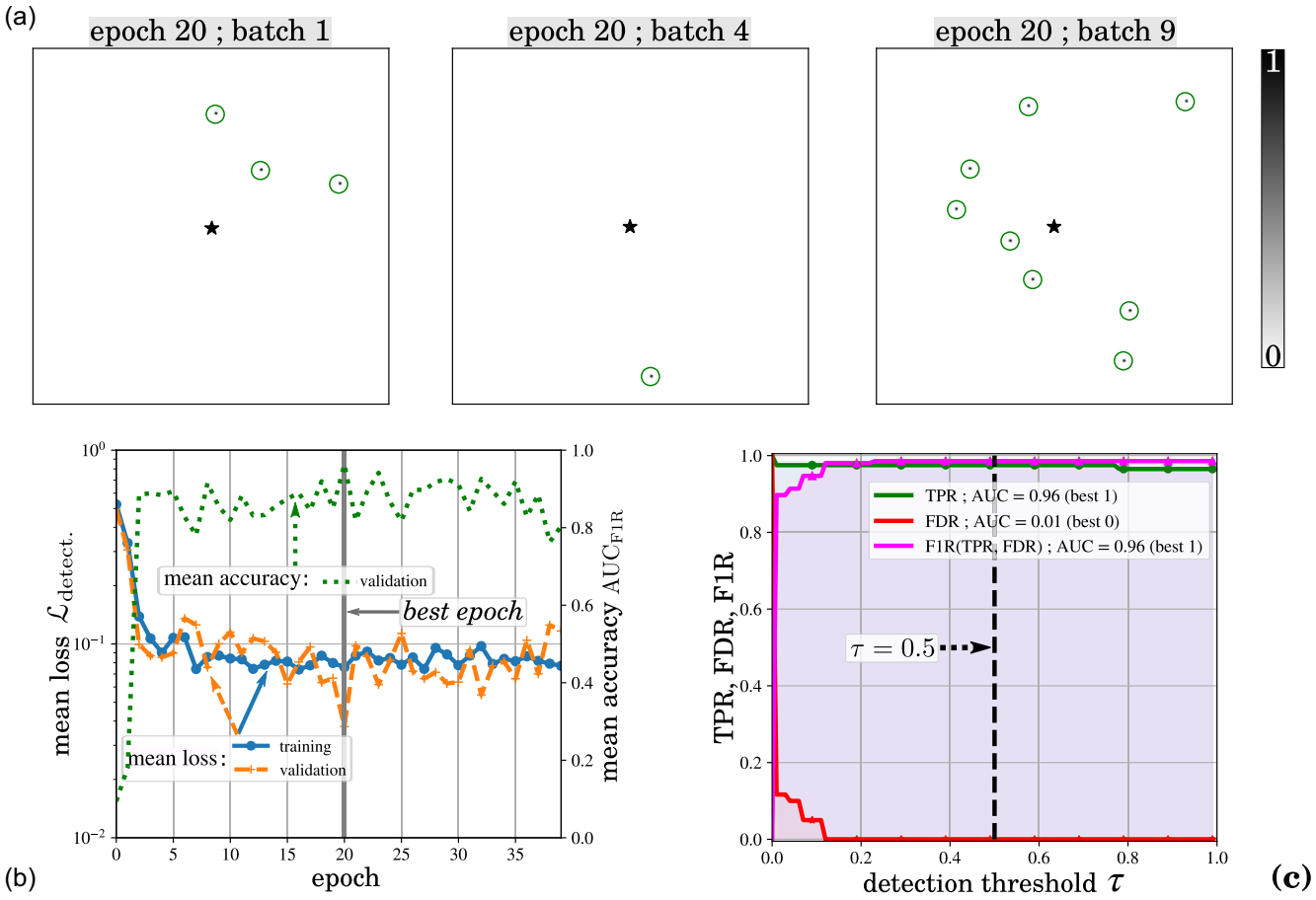

Figure 5. Example of training and validation results. (a) Examples of detection maps obtained at validation time for the best epoch (number 20). (b) Evolution of the loss function at training and validation time, as well as the evolution of the F1R accuracy metric at validation time. (c) ROCs representing the TPR, FDR, and F1R as a function of the prescribed detection threshold τ for the best epoch (number 20, symbolized by the grey vertical line in panel b). Data set: HD 95086 (2015 May 05), see Section 4 for the recording logs.

$[0;\,1]^{\mathcal{M}}$ thresholded at τ in $[0;\,1]$, we count the number of true positives (TP, i.e. true detections), false positives (FP, i.e. false alarms), and false negatives (FN, i.e. missed detections). Following standard practice in direct imaging (see e.g. Gonzalez, Absil & Van Droogenbroeck 2018; Flasseur et al. 2018b; Cantalloube et al. 2020), detections are treated as blobs of one resolution element radius which corresponds to the expected spatial extent of an exoplanetary signature in the data. From TP, FP, and FN, we derive the true positive rate (TPR, i.e. the recall), the false discovery rate (FDR, i.e. the precision), and the F1R score, which is the harmonic mean between TPR and FDR, that we use as an overall measure of the precision-recall trade-off:

$$\begin{cases} \text{TPR} = \frac{\text{TP}}{\text{TP}+\text{FN}} \in [0;\,1], \\ \text{FDR} = \frac{\text{FP}}{\text{FP}+\text{TP}} \in [0;\,1], \\ \text{F1R} = \frac{2}{\frac{1}{\text{TPR}}+\frac{1}{\text{FDR}}} = \frac{2\text{TP}}{2\text{TP}+\text{FN}+\text{FP}} \in [0;\,1]. \end{cases} \tag{13}$$

From TPR, FDR, and F1R, receiver operating curves (ROCs; Kay 1993) are built. ROCs are obtained by evaluating the figures of merit defined in equation (13) as a function of the detection threshold τ. Finally, the area under the curve (AUC) for the F1R score is computed as an aggregate score of the model performance (best when close to 1). Gonzalez et al. (2016), Gonzalez, Absil & Van Droogenbroeck (2018), Flasseur et al. (2018b), Cantalloube et al. (2020), Dahlqvist, Cantalloube & Absil (2020) and Daglayan et al. (2022) exemplified the relevance of ROCs in high-contrast imaging to derive an aggregate measurement of the targeted precision-recall trade-off.

Fig. 5(a) displays some examples of detection maps obtained at validation time for the best validation epoch. These maps illustrate qualitatively the ability of our model to detect synthetic sources while simultaneously avoiding false alarms. Fig. 5(b) shows the evolution of the empirical risk (see equation 12) at training and validation time as well as the evolution of the F1R accuracy metric (see equation 13) at validation time. The loss function does not exhibit significant discrepancy between training and validation steps and the convergence is reached in a few epochs.[7] Besides, the accuracy score is high and well anticorrelated with the loss. This latter observation illustrates that the loss function is a satisfactory estimate of the overall accuracy metric [see targeted property (iii) at the beginning of Section 2.2.3]. Finally, Fig. 5(c) gives an illustration of validation

[7]In machine or deep learning, an epoch refers to a group of multiple training sets from which the network weights are optimized sequentially by stochastic gradient descent. Multiple epochs, formed by a random selection and ordering of some training sets taken from the training base, are generally needed to reach convergence of the network weights. Learning and optimization hyper-parameters can also be tuned between two consecutive epochs according to a pre-defined scheduling, see Section 2.2.4.

ROC obtained for the best epoch (symbolized by a grey vertical line in Fig. 5b). The validation ROC confirms the good ability of the trained model to discriminate (synthetic) point-like sources from the nuisance component.

2.2.4 *Implementation details*

Pairs of samples $\{\check{\boldsymbol{r}}^{[s]}; \boldsymbol{y}^{[s]}\}$ are generated on the fly at training and evaluation time following the procedure described in Section 2.2.1. To avoid overfitting, each realization s is unique with no repetition for the different epochs. The notion of *epoch* is used only as a way to evaluate regularly the performance of the model with the validation procedure described in Section 2.2.3, and also to adapt the learning rate through a pre-defined scheduling. The optimization of the network weights is performed with an iterative stochastic gradient-descent strategy on minibatches of samples formed (possibly) by the concatenation of multiple training sets. It means that, at each iteration, the model weights are updated in the opposite direction to the gradient of the loss. The loss is evaluated from the current minibatch of samples with respect to the model weights. Since we work on series of T images, with T typically lying between 50 and 100 images, the batch size (i.e. the number of training sets comprised within a minibatch) is fixed at 1 given memory constraints. This setting does not affect the overall performance of the method and only requires to perform more iterations to reach convergence since the cost function is quite noisy, see Fig. 5(a). Even under this setting, the convergence is typically reached in a few epochs, see Section 2.2.3 and Fig. 5(b). In practice, for each training epoch, $S = 100$ pairs of samples $\{\check{\boldsymbol{r}}^{[s]}; \boldsymbol{y}^{[s]}\}$ are generated and fed sequentially as input of the network. For each validation epoch, S is fixed at 10 given the computational burden required to build ROCs representing the F1R score as a function of the detection threshold τ. The training process stops when the accuracy metric $\mathrm{AUC}_{\mathrm{F1R}}$ (i.e. AUC under ROC representing the F1R score as a function of the detection threshold τ at validation time) evolves in less than 2 per cent during the 10 previous epochs. The model optimization is performed with the adaptive gradient algorithm AMSGrad (Reddi, Kale & Kumar 2019) which is a variant of the Adam (Kingma & Ba 2014) optimizer with a longer term memory of past gradients. The parameters of the optimizer and of the scheduler have been fine tuned on two data sets and are kept constant for all our experiments. In practice, we observed that the optimized values are quite robust with respect to the data set diversity. The weight decay[8] is fixed at 10^{-5} and the initial learning rate is set to 10^{-3} with a regular decrease by 10 per cent every 10 epochs. The optimization of the network weights is performed with the high-performance deep learning library PyTorch (Paszke et al. 2019) on GPUs server with NVIDIA system equipped with either Tesla V100 or GTX 1080 Ti cards. The pre-processing step is highly parallelized and has a double implementation so that it can performed either on CPUs or on GPUs depending on the number of available CPUs cores and on the server specifications.

3 SOURCE CHARACTERIZATION ALGORITHM

Once sources have been detected, they can be characterized in terms of astrometry and photometry. In this section, we present a new method based on supervised deep learning to estimate the photometry of detected point-like sources. The subpixel estimation of the astrometry is not addressed in this paper because it requires a subpixel estimation of the detection criterion as well as statistical guarantees on its significance. These specific developments are left for future work, and the proposed characterization algorithm can be use to estimate the photometry at the pixel barycentre of (candidate) sources revealed by the detection stage presented in Section 2. As in Section 2, we successively discuss the pre-processing stage, the formalization of the problem as a regression task, the model, and the underlying architecture, the metric we use for training and evaluation, and some implementation details.

3.1 Pre-processing aspects

We adopt a simple patch-based approach, in which we predict the flux of a putative source from a unique (reduced) patch centred on the approximate location of the source. We propose to parametrize the mapping between the input patch and the flux with a CNN trained in a supervised fashion, from the data set of interest (i.e. the underlying model is data-dependent, as for the detection stage of this paper).

The data set is first reduced to a single frame, from which the input patches are extracted. This pre-processing consists in three steps. First, the temporal mean is computed and subtracted pixelwise in order to remove most of the quasi-static speckles. Second, the data set is derotated by the opposite of the parallactic angles to co-align the signal of the sources. Third, the data set is averaged temporally, resulting in a single-averaged frame. This last step allows to obtain an efficient training procedure as it reduces the size of the input data by a factor T. Besides, we observed empirically that this step is beneficial to improve the estimation accuracy as it acts as a simple denoiser: the source signal is constant along time, while residual speckles are not quasi-static after cube derotation, thus cancelling out. Keeping the notation introduced in Section 2, these operations transform a given (intermediate) training data set $\bar{\boldsymbol{r}}$ in $\mathbb{R}^{N\times T}$ with injected synthetic sources as follows:

$$\begin{cases} \check{r}_n = \bar{r}_n - \widehat{m}_n\,, \forall n \in \mathbb{P}, & \text{(step 1)}, \\ \check{\boldsymbol{r}} = \frac{1}{T}\sum_{t=1}^{T} [\mathbf{D}\,\check{r}]_t\,, & \text{(steps 2 and 3)}. \end{cases} \tag{14}$$

In this framework, and unlike the detection stage, we do not apply a whitening of the spatial correlations at step 1. Indeed, we observed empirically that whitening the data set between steps 1 and 2–3 degrades the performance of our model, as illustrated by Fig. 6. More quantitatively, the absolute error of estimation is increased, whatever the angular separation, by a factor between 3 and 5. Besides, keeping a whitening step for photometry estimation does not allow to obtain better results than PACO for most of the field of view. This is expected as the whitening distorts the shape and the norm of the source signal, thus hampering the recovery of its flux. The detection algorithm is not subject to this constraint as its task is to determine whether a source is present or not, regardless of its flux. This fact illustrates that deriving a quantitative result (as a flux estimate) is a more complex task than providing a qualitative result (as related to the presence or to the absence of a source) with our algorithmic setting. We can expect that building the model from several data sets of observations (instead of a single one in this work) would relax these constraints. These specific developments are left for future work, see also Section 5 for a discussion. After pre-processing, square patches $\check{\boldsymbol{p}}^{[p]} \in \mathbb{R}^J$ are finally extracted around the location of each injected synthetic source p during the training and validation steps, or around each (candidate) real point-like source p at inference time. The patch size J is an hyper-parameter whose setting is discussed in more details in Section 3.2.4.

[8]In machine or deep learning, the weight decay refers to a regularization technique reducing the complexity of a model to prevent overfitting.

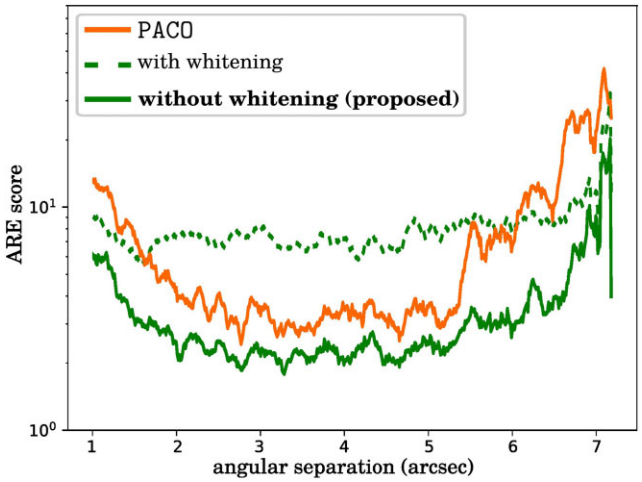

Figure 6. Influence of the whitening of the spatial correlations during the pre-processing step of the characterization algorithm. ARE (see Section 3.2.3 for the definition of the metric) on the estimated photometry is reported as a function of the angular separation, with and without whitening of the spatial correlations. The performance of PACO are displayed as a purpose of comparison. The results are averaged azimuthally for 40 000 sources of flux drawn uniformly between 1×10^{-6} and 3×10^{-5}. The known real sources were masked out and were not considered. Data set: HD 95086 (2015 May 05), see Section 4 for the recording logs.

3.2 Regression by supervised deep learning

3.2.1 Construction of the training basis

As with the detection procedure of Section 5, we resort to massive injections of synthetic sources with various fluxes to build our training basis.

Prior to the injections, the first step consists in masking any real and/or synthetic detected sources that we aim to estimate the photometry at inference time. This operation prevents, as best as possible, data leakages between training and test sets so that training patches do not contain any pixel from patches considered at inference time. In practice, source masking is performed by replacing, for each temporal frame, the local area impacted by the signal of the sources of interest by their pixelwise temporal mean. Like for the generation of the training basis of the detection stage, we also apply, as a data-augmentation strategy, a random permutation of the temporal frames prior to the construction of a training set s from which training samples with injected sources are extracted.

We then build each training set s by injecting a dozen of synthetic sources, with a relative flux (i.e. exoplanet to star contrast) ranging from 1×10^{-6} to 3×10^{-5} with respect to the flux of the host star. Given the current instrumental and processing performance in direct imaging, this setting corresponds to sources with relatively low flux, for which the estimation is usually the most flawed. This operation range can be easily modified in the algorithm to characterize (expected) sources with a lower or higher contrast level, if needed. The contrast of synthetic sources is drawn uniformly in the above-mentioned range, regardless of the angular separation. After injection, each training set is pre-processed and the input patches are extracted following the method described in Section 3.1. This procedure is repeated to get a prescribed number P of training patches $\{\check{\boldsymbol{p}}^{[p]} \in \mathbb{R}^J\}_{p=1:P}$. The number of training patches, hence the total number P of injected synthetic sources, is an additional hyper-parameter whose setting is discussed in more details in Section 3.2.4.

Table 2. Architecture of the proposed CNN for source characterization. The shapes of the layers are indicated for a unit batch size.

Layer	Shape
Input	$1 \times 31 \times 31$
Conv2D + ReLU	$128 \times 25 \times 25$
Conv2D + ReLU	$128 \times 21 \times 21$
Conv2D + ReLU	$256 \times 17 \times 17$
MaxPooling	$256 \times 1 \times 1$
DenseLayer + ReLU	$256 \times 1 \times 1$
DenseLayer	$1 \times 1 \times 1$

3.2.2 Model and architecture

We built a custom network based on VGG, an architecture initially proposed for image classification (Simonyan & Zisserman 2014). The underlying model has 1.2 million of free parameters, and its detailed architecture is described in Table 2. We use a stride[9] of one for all convolutional layers. We have also tested alternative models, both with a deeper and shallower architecture, all leading to worst estimation performance than the selected one. In particular, we experienced a significant degradation of the performance at short angular separations with deeper architectures. The later are the more subject to overfitting (due to the increase in terms of model complexity), especially in the absence of a whitening procedure preventing memorization of the nuisance structures by the network.

From an input patch $\check{\boldsymbol{p}}^{[p]} \in \mathbb{R}^J$, the network produces a single scalar $\widehat{\alpha}^{[p]} \in \mathbb{R}_+$, representing the estimated source's photometry.

3.2.3 Loss function and accuracy metric

Our choice of the loss function $\mathcal{L}_{\text{carac.}}$ is driven by the two following criteria: (i) being computationally efficient, and (ii) matching the astrophysical goals. We chose the absolute relative error (ARE) between the ground-truth and the predicted photometry, which is a classical loss for regression problems:

$$\mathcal{L}_{\text{carac.}}\left(\alpha^{[p]}, \widehat{\alpha}^{[p]}\right) \ (\%) = 100 \times \frac{\left|\alpha^{[p]} - \widehat{\alpha}^{[p]}\right|}{\alpha^{[p]}}. \tag{15}$$

The ARE has the advantage of giving the same contribution to each individual source p regardless of its flux, when it is averaged over multiples ones. Since this metric is computationally very efficient, we also use it to evaluate the overall performance of the method at validation time.

3.2.4 Implementation details

In this section, we successively discuss the setting of the patch size, the number of training sources, the sampling strategy of injected synthetic sources, and some optimization aspects.

Due to the pre-processing described in Section 3.1, part of the signal of the sources can be encoded in the temporal mean that is subtracted to the full frames in order to attenuate the quasi-static speckles. As a result, a (negative) contribution, taking the form of an arc, can spread out along the trajectory of the sources in the reduced frame. This well-known phenomenon in direct imaging is usually referred as *self-subtraction*. As a consequence, we can expect that the performance of the predictor would increase with the patch size

[9]In deep learning, the convolutional stride (in pixels) set how far the convolutional filters move from one node of the image grid to the next one.

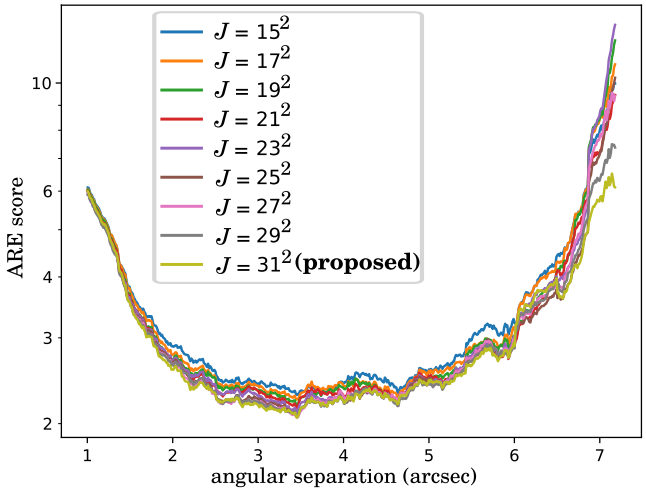

Figure 7. Influence of the patch size J in the characterization algorithm. ARE (see Section 3.2.3 for the definition of the metric) on the estimated photometry is reported as a function of the angular separation, for patches between $J = 15^2$ and 31^2 pixels area. The results are averaged azimuthally for 40 000 sources of flux drawn uniformly between 1×10^{-6} and 3×10^{-5}. The known real sources were masked out and were not considered. Data set: HD 95086 (2015 May 05), see Section 4 for the recording logs.

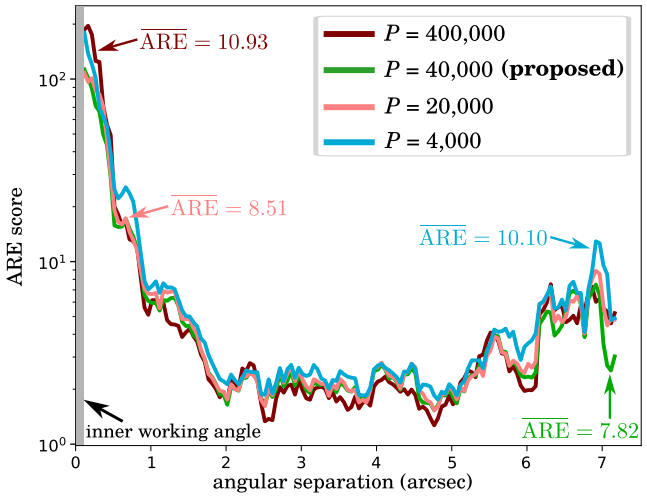

Figure 8. Influence of the number of training sources P in the characterization algorithm. ARE (see Section 3.2.3 for the definition of the metric) on the estimated photometry is reported as a function of the angular separation, for $P \in \{400\,000; 40\,000; 20\,000; 4000\}$ sources. The results are averaged azimuthally for sources of flux drawn uniformly between 1×10^{-6} and 3×10^{-5}. The mean ARE (denoted by $\overline{\text{ARE}}$) averaged over the angular separation is also reported as an overall measure of the performance. The known real sources were masked out and were not considered. Data sets: the 11 SPHERE-IRDIS data sets considered in this work, see Section 4 for the recording logs.

J, to be able to capture the (extended) signature of sources induced by self-subtraction. Besides, increasing the patch size increases the context (i.e. local realizations of the nuisance component) that could be beneficial to unmix the different contributions. As shown by Fig. 7, increasing the patch size is indeed beneficial as it reduced the mean relative error of estimation. The gain is more pronounced as the angular separation increases, which could be due to the larger extent of the self-subtraction signature (as the apparent motion of sources induced by ADI increases with the angular separation). However, large patches are not convenient in the case of adjacent sources, as both signals will be contained in both input patches. As a trade-off, we chose a patch size of $J = 31^2$ pixels, as it encompasses the core of the signal of the source without being impractical when multiple sources are present in the field of view.

The number P of synthetic sources (which also corresponds to the number of patches) used at training time is an additional hyper-parameter obeying a trade-off. On the one hand, it should be large enough to be representative of the variety of real sources in terms of flux and locations. On the other hand, it should be sufficiently small to avoid overfitting on the training set. These expected behaviours are confirmed by numerical experiments presented in Fig. 8. The error at small angular separations increases with P, since this is the area of the field of view the more subject to data leakages when generating multiple samples from a few tens of pixels only. The overall performance also degrades when P is too small. Based on this study, we include $P = 40\,000$ patches in our training set since it leads to the smallest ARE averaged over the whole of view.

Concerning the source's sampling strategy, we are interested at evaluation time in assessing the performance of the model per angular separation. As such, it is natural to sample test sources uniformly in the polar coordinates system. However, polar sampling is detrimental during the training phase, as pixels at short angular separations would be over-represented in the training set, leading to an overfitting of the model in this area. It can be noted that this effect does not occur in the detection stage of the proposed approach given that the training sets consist of full frames of pre-processed observations; each pixel of the field of view being equally represented in the training base. We experimentally observed that sampling training sources uniformly in the Cartesian coordinates system reduces significantly overfitting at short angular separations, without degrading performance in the rest of the field of view. As a result, we opt during training for a uniform sampling of the coordinates of synthetic sources in the Cartesian system.

Concerning the optimization process, pairs of samples $\{\tilde{\boldsymbol{p}}^{[p]}; \boldsymbol{\alpha}^{[p]}\}$ are pre-generated before training and evaluation. As for the detection stage, each realization p is unique to limit overfitting as best as possible. Unlike the detection stage, the characterization stage operates on a single patch (instead of the T temporal images), thus allowing to chose a batch size higher than one to improve the stability and to reduce the computation time of the optimization process. In practice, the batch size is fixed at 1024. The number of epochs is fixed at 300, which shown to be sufficient to reach convergence of the network weights for all the considered observations. For each training epoch, the full set of training samples is fed as input of the network in a random order. The model optimization is performed with Adam (Kingma & Ba 2014) with a learning rate of 10^{-3}. As for the detection stage, the optimization of the network weights is done in PyTorch (Paszke et al. 2019) on either Tesla V100 or GTX 1080 Ti cards. The pre-processing step is highly parallelized, and it is run on CPUs.

4 RESULTS

4.1 Data sets description and reduction strategies

For our comparative analysis, we have selected 15 data sets recorded with the SPHERE-IRDIS instrument.

Three of the 15 SPHERE-IRDIS data sets were extracted from the *exoplanet imaging data challenge* (EIDC) initially designed to ground the detection performance of existing post-processing

Table 3. Observing parameters of ADI sequences from the VLT/SPHERE-IRDIS instrument considered in this paper. Columns are: target name, ESO survey ID, observation date, spectral filter λ, number T of available temporal frames, total apparent rotation Δ_{par} of the field of view, number NDIT of subintegration exposures, individual exposure time DIT, average coherence time τ_0, average seeing, and the first paper reporting analysis of the same data. All the observations are performed with the apodized Lyot coronagraph (Carbillet et al. 2011) of the VLT/SPHERE instrument.

Target	ESO ID	Obs. date	λ (μm)	T	Δ_{par} (°)	NDIT	DIT (s)	τ_0 (ms)	Seeing (arcsec)	Related paper
VLT/SPHERE-IRDIS observations from the EIDC challenge										
IRDIS 1[a]	[a]	[a]	1.625	252	40.3	[a]	[a]	[a]	[a]	Cantalloube et al. (2020)
IRDIS 2[a]	[a]	[a]	1.593	80	31.5	[a]	[a]	[a]	[a]	Cantalloube et al. (2020)
IRDIS 3[a]	[a]	[a]	1.593	228	80.5	[a]	[a]	[a]	[a]	Cantalloube et al. (2020)
VLT/SPHERE-IRDIS observations										
HD 95086	095.C-0298(A)	2015-05-05	2.110	52	18.2	4	64	2.3	0.89	Chauvin et al. (2018)
HD 95086	1100.C-0481(E)	2018-01-05	2.110	70	41.0	10	96	7.8	0.32	Desgrange et al. (2022)
HD 95086	106.21VL.001	2021-03-11	2.110	104	41.4	2	32	7.0	0.77	[b]
HIP 88399	095.C-0298(A)	2015-05-10	1.593	46	34.3	4	64	1.2	1.05	Langlois et al. (2021)
HIP 88399	097.C-0865(A)	2016-04-16	1.593	54	37.3	5	64	2.0	1.45	Langlois et al. (2021)
HIP 88399	1100.C-0481(F)	2018-04-11	1.593	40	31.9	10	96	5.5	0.74	Langlois et al. (2021)
HD 131399	095.C-0389(A)	2015-06-12	2.110	92	36.7	6	16	1.9	0.90	Wagner et al. (2016)
HD 131399	296.C-5036(A)	2016-05-07	2.110	56	39.5	7	32	3.6	0.98	Wagner et al. (2016)
HIP 65426	198.C-0209(E)	2017-02-09	2.110	55	49.1	4	64	4.4	0.82	Chauvin et al. (2017)
HIP 65426	1100.C-0481(G)	2018-05-13	2.110	40	31.7	10	96	4.3	0.81	Cheetham et al. (2019)
HIP 72192	095.C-0389(A)	2015-06-11	2.110	96	17.3	6	16	1.9	1.03	Flasseur et al. (2018b)
HR 8799[c]	095.C-0298(C)	2015-07-04	2.110	112	17.9	8	32	2.3	0.94	Langlois et al. (2021)

Notes. [a]Since the EIDC aimed to perform a *blind* benchmark, information that would allow to identify the data sets are not known. [b]This data set was recorded with the star-hopping technique recently available for the VLT/SPHERE instrument (Wahhaj et al. 2021) and its analysis is not reported yet. We do not exploit the data set associated to the observation of the reference star. The data set associated to the target star (HD 95086) is processed like all other data sets considered in this paper. [c]This data set is only used for additional experiments conducted in Appendix A2.

algorithms for high-contrast imaging (Cantalloube et al. 2020). These data sets are used as a sanity check to assess the ability of the proposed algorithm to detect injected sources at moderate levels of contrast.

To study in more details the performance of the proposed method, we selected 12 additional data sets, mostly part from the SHINE survey of the SPHERE-IRDIS instrument (Desidera et al. 2021; Langlois et al. 2021; Vigan et al. 2021). They were obtained by the observation of the following stars:

1. HD 95086, an A8III-type star of the Carina constellation, hosting an exoplanet discovered by direct imaging with the SPHERE instrument (Rameau et al. 2013a, b). Ten known background point sources are also within the SPHERE-IRDIS field of view. In addition, based on the analysis of `PACO` and `deep PACO` detection maps, we have identified two additional (unbounded) candidate point-like sources. Given their unknown status, we exclude these sources from our general analysis (i.e. there are not considered as true detections or false alarms). We briefly discuss there status in Section 4.2.3.

2. HIP 88399, an F6V-type star of the Vela constellation, without known bounded exoplanet. However, six faint background sources fall within the SPHERE-IRDIS field of view.

3. HD 131399 A, an A1V -type star of a triple system located in the Centaurus constellation, with a known faint point source (HD 131399 Ab) discovered by direct imaging (Wagner et al. 2016). While first supposed to be an exoplanet, follow-up analysis of its astro-photometry show that HD 131399 Ab is more likely a background brown dwarf (Nielsen et al. 2017). Besides, a bright background star falls within the SPHERE-IRDIS field of view.

4. HIP 65426, an A8III-type star of the Carina constellation, hosting an exoplanet discovered by direct imaging with the SPHERE instrument (Chauvin et al. 2017).

5. HIP 72192, an A0V-type star of the Lupus constellation, without known bounded exoplanet. However, two faint background sources fall within the SPHERE-IRDIS field of view.

Considered SPHERE-IRDIS observations were obtained in *H2–H3* (i.e. $\lambda_{H2} = 1.593\,\mu m$ and $\lambda_{H3} = 1.667\,\mu m$) or in *K1–K2* (i.e. $\lambda_{K1} = 2.110\,\mu m$ and $\lambda_{K2} = 2.251\,\mu m$) dual spectral bands. The main parameters of each observation are summarized in Table 3. The diversity in the experienced observing conditions is quite representative of the SPHERE observations.

The raw observations were pre-reduced[10] with the DRH pipeline (Pavlov et al. 2008) of the SPHERE instrument, which performs thermal background subtraction, flat-field correction, anamorphism correction, compensation for spectral transmission, flux normalization, bad pixels identification and interpolation, frame centring, true-North alignment, wavelength calibration, astrometric calibration, and frame selection. These operations are complemented by custom routines implemented in the SPHERE data centre (Delorme et al. 2017), in particular to improve bad pixels correction. Finally, the SPHERE data centre combines the pre-reduced observations and delivers the calibrated ADI data sets we consider in this work. ADI reduction is performed by considering the first spectral channel (i.e. either at $\lambda_{H2} = 1.593\,\mu m$ or at $\lambda_{K1} = 2.110\,\mu m$) given that the contrast is significantly more favourable in this channel than in the second one.

To ground the performance of the proposed algorithm, we resort to massive injections of synthetic sources, as done to train the deep model of the proposed algorithm, see Section 2.2.1. Given the simplicity of the signature of the sought objects (i.e. taking the form of a blob spatially correlated over only a few pixel width), we did not find significant bias in using the same injection procedure for

[10]In this paper, we use the term *pre-reduction* to stand for the extraction, mapping, and correction of the raw data. We use the term *pre-processing* to stand for the centring and whitening of the pre-reduced observations with the statistical model detailed in Section 2.1, that serve as inputs of the supervised deep learning stage detailed in Section 2.2.

the training and evaluation steps. Injections of synthetic sources in the EIDC benchmark were performed by there authors with the VIP pipeline.

The detection performance of the proposed method are compared in Section 4.2 with the cADI, PCA, and PACO algorithms (see Introduction for their respective principle). For cADI, we have re-implemented the original method (Marois et al. 2006) based on a full-frame estimation of the off-axis PSF and of the S/N map, that is, without angular-specific processing. We have also used the refined implementation of cADI available in the VIP package (Gonzalez et al. 2017), which includes a protection angle strategy accounting for a minimal field rotation between successive images when building the off-axis PSF in order to limit the self-subtraction effect. After computation of the off-axis PSF, an S/N map is derived by accounting for an annular-based estimation of the noise in the residual images. We also applied the VIP implementation of the PCA-based algorithm combined with the same protection angle strategy and the annular-based computation of the S/N. For PCA reductions, the number of modes has been optimized in annuli by maximizing the S/N of synthetic sources with similar ranges of contrast than the ones we consider for our comparisons. The other parameters of the VIP implementation of cADI and PCA are less critical and are fixed at pre-set values (Gonzalez et al. 2017). For PACO, we performed the data reduction with our fully unsupervised processing pipeline (Flasseur et al. 2018a).

Concerning the photometry estimation, we compare in Section 4.3 the performance of the proposed method with PACO and the VIP implementation of the PCA. PACO parameters are estimated automatically in a data-driven fashion. In a nutshell, the characterization of a point-like source p is performed by a joint estimation of (i) the statistics (i.e. mean $\boldsymbol{m}_n$ and covariances $\mathbf{C}_n$) of the nuisance component $\boldsymbol{f}$, and (ii) of the photometry (i.e. flux α_p) of the given source, see equation (1). For the PCA, we resort to a similar procedure than for the detection step to set the parameters. In particular, the number of modes is optimized for each injected source to be characterized by maximizing its S/N. Once the setting fixed, the flux of a given source is estimated by minimizing the residuals through the injection of negative fake companions (Wertz et al. 2017). This is performed with a two steps procedure, as recommended in the VIP package: (i) a first guess estimate is obtained by performing a grid search, and (ii) a local optimization is performed with a Nelder–Mead simplex algorithm (Nocedal & Wright 1999). Given computational constraints (largely dominated by the PCA), the exact (known) subpixel location ϕ_p of each injected source p is provided to the different algorithms (i.e. it is not optimized), and the photometry is estimated at this ground-truth position. When estimating the photometry of real sources, both the astrometry and the photometry are optimized by the different algorithms.

4.2 Detection results

4.2.1 Detection of known real sources

A first classical sanity test to evaluate the detection performance of a post-processing algorithm is to study qualitatively its ability to re-detect real known sources initially detected with different algorithms, and possibly from different data sets. We present detailed results for one data set (HD 95086, 2015 May 05, see Table 3) selected among the eleven SPHERE-IRDIS observations we consider because HD 95086 is the star having the larger number of known real sources in the SPHERE-IRDIS field of view. Results obtained for the 10 other data sets are given in Supporting Information. Figs 9 and 10 give detection maps produced with the five tested algorithms. The detection threshold is set to $\tau = 5$ for the algorithms producing an S/N map (i.e. cADI, cADI (VIP), PCA (VIP), and PACO), and to $\tau = 0.5$ for the proposed method producing a pseudo-probability map. Due to the binary pixelwise classification task, we consider for the training step of the proposed method (see Section 2.2), its detection map is almost binary (i.e. each pixel value is close either to 0 or 1) so that the setting of the threshold τ is quite flexible. Based on the analysis of the detection maps, PACO and the proposed method lead to the best qualitative results since there are the only algorithms able to detect all real known sources without any false alarm in most of the field of view. With the proposed method, only two false alarms occur very near the borders of the field of view due to the limited number (much lower than T) of temporal samples available in this area to build a consistent model of the nuisance. This claim is also supported by the PACO detection map that displays a few false alarms in the same area.

Fig. 11 gives a more quantitative analysis of the previous results through ROCs representing the TPR as a function of the FDR (see Section 2.2.3 and equation 13) for the same data set of HD 95086 (2015 May 05). This type of representation gives a comparison of the precision-recall trade-off reached by each method, regardless the detection quantity (S/N or pseudo-probability) they produce. These curves are obtained by counting the number of TPs and FAs for the full range of possible detection thresholds, that is, $\tau \in [0; 1]$ for the proposed method, and $\tau \in \left[\min(\widehat{\boldsymbol{y}}); \max(\widehat{\boldsymbol{y}})\right]$ for cADI, PCA, and PACO. Table 4 presents averaged results over the 11 SPHERE-IRDIS data sets we consider in this study, and the detailed scores for each data set are given in Table S1 of the Supporting Information. These results illustrate the benefits of the proposed method in terms of precision-recall trade-off: the AUC under ROC is improved by at least 7 per cent with respect to the comparative algorithms.

As a final study based on the detection of known real sources, we evaluate the importance of our pre-processing step by resorting to model ablation. Removing the whitening procedure and keeping only the temporal centering in the pre-processing step does not allow to reach convergence of the network weights at training time. This is due to the high dynamics and to the high spatial non-stationarity of the residual images. We also test to account only for the pixel variances in the whitening procedure, that is, we neglect the spatial covariances so that matrices $\widehat{\mathbf{S}}_n$ in equation (2) are considered diagonal. Fig. 12 compares the precision-recall trade-off of this downgraded model to the model of the proposed approach (accounting locally for the spatial covariances). When neglecting covariances, the overall precision-recall trade-off of the detector is decreased by 8 per cent in average and the sensitivity of detection is especially lowered for low false discovery rates (which is, in practice, the most useful regime in high-contrast imaging). While several whitening methods could be used in conjunction with our method, this study confirms the importance of our custom whitening procedure accounting for the spatial covariances of the nuisance in order to reach the best performance of a detector built by supervised deep learning. This observation is also in agreement with studies performed in other works through alternative whitening procedures. For instance, the SODINN algorithm (see Introduction) that also trains a CNN to perform a detection task by supervised deep learning is sometimes prone to a large false alarm rate (Cantalloube et al. 2020). Likely, this side effect is (at least in part) due to the embedded pre-processing step that builds an empirical model of the nuisance component through PCA; an approach that does not explicitly model the spatial covariances of the nuisance, thus leading to spatially non-stationary residual images used at training time.

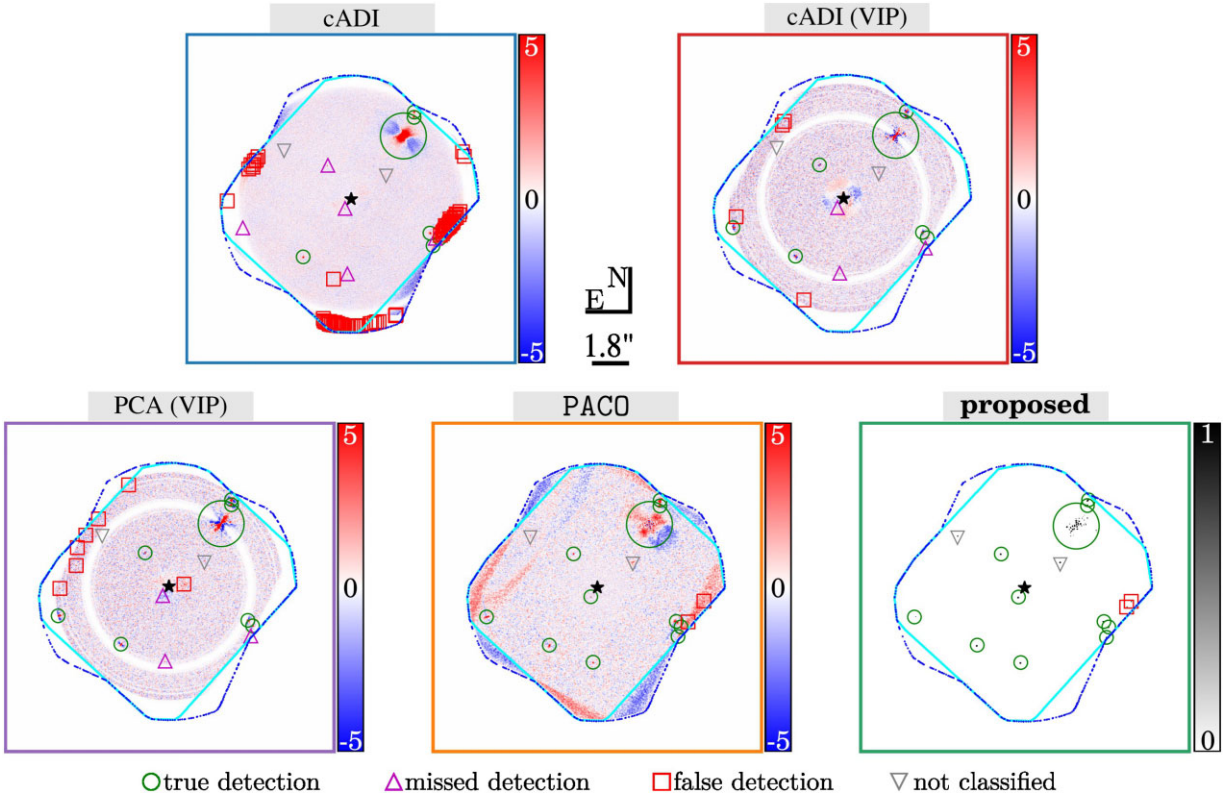

Figure 9. Detection maps obtained with the selected algorithms (see Section 4.1). Sources are classified as true, missed, and false detections. The two additional candidate point-like sources whose identification is discussed in Section 4.2.3 are not classified. The detection threshold is set to $\tau = 5$ for cADI, cADI (VIP), PCA (VIP), and PACO. It is set to $\tau = 0.5$ for the proposed algorithm. The light blue line represents the sensor field of view (encompassed withing N-pixels square support), while the dashed blue line represents the extended field of view (encompassed withing M-pixels square support) on which the detection can be performed due to the apparent rotation of the field induced by ADI. Data set: HD 95086 (2015 May 05), see Section 4 for the observation logs.

4.2.2 Detection of synthetic sources

In this section, we evaluate quantitatively the detection performance of the proposed method with ROCs and contrast curves built from massively injected synthetic sources.

As a first step, we apply the proposed detection algorithm on the three SPHERE-IRDIS data sets from the public EIDC data challenge (Cantalloube et al. 2020). The resulting detection maps are given in Fig. 13. Using the same procedure than Cantalloube et al. (2020) to benchmark 22 post-processing algorithms (including cADI, PCA, and PACO), we compute the F1R score at the set detection threshold ($\tau = 0.5$), and ROCs for the TPR and the FDR scores, as defined in equations (13), by varying the detection threshold. The AUC is then computed from ROCs. Table 5 summarizes the results we obtained with the proposed algorithm. As a purpose of comparison, we also report the PACO results that have been published in Cantalloube et al. (2020). It emphasizes the interesting precision recall of the proposed approach that performs, on these data sets, on par with or better than the 22 post-processing algorithms considered in the EIDC data challenge. However, these results should be taken with caution since they are based on a few data sets, with only six injected sources at relatively bright levels of contrast. Besides, several algorithms (including PACO) are also able to detect the injected sources without any false alarm in the field of view for a sufficiently large detection threshold. At this stage, the better performance of deep PACO in terms of false alarms rejection (quantified by the AUC_{FDR} metric) are mostly explained by the fact that the proposed algorithm produces detection maps with almost binary values. It can be noted that this effect is also encountered for all the algorithms of the EIDC data challenge producing the same type of outputs like RSM (Dahlqvist, Cantalloube & Absil 2020) or SODINN (Gonzalez, Absil & Van Droogenbroeck 2018).

To ground in details the precision-recall trade-off of the proposed detection algorithm, it is necessary to build consistent ROCs and contrast curves by resorting to massive injections of synthetic sources – unknown at training time but that we aim to detect at inference time – for various levels of contrast. For that purpose, the most realistic procedure, hereafter called *reference procedure*, consists in (i) splitting the whole set of synthetic sources in small subsets, (ii) injecting synthetic sources of one subset in the data set of interest so that injected fake sources mimic the behaviour of real (possibly unknown) sources, (iii) training the detection model with the procedure described in Section 2, and (iv) applying the trained model to the data set containing the synthetic sources injected in step (ii). Steps (ii)–(iv) are repeated for all subsets of synthetic sources. This procedure simulates the real situation when we face a new data set with real unknown exoplanets that we aim to detect at inference time. Due to the computational burden of step (iii), repeating this full procedure for all subsets of synthetic sources that we aim to detect at inference time is not realistic to build ROCs and contrast curves. To circumvent this issue, we resort to a *proxy procedure*: instead of training a different model for each subset of injected sources, we train a unique model without synthetic sources mimicking the

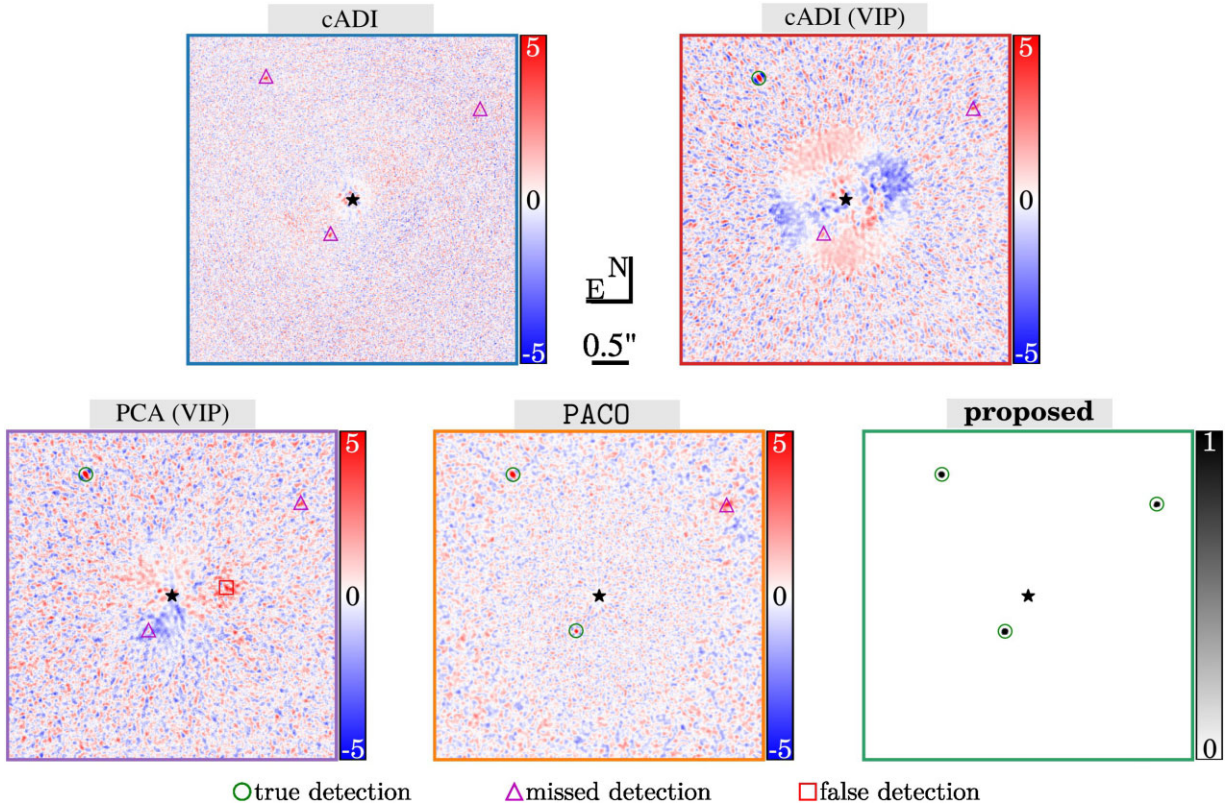

Figure 10. Same caption than Fig. 9. Zoom near the host star. Data set: HD 95086 (2015 May 05), see Section 4 for the observation logs.

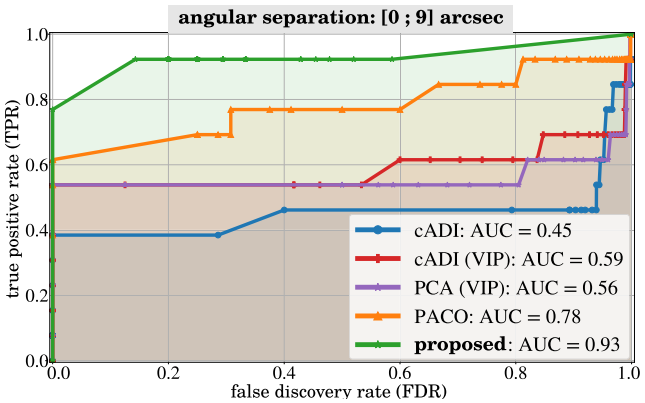

Figure 11. ROCs showing the TPR as a function of the FDR for real sources. Data set: HD 95 86 (2015 May 05), see Section 4 for the observation logs.

Table 4. Mean results of AUC for ROCs giving the TPR as a function of the FDR. The scores are averaged over the 11 SPHERE-IRDIS data sets considered in this paper, see Section 4 for the observation logs. Only the 59 known real sources present in these data sets were considered. Fig. 7 of Flasseur et al. (2022) display the corresponding ROC from which these mean results were aggregated. The best score is emphasized in bold font.

Sep. (arcsec)	cADI	cADI (VIP)	PCA (VIP)	PACO	Proposed
[0; 9]	0.38	0.72	0.66	0.88	**0.95**

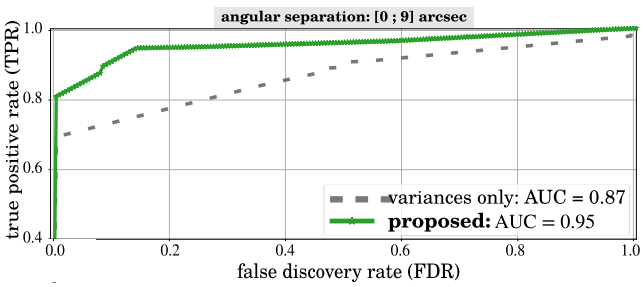

Figure 12. Ablation study on the influence of the whitening procedure of the pre-pre-processing step. The ROCs show the TPR as a function of the FDR built from the 11 SPHERE-IRDIS data sets we consider in this work (containing 59 known real sources), see Section 4 for the observation logs.

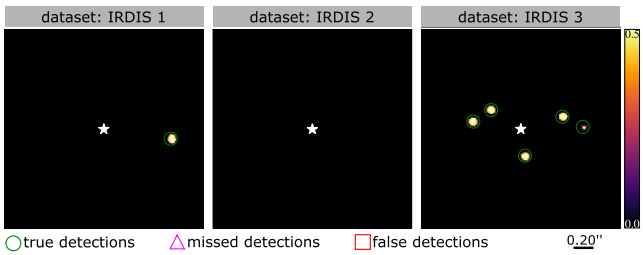

Figure 13. Detection maps obtained with the proposed algorithm on the three SPHERE-IRDIS data sets from the EIDC data challenge, see Section 4 for the observation logs. Sources are classified as true, missed, and false detections. The detection threshold is set to $\tau = 0.5$.

Table 5. Detection scores: F1R at detection threshold $\tau = 0.5$ (the higher, the better), AUC under the ROC representing the TPR as a function of τ (the higher, the better), and AUC under the ROC representing the FDR as a function of τ (the lower, the better). Scores reported for PACO are extracted from the EIDC data challenge (Cantalloube et al. 2020), and deep PACO scores are computed with a similar procedure for the three SPHERE-IRDIS data sets from the EIDC.

	IRDIS 1	IRDIS 2	IRDIS 3	Mean	Rank	IRDIS 1	IRDIS 2	IRDIS 3	Mean	Rank
			PACO					deep PACO		
F1R	1.00	[a]	1.00	1.00	1st/22 (on par)	1.00	[a]	1.00	1.00	1st/23 (on par)
AUC_{TPR}	1.00	[a]	0.93	0.97	1st/22 (on par)	1.00	[a]	0.96	0.98	1st/23
AUC_{FDR}	0.39	[a]	0.32	0.36	6/22	0.01	[a]	0.01	0.01	1st/23

Notes. [a]Metrics can not be computed since there is no injected source in this data set.

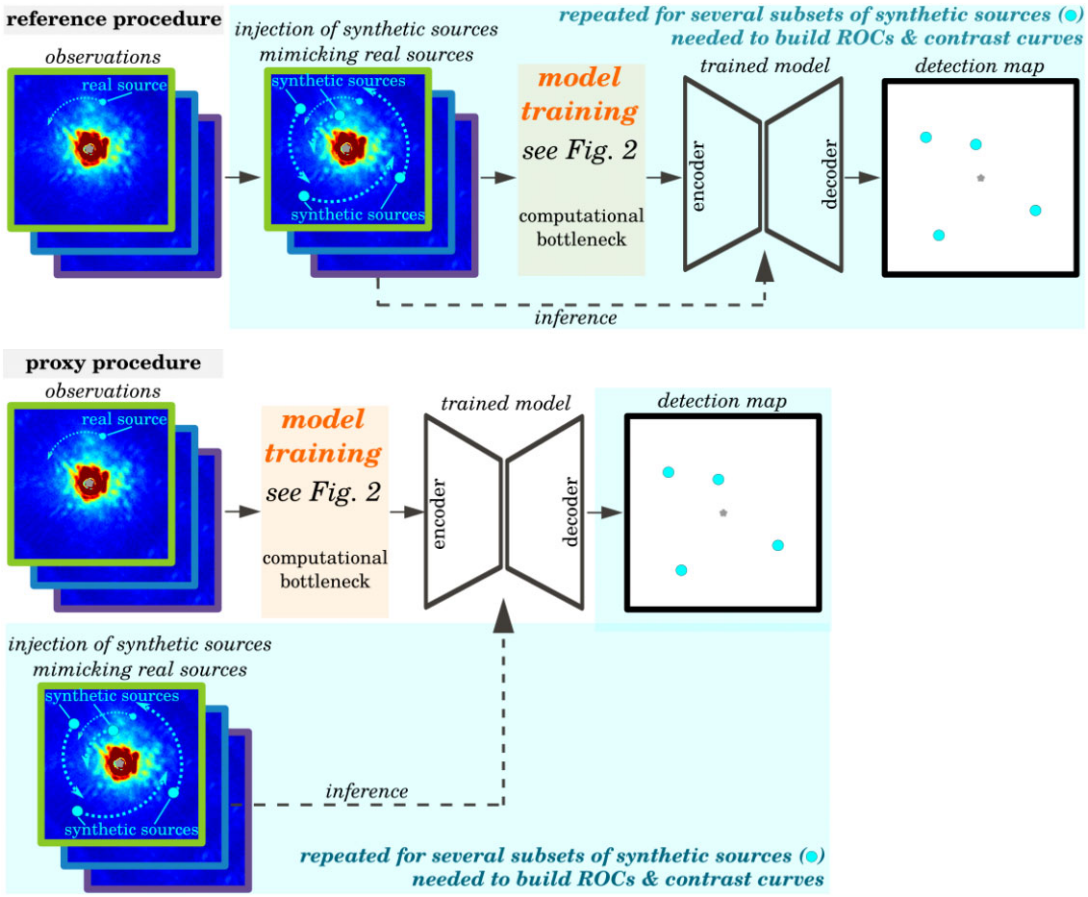

Figure 14. Schematic representation of the reference and proxy procedures used to evaluate the performance of the proposed approach through massive injections of synthetic sources mimicking the behaviour (i.e. same apparent motion) of real sources.

behaviour of real (possibly unknown) sources. Synthetic sources are injected a posteriori, that is, after completing the training step (iii), and the trained model is applied on the resulting data set. Fig. 14 illustrates the principle of the reference and proxy procedures. The proxy procedure leads to an improvement in terms of algorithmic complexity by a factor equal to the number of subsets (typically 2000, with in average five synthetic sources in each subset), that is determinant to be used in practice. At this stage, we still need to show that the proxy procedure we defined leads to reliable results so that our comparisons with state-of-the-art algorithms are fair. In particular, we aim to show that the model resulting from the proxy procedure is not prone to overfitting induced by an imperfect separation between training and testing data. In the proxy procedure, overfitting could possibly occur if the model partially memorizes the nuisance component that is seen by the network without the injected sources we aim to detect at inference time. Fig. 15 shows examples of detection maps obtained with the procedure of reference described previously and its proxy version on the three data sets of HD 95086 (2015 May 05, 2018 February 23, and 2021 March 11) considered in this work. In these experiments, we considered more than 700 synthetic sources spread over the whole field of view. Table 6 compares quantitatively the two approaches in terms of detection performance. These results show that the proxy procedure leads to reliable and conservative estimations of the overall performance of `deep PACO`. In addition, since synthetic sources mimicking the behaviour of real unknown sources are recovered with comparable rates between the two procedures, it emphasizes that the custom data-augmentation strategy of the training step, including a random permutation of the images of the data series, is efficient to circumvent the absence of ground-truth about real sources. In the following, we safely use the

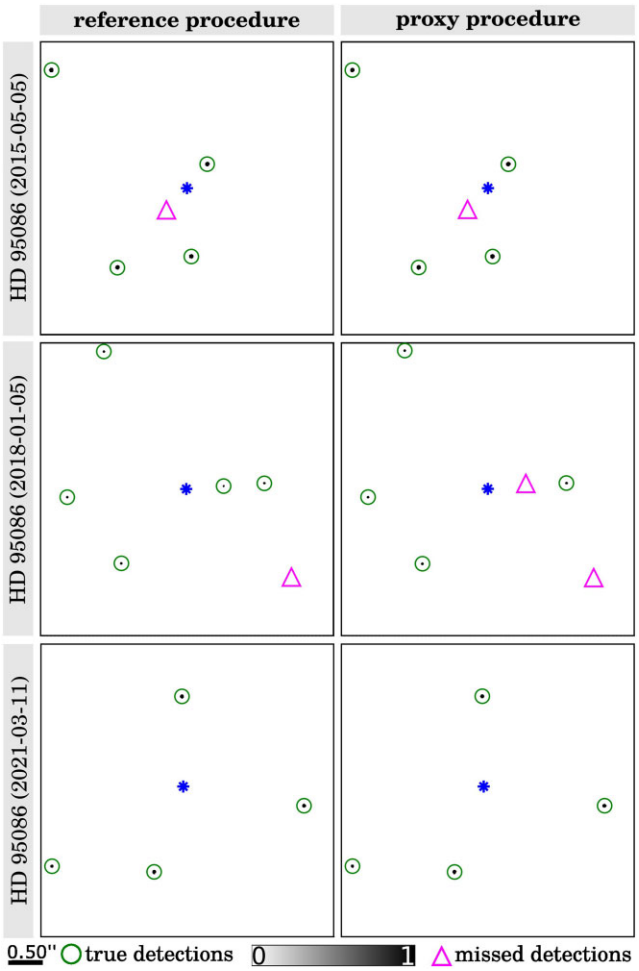

Figure 15. Comparison between the reference procedure and its proxy version for evaluation of the detection performance of the proposed algorithm. Synthetic sources are classified as true or missed detections using a detection threshold at $\tau = 0.5$. Examples of detection maps obtained in the presence of injected synthetic sources are given on three data sets of HD 95086 (2015 May 05, 2018 February 23, and 2021 March 11), see Section 4 for the observation logs.

Table 6. Comparison between the reference procedure and its proxy version for evaluation of the detection performance of the proposed algorithm. Synthetic sources are classified as true, missed, or false detections using a detection threshold at $\tau = 0.5$. Mean detection results are averaged for the three data sets of HD 95086 (2015-05-05, 2018-02-23, and 2021-03-11), see Section 4 for the observation logs.

	Reference procedure	Proxy procedure
True detections	583/728 (80.0 per cent)	564/728 (77.5 per cent)
Missed detections	145/728 (20.0 per cent)	162/728 (22.5 per cent)
False detections	0	0

proxy procedure to compare the detection capability of the proposed `deep PACO` algorithm with other post-processing methods.

Following the previously defined proxy procedure, we present detailed results obtained on HIP 88399 (2018 April 11), which is a SPHERE-IRDIS data set representative of the mean results we obtained over the eleven ones we consider in this work. Results for the 10 other data sets are reported in the Supporting Information. Fig. 16 shows detection results on a sample of 10 000 synthetic sources

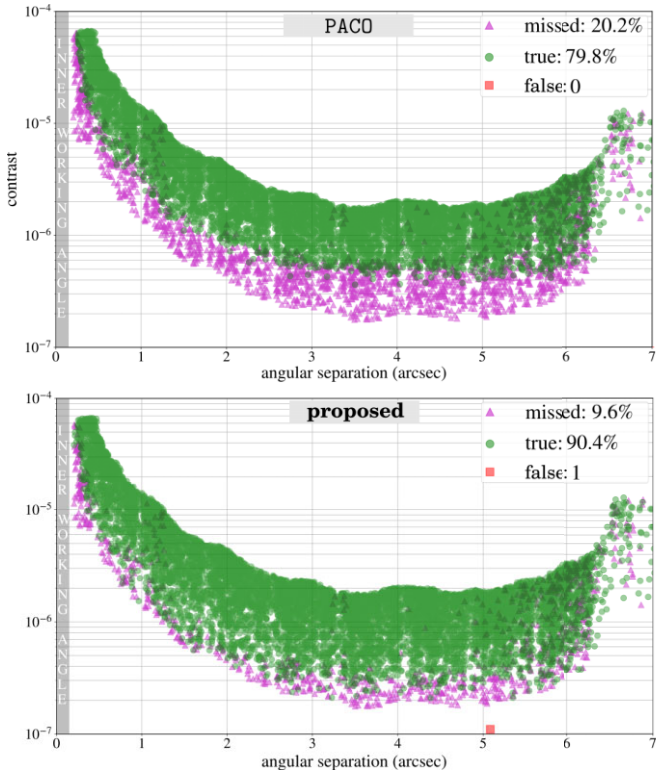

Figure 16. Detection results for 10 000 synthetic sources in a diagram plotting contrast *versus* angular separation. Each synthetic source is classified as missed, true or false detection. The angular separation of false alarms is reported on the x-axis, and they are not associated to contrast value since they do not correspond to injected synthetic sources. Data set: HIP 88399 (2018 April 11), see Section 4 for the observation logs.

in a diagram contrast *versus* angular separation for `PACO` and the proposed method. Each synthetic source is classified as missed, true, or false detection using the detection thresholds defined in Section 4.2.1. For `PACO`, setting the detection threshold at $\tau = 5$ corresponds, in average, to a realistic control (Flasseur et al. 2018a, b, c, 2020a) of the probability of false alarms (PFA) at 5σ (i.e. PFA $\simeq 3 \times 10^{-7}$). While the PFA should theoretically be controlled by the other algorithms producing an S/N map (cADI and PCA), we have shown in previous works (Flasseur et al. 2018a, b, c, 2020a) that the contrast curves are overoptimistic for these algorithms (i.e. there are significantly more false alarms than expected) due to a mismodelling of the nuisance component. This claim is also supported by the detection maps given in Figs S1–S20 of the Supporting Information for which the number of experienced false alarms is significantly higher than expected at S/N = 5. For the proposed method, converting pseudo-probabilities into S/N scores is not feasible given that the pseudo-probabilities are very close either to 0 or 1 due to the underlying binary pixelwise classification task considered at training time. For this reason, we can only check empirically that the targeted false alarm rate at 5σ is satisfied. To do so, we capitalize on our experiments with synthetic sources by counting the number of false alarms, that is, detection blobs above the threshold $\tau = 0.5$, that do not correspond to the location of an injected synthetic source. The number of false alarms is then converted into an empirical PFA by dividing the number of counts by the total number of possible detection blobs (each with a radius of one resolution element, i.e. four pixels radius for SPHERE-IRDIS observations) within the detection maps. By applying this procedure on several dozens of detection

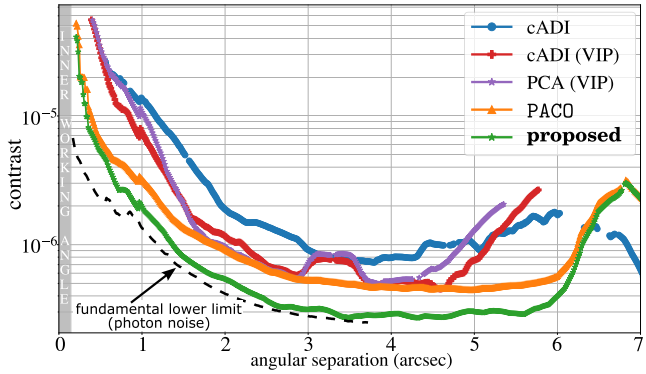

Figure 17. Contrast as a function of the angular separation. These results are based on the classification between true and missed detections of massively injected synthetic sources at various levels of contrast, see Fig. 16. It can be noted that the contrast curves of cADI and PCA are overoptimistic (see the text), since the targeted PFA is not reached. For that reason, the angular separations leading to a PFA locally higher than ten times the targeted PFA at 5σ are not reported. The black dashed line represents the ultimate detection limit driven by the photon noise. Data set: HIP 88 399 (2018 April 11), see Section 4 for the observation logs.

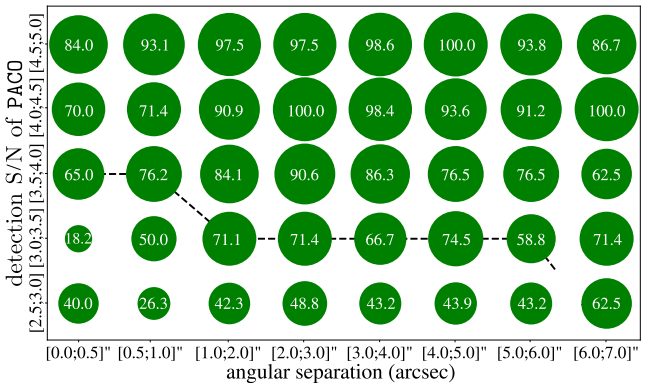

Figure 18. TPR (in per cent) of deep PACO for synthetic sources missed by PACO (i.e. for S/N $\leq$ 5) as a function of the angular separation (on the x-axis) and of the PACO's S/N of detection (on the y-axis). The black dashed line represents the equivalent PACO's detection threshold to reach TPR = 50 *per cent* with deep PACO (we recall that the classical detection threshold at 5σ with PACO also corresponds to TPR = 50 *per cent*). Data set: HIP 88399 (2018 April 11), see Section 4 for the observation logs.

maps obtained with the proposed algorithm, we experienced in average a lower or equal empirical PFA than statistically expected at 5σ. As a conclusion of this study, the contrast estimates we will derive for the proposed approach can be fairly compared with the PACO results since they correspond to similar probabilities of false alarms and of detections. Fig. 16 illustrates the capability of the proposed method to detect fainter sources than PACO. From the large number of synthetic sources classified as true, missed, and false detections in Fig. 16, we now derive the contrast curve of each algorithm. Concerning the proposed approach, given that we showed empirically that the PFAs is controlled at 5σ, it simply remains to compute the contrast level for which and equal amount of true and of missed detections is experienced. This procedure is repeated for the full range of angular separations with a sliding window of 0.05 arcsec wide. The same procedure is applied for the other algorithms. Fig. 17 summarizes the resulting contrast curves obtained with the five considered algorithms. The proposed method achieves the best detection sensitivity with an improvement in contrast up to a factor 4 with respect to the PACO algorithm. We also compare the detection sensitivity of deep PACO with the fundamental detection limit driven by the photon noise. The procedure to compute the photon noise limit is based on a careful evaluation of the contribution of the different sources of noise (i.e. photon, thermal background, and detector readout noise) combined with a statistical evaluation of the underlying S/N. This procedure will be described in details in a paper currently in preparation. We observe that deep PACO can reach for some data sets (see e.g. HIP 88 399 (2018 April 11) in Fig. 17, and HIP 88399 (2015 May 10) as well as HIP 88399 (2016 April 16) in the Supporting Information) at large separations the best achievable detection limit driven by the photon noise, which corresponds to an optimal unmixing between the signal of the sources of interest and the nuisance component. Near the star, an important gap remains (by a factor 5–10) between the actual performance and the theoretical lower limit, which a sign of a lack of angular diversity in this area. In practice, the gap between the actual contrast and the lower bound limit depends on several characteristics such as the quality and the stability of the observing conditions, the total amount of parallactic rotation, the number T of temporal frames, etc., as illustrated in Figs S21 and S22 of the Supporting Information. Reducing this gap at close separation could be addressed by investigating ways to perform the training step of our model from various data sets for which the presence of sources at similar locations is very unlikely. This adaptation requires specific developments that are left for future work.

Fig. 18 focuses on the comparison between PACO and the proposed deep PACO algorithm. It represents the TPR of deep PACO for synthetic sources missed by PACO (i.e. for S/N $\leq$ 5) as a function of the angular separation and of the PACO's S/N of detection. For instance, on this data set about 65 per cent of sources below 0.5 arcsec with an S/N of detection between 3.5 and 4.0 with PACO are detected with deep PACO (i.e. above the detection threshold $\tau = 0.5$). Similarly, more than 86 per cent of sources between 2.0 and 4.0 arcsec with an S/N of detection between 3.5 and 4.0 with PACO are detected with deep PACO (above the detection threshold $\tau = 0.5$). For these two examples, achieving the same TPR with PACO would require to decrease the detection threshold at the price to an increase of the FPR up to a mean factor of 800. Figs S23 and S24 of the Supporting Information give similar type of representation than Fig. 18 for the 10 other SPHERE-IRDIS data sets analysed in this work.

Fig. 19 gives ROCs representing the TPR as a function of the FDR for the HIP 88399 (2018 April 11) data set. There results are obtained with the 10 000 synthetic sources considered for results presented in Fig. 16. The results are split in four different angular separation ranges: [0; 2], [2; 4], [4; 6], and [6; 7] arcsec. Table 7 complements this study by presenting averaged results over the 11 SPHERE-IRDIS data sets we consider in this study, and the detailed scores for each data set are given in Table S2 of the Supporting Information. These results illustrate again the benefits of the proposed method in terms of precision-recall trade-off: the AUC under ROC is improved by 9 per cent–17 per cent with respect to the best comparative algorithm for the four angular separation ranges we consider.

4.2.3 Identification of candidate background point-like sources

In this section, we take the example of a joint analysis of data sets of HD 95086 considered in this work to illustrate the ability of the proposed approach to detect candidate faint point-like sources. Fig. 20 shows the detection maps obtained with PACO and the proposed deep PACO approach on the 2015 and 2018 observations of HD 95086. It emphasizes that two candidate point-like sources

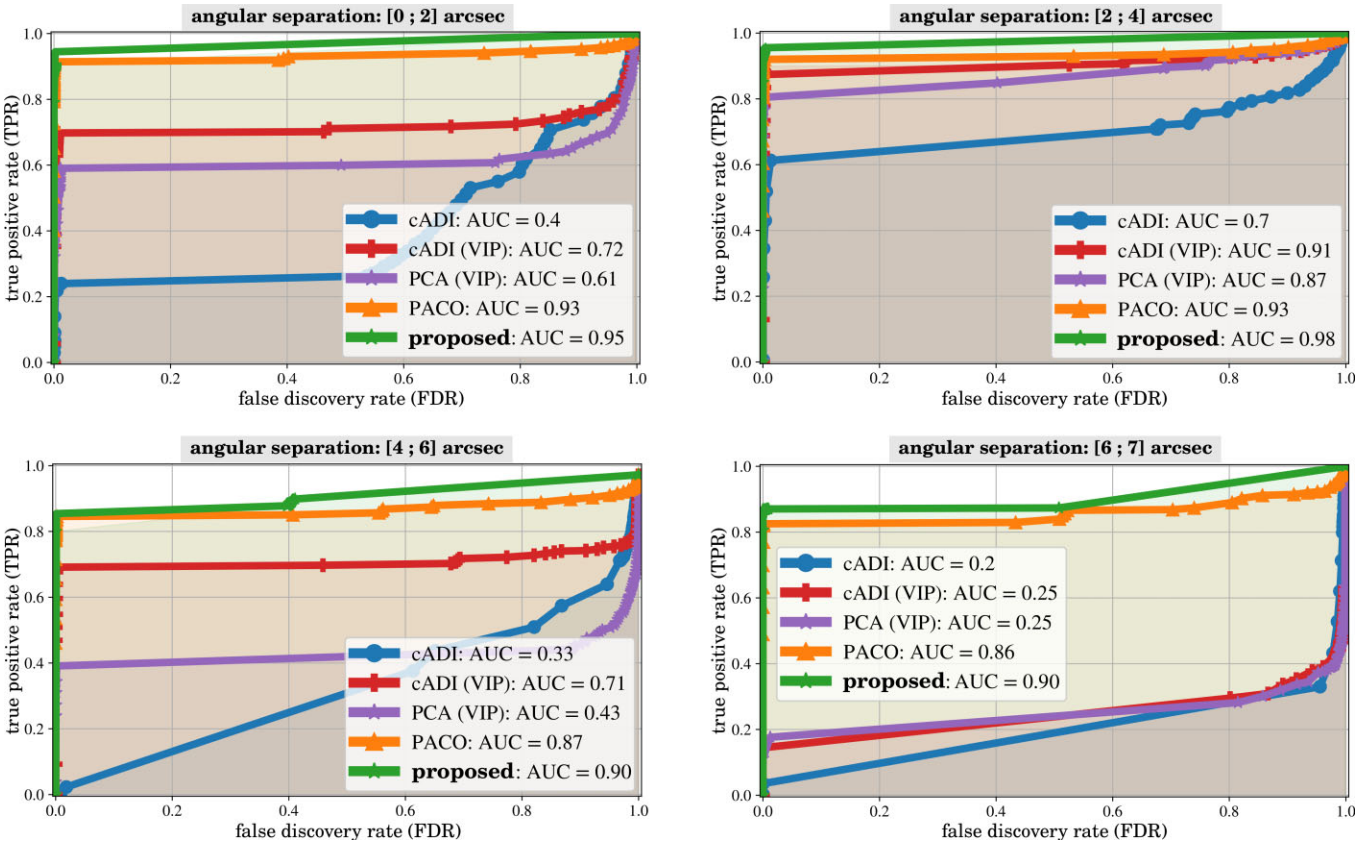

Figure 19. ROCs showing the TPR as a function of the FDR for injected synthetic sources. Data set: HIP 88399 (2018 April 11), see Section 4 for the observation logs.

Table 7. Mean results of AUC for ROCs giving the TPR as a function of the FDR. The scores are averaged over the 11 SPHERE-IRDIS data sets considered in this paper, see Section 4 for the observation logs. Only massively injected sources (10 000 per data set) were considered. Figs 8 of Flasseur et al. (2022) display the corresponding ROC from which these mean results were aggregated. Best scores are emphasized in bold font.

Sep. (arcsec)	cADI	cADI (VIP)	PCA (VIP)	PACO	Proposed
[0; 2]	0.53	0.60	0.62	0.81	**0.91**
[2; 4]	0.60	0.65	0.67	0.83	**0.92**
[4; 6]	0.11	0.59	0.55	0.71	**0.88**
[6; 7]	0.15	0.38	0.33	0.77	**0.90**

(respectively, denoted CC-11 and CC-12) are detected with `PACO` at an S/N above 5 only for one of the two epochs (in 2018, which is also one of the best SPHERE-IRDIS observations of this star) while they remain just below the classical detection threshold at 5σ on the 2015 epoch. These two point-like sources are detected with `deep PACO` on the same two epochs. Based on their astrometry, these two faint point sources are co-moving with the other background stars seen in the projected field of view of the instrument, so that they could be background stars too.[11] These candidate point-like sources are not reported in the last detailed study of the HD 95086 architecture based on the analysis of 10 SPHERE-IRDIS data sets (including the 2015 epoch but not the 2018 epoch) with classical post-processing algorithms (i.e. cADI, PCA, and TLOCI), see fig. 2 of Chauvin et al. (2018). The analysis of more recent and better data sets (including the 2018 one) in the SHINE survey (Langlois et al. 2021) of the SPHERE instrument allows to identify by visual inspection (i.e. not based on a strict measure of S/N above the classical detection threshold at 5σ) CC-11 while CC-12 has not been identified yet. The detection of the two CCs in the worst epoch data (2015 May 05) emphasizes again the benefits of the proposed `deep PACO` algorithm. This example also illustrates the complementarity of `PACO` and `deep PACO`. Even if `deep PACO` does not provide tight confidence scores (i.e. its outputs can not be directly interpreted in terms of an S/N), it allows to identify candidate point-like sources. Given the locations found by `deep PACO`, an estimation of the PFA can be derived from the S/N extracted on the `PACO` detection maps at the same locations. Using this procedure, the theoretical PFA (not including systematic sources of errors) for the co-localization of the candidate point sources CC-11 and CC-12 in the two epochs is small. None of CC-11 and CC-12 are detected in the 2021 epoch of HD 95086, as shown by Figs S3 and S4 of the Supporting Information. This observation is consistent with the fact that the achievable contrast is worse for this data set than for the 2015 and 2018 observations, likely because the observing conditions were quite average for the 2021 observations and that only half of the total integration time was spent on the target star (the 2021s epoch being recorded with the star-hopping technique, and the resulting reference data set remaining not exploited in this paper).

[11]The goal of this paper is not to study in details these CCs, but rather to present a new post-processing algorithm. These CCs should be taken with caution, and a detailed analysis is needed, in particular to exclude systematic sources of errors.

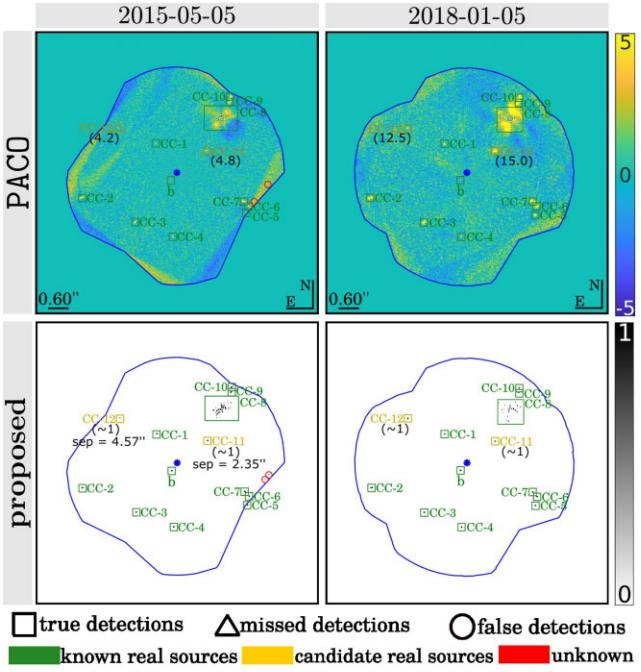

Figure 20. Detection maps obtained with PACO and deep PACO on two epochs of the HD 95086 star observed with SPHERE-IRDIS. Symbols classify sources as true, missed, and false detections based on our analysis. The detection threshold was set to $\tau = 5$ for PACO producing an S/N map, and to $\tau = 0.5$ for deep PACO producing a pseudo-probability map. Colours classify sources as known real sources, candidate (background) sources, and detections with unknown status. Data sets: HD 95086 (2015 May 05 and 2018 January 05), see Section 4 for the observation logs.

PACO and deep PACO detection maps from the 2015 and 2018 epochs also display a few blobs, circled in red, above the detection threshold that are not consistently detected on multiple epochs. These detections are located very near the borders of the (extended) field of view. They are likely artefacts due to the lack of temporal samples in this area to estimate the model parameters. This hypothesis is also supported by the fact that PACO detection maps are not stationary and does not follow a centred Gaussian distribution with unit variance in this area. deep PACO detection map from the 2021 epoch display a blob at 0.25 arcsec, that we likely identified as a false alarm since we have experienced a few false alarms in the same area during training time due to a strong stellar leakage (see Figs S3, S4, and S22b of the Supporting Information).

4.3 Characterization results

In Section 4.3.1, we ground the performance of the proposed photometry estimation module on the same data sets than the considered ones for the detection stage by resorting to massive injections of synthetic sources. For that purpose, we use the ARE metric (equation 15) averaged per angular separation. We also put in perspective the results of the detection and of the characterization stages to ground the global gain brought by the proposed algorithm. In Section 4.3.2, we propose a fast (and approximate) evaluation procedure suited when large amounts of synthetic sources are considered. Finally, in Section 4.3.3, we briefly discuss and compare photometry estimations provided by our method with estimations coming from the literature on some known sources (either exoplanets, brown dwarfs or background sources) present in the eleven SPHERE-IRDIS data sets considered in this paper.

4.3.1 *Characterization of synthetic sources*

To assess the performance of our method, we reproduce a *real* setting, hereafter called *reference procedure*, in which we estimate the flux of injected synthetic sources representative of (possibly unknown) exoplanets. To do so, we first inject synthetic test sources in the considered data set, and then extract the patches used for training the model. It is crucial to perform these steps in that order otherwise the model could have the opportunity to learn the residual nuisance component below the test sources, and a bias could be introduced.

Following this strategy, Fig. 21 shows the ARE score on the estimated photometry of synthetic sources massively injected with the reference procedure as a function of the their contrast and of their angular separation. Fig. S26 of the Supporting Information gives the same type of plots for the throughput (i.e. the ratio $\widehat{\alpha}/\alpha$ between the retrieved and ground-truth source's contrast). In both cases, results are averaged over the 11 SPHERE-IRDIS data sets of this study. Fig. 22 complements this study by aggregating the ARE score over the source contrast. These results show that the proposed algorithm leads, in average, to better characterization performance than PCA (VIP) and PACO for angular separations larger than 0.8 arcsec, with a reduction of the ARE by a factor between 1.10 and 10 with respect to PCA (VIP), and by a factor between 1.10 and 5 with respect to PACO. Closer to the star, the advantage is on average to PACO and to the proposed approach, the best of the two algorithms depending on the data set and of the source location. The fact that PACO can perform better than the proposed algorithm is due to the complexity to train a deep model, without leak between the train and the test sets, from a unique data set of interest. This effect occurs only at short angular separations since this is the region of the field of view where generating multiple non-redundant training samples is the most tricky. As illustrated in Section 4.2.2, this side effect does not occur in the detection stage of the proposed approach thanks to the included whitening procedure, which removes most of the quasi-static speckles (i.e. only residual structures non-captured by the statistical model remain). Except these residual structures that we aim to capture by deep learning, each new training set thus contain, before injection of synthetic sources, a quasi-random realization of uncorrelated Gaussian noise. This key property prevents a leak between the train and the test sets as well as the memorization of the nuisance structures by the network during its training.

4.3.2 *Efficient (and approximate) evaluation procedure*

When the massive injection of synthetic sources withing multiple data sets is needed to ground the performance of a built detector, the reference procedure described and applied in Section 4.3.1 is computationally expensive. The computational bottleneck is related to the need to train a new model for every test data set containing a dozen of synthetic test sources. As an illustration, 30 different models must be trained to get a test set of only 300 samples. In this context and in a similar fashion to the detection stage (see Section 4.2.1 and Fig. 14), we also propose a fast and approximate version of the reference procedure, that is referred as the *proxy procedure* in the following. Instead of training a different model for each subset of injected sources, we train a unique model without additional synthetic sources mimicking the behaviour of real sources. Synthetic sources are injected a posteriori, that is,

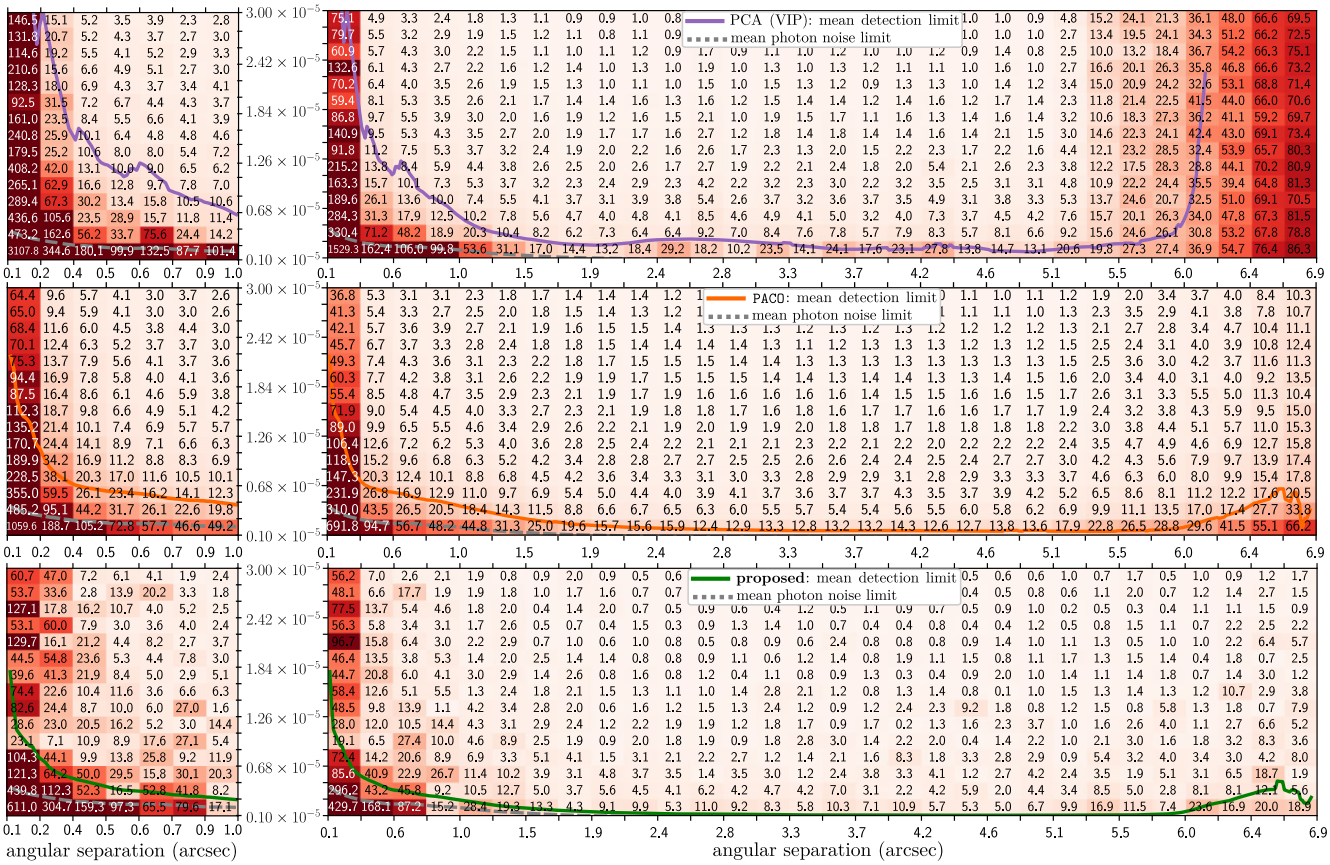

Figure 21. Mean ARE score on the estimated photometry of injected synthetic sources as a function of the their contrast and of their angular separation. From top to bottom, the panels corresponds respectively to PCA (VIP), PACO, and the proposed algorithm. For each panel, the mean detection limit (straight line) and the mean photon noise limit (dashed line) are superimposed. The results are averaged azimuthally for 40 000 sources of flux drawn uniformly between 1×10^{-6} and 3×10^{-5}. Data sets: the 11 SPHERE-IRDIS data sets considered in this work, see Section 4 for the recording logs.

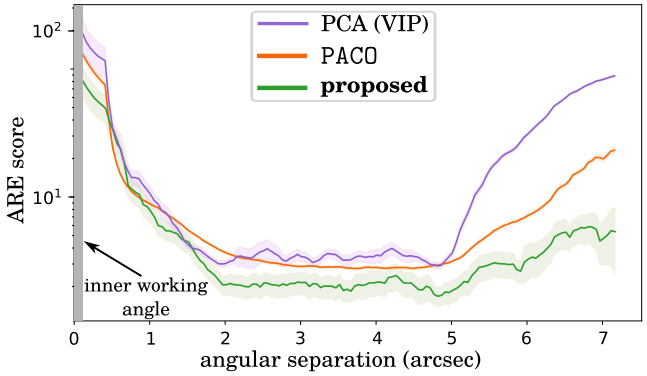

Figure 22. Mean ARE score on the estimated photometry of injected synthetic sources as a function of their angular separation. The results are averaged azimuthally for sources, of flux drawn uniformly between 1×10^{-6} and 3×10^{-5}, that have been considered in Fig. 21. Note that because of the azimuthal average, some sources belonging below the detection limit contribute to the display results, especially at short angular separations, see Fig. 21. Data sets: the 11 SPHERE-IRDIS data sets considered in this work, see Section 4 for the recording logs.

after training. Finally, the trained model is applied on the resulting data set. The gain in terms of algorithmic complexity (about a factor 2000 with in average five synthetic sources in each subset) is similar to the one brought by the proxy procedure of the detection step.

At this stage, we have to measure the ability of the proxy procedure to provide a fair estimate of the real performance that would be achieved with the reference procedure. Fig. 23(a) compares the performance of the two procedures in terms of mean ARE on the 11 SPHERE-IRDIS data sets considered in this paper. In these experiments, we considered around 3300 synthetic sources spread over the whole field of view. We observe that the proxy procedure leads to results very close to the ones provided by the reference procedure, excepted near the star where the proxy procedure is overoptimistic, that is, a positive bias is present (up to a factor 5, at worst). This bias at short angular separations can be attributed to *data leakages* between the train and the test sets in the absence of whitening procedure. Basically, as for the detection part, the model is data-dependent and the absence of whitening procedure as well as of the associated temporal shuffling of the frames induces that some parts of the nuisance component are seen and memorized by the network using the proxy procedure. This effect occurs only at short angular separations since this is the area of the field of view where training patches contain most likely some similar parts of the nuisance component. This hypothesis is also supported by the absence of bias between the proxy and the reference procedures of the detection stage. The latter encompasses a whitening procedure and a temporal shuffling of the frames that prevents data leakages between the train and test sets.

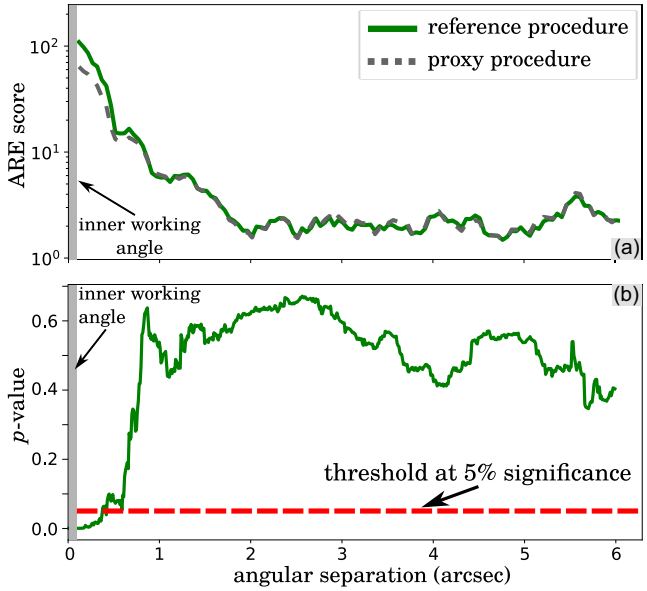

Figure 23. Comparison between the reference procedure and its proxy version for evaluation of the characterization performance of the proposed algorithm. (a) Mean ARE scores (see definition in equation 15) are reported as a function of the angular separation. (b) p-values are reported as a function of the angular separation, and can be compared with the 5 per cent significance threshold given by the dashed curve. Data sets: the 11 SPHERE-IRDIS data sets considered in this work, see Section 4 for the recording logs.

So, it remains to decide for a suited cut-off in terms of angular separation to switch from the reference to the proxy procedure. For that purpose, we resort to a binary hypothesis test. It performs a paired t-test (Kendall, Stuart & Ord 1948), where the $\mathcal{H}_0$ (null) hypothesis represents the equality between the reference and the proxy procedures while the $\mathcal{H}_1$ (alternative) hypothesis represents better results with the proxy procedure. From the results presented in Fig. 23(a), we conduct in Fig. 23(b) this statistical test on the same data sets. It displays the resulting p-values as a function of the angular separation, and compares them with a 5 per cent significance threshold. For the prescribed significance level, the two procedures can be considered as equivalent for angular separations larger than 0.5 arcsec.

As a conclusion of this study, when the number of synthetic sources for which the photometry should be evaluated at inference time constitutes a computational bottleneck, the efficient and approximate proxy procedure can be safely used in the main part of the field of view and the more expensive procedure of reference should be used near the star.

4.3.3 Characterization of real sources

In this section, we compare the photometry estimations obtained with the tested algorithms on some real known sources present in the considered data sets. When available, we also compare them with measurements published in the literature that have been obtained either with the SPHERE Data Center implementations of TLOCI and PCA or with ANDROMEDA (for HD 95086 b only). For PCA (VIP), the reported uncertainties are computed from the residual combined image (i.e. after subtraction of the on-axis PSF, derotation, and stacking) in an annulus at the same angular separation than the source of interest. For PACO, the photometry is estimated by accounting for the spatial correlations of the nuisance component. As concerns the proposed algorithm, and as for its detection stage, the control of the uncertainties is an intricate task given that the core of the deep model works as a *black-box*. For that reason, we did not produce estimations of the standard deviation, which is left for future work. For the experiments we conduct with PCA (VIP), PACO and the proposed method, we use the same pre-reduction of the raw observations and the same off-axis PSF template that has been measured prior the science observations. It is not ensured that these specificity also hold for photometry estimates extracted from the literature, which could induce (unknown) systematic variations that are not taken into account. Similarly, given the variability of the observing conditions between different observations, absolute comparisons between multiple epochs of the same target should be done with caution.

Table 8 reports the photometry estimations on known real sources. Even in the absence of absolute ground truth, we can make some relative comparisons for a given data set and for a given source. In that view, photometry estimates produced by the proposed method are compatible with published and/or obtained results with PCA (VIP) and PACO for almost all sources. The largest discrepancy is for CC5 of HIP 88399 (2018 April 11), where a factor 6 lies between the estimates from PCA (VIP) and the proposed algorithm. Very likely, the PCA (VIP) estimate is not reliable as the source is located at large angular separation, that is, in an area of the field of view where PCA (VIP) is prone to large errors (see Fig. 21). Besides, PACO and the proposed method lead to quite close estimations, which also supports the previous claim.

Photometry estimates of CC-12 that we identified from HD 95086 data sets in Section 4.2.3 are consistent among the different algorithms. They are also consistent between the two epochs where CC-12 has been detected. For CC-11 that we identified from the same data sets, we note a discrepancy by about 40 per cent between the estimates obtained (i) by the proposed method, and (ii) by the other tested algorithms. However, this two groups of estimations are consistent between the 2015 and 2018 epochs. The discrepancy could be attributed to the presence of a bright background source (denoted by CC-8 in Fig. 20) that could lead to an overestimation of the photometry with algorithms that do not rely on a training step with synthetic sources. In any case, these two candidate point-like sources should be considered with caution.

5 CONCLUSION

We have described the key principles of a new algorithm for detecting and characterizing point-like sources at high contrast from ADI observations. The detection stage combines the statistics-based model of PACO with deep learning in a three step procedure: (i) the data are centred and whitened using the PACO framework, (ii) a CNN is trained to detect synthetic sources from the pre-processed images, and (iii) a detection map is inferred. While the CNN itself works as a black-box approach, the proposed method encompasses prior domain knowledge such as the apparent motion of sources and the expected shape of the exoplanetary signal inside the ADI data sets. More importantly, the proposed detection approach capitalizes on the statistical model of the nuisance component embedded in PACO to improve the stationarity and the contrast during a pre-processing step. Once a candidate source has been identified, its photometry can be estimated using a dedicated deep learning module, also trained in a supervised fashion.

Tested on 11 SPHERE-IRDIS data sets, the proposed detection method performs better than standard algorithms of the field like PCA as well as PACO in terms of precision-recall trade-off. The detection

Table 8. Estimated photometry on some known real sources present in the considered data sets. For HD 95086, the (candidate) point-like sources are denoted as in Fig. 20. For HIP 88399 (respectively, HIP 72192), point-like sources are denoted by CC1–CC5 (respectively, CC1 and CC2) through increasing angular separations. The astrometry (angular separation and true-North angle) has been estimated with PACO.

Identification					Estimated photometry $\widehat{\boldsymbol{\alpha}}$ (contrast)			
Target	Obs. date	Source	Ang. sep. (arcsec)	Angle (°)	Literature	PCA (VIP)	PACO	Proposed
HD 95086	2015-05-05	b	0.619 ± 0.003	148.4 ± 0.2	[a] $(1.13 \pm 0.46) \times 10^{-5}$ [b] $(1.10 \pm 0.55) \times 10^{-5}$ [c] $(1.58 \pm 1.75) \times 10^{-5}$	$(1.06 \pm 0.06) \times 10^{-5}$	$(1.14 \pm 0.17) \times 10^{-5}$	1.11×10^{-5}
HD 95086	2015-05-05	CC1	2.207 ± 0.003	35.1 ± 0.1	[a] $(6.19 \pm 1.94) \times 10^{-6}$	$(6.33 \pm 0.57) \times 10^{-6}$	$(7.05 \pm 0.72) \times 10^{-6}$	6.41×10^{-6}
HD 95086	2015-05-05	CC11	2.346 ± 0.006	305.4 ± 0.2	–	$(2.06 \pm 4.61) \times 10^{-6}$	$(3.14 \pm 0.70) \times 10^{-6}$	1.80×10^{-6}
HD 95086	2015-05-05	CC12	4.546 ± 0.010	52.3 ± 0.1	–	$(2.53 \pm 1.31) \times 10^{-6}$	$(2.70 \pm 0.90) \times 10^{-6}$	1.94×10^{-6}
HD 95086	2018-01-05	b	0.617 ± 0.002	145.7 ± 0.2	[b] $(1.20 \pm 0.32) \times 10^{-5}$ [c] $(1.32 \pm 0.35) \times 10^{-5}$	$(1.23 \pm 0.23) \times 10^{-5}$	$(1.20 \pm 0.06) \times 10^{-5}$	1.18×10^{-5}
HD 95086	2018-01-05	CC1	2.241 ± 0.002	37.4 ± 0.1	–	$(6.25 \pm 1.88) \times 10^{-6}$	$(6.59 \pm 0.37) \times 10^{-6}$	6.39×10^{-6}
HD 95086	2018-01-05	CC11	2.262 ± 0.004	305.9 ± 0.1	–	$(2.56 \pm 0.28) \times 10^{-6}$	$(2.75 \pm 0.35) \times 10^{-6}$	1.90×10^{-6}
HD 95086	2018-01-05	CC12	4.606 ± 0.006	53.0 ± 0.1	–	$(2.73 \pm 3.00) \times 10^{-6}$	$(2.39 \pm 0.40) \times 10^{-6}$	2.40×10^{-6}
HD 95086	2021-03-11	b	0.630 ± 0.003	142.4 ± 0.2	–	$(1.21 \pm 0.23) \times 10^{-5}$	$(1.25 \pm 0.18) \times 10^{-5}$	1.20×10^{-5}
HD 95086	2021-03-11	CC1	2.293 ± 0.005	40.8 ± 0.1	–	$(5.66 \pm 0.73) \times 10^{-6}$	$(6.30 \pm 1.56) \times 10^{-6}$	5.48×10^{-6}
HD 95086	2021-03-11	CC11	S/N too low	S/N too low	S/N too low	S/N too low	S/N too low	S/N too low
HD 95086	2021-03-11	CC12	S/N too low	S/N too low	S/N too low	S/N too low	S/N too low	S/N too low
HIP 88399	2015-05-10	CC1	1.302 ± 0.002	210.4 ± 0.1	–	$(7.09 \pm 0.72) \times 10^{-6}$	$(7.39 \pm 0.66) \times 10^{-6}$	6.67×10^{-6}
HIP 88399	2015-05-10	CC2	3.488 ± 0.006	142.4 ± 0.1	[a] $(7.94 \pm 2.10) \times 10^{-6}$	$(7.10 \pm 0.36) \times 10^{-6}$	$(7.56 \pm 0.27) \times 10^{-6}$	7.61×10^{-6}
HIP 88399	2015-05-10	CC3	4.281 ± 0.006	227.7 ± 0.1	–	$(2.36 \pm 0.17) \times 10^{-7}$	$(3.42 \pm 1.03) \times 10^{-7}$	2.52×10^{-7}
HIP 88399	2015-05-10	CC4	5.208 ± 0.003	322.1 ± 0.1	[a] $(5.11 \pm 6.00) \times 10^{-7}$	$(5.57 \pm 5.06) \times 10^{-7}$	$(4.89 \pm 0.95) \times 10^{-7}$	4.94×10^{-7}
HIP 88399	2015-05-10	CC5	S/N too low	S/N too low	S/N too low	S/N too low	S/N too low	S/N too low
HIP 88399	2016-04-16	CC1	1.237 ± 0.003	212.8 ± 0.1	[a] $(9.38 \pm 6.99) \times 10^{-6}$	$(1.04 \pm 0.13) \times 10^{-5}$	$(1.22 \pm 0.18) \times 10^{-5}$	1.11×10^{-5}
HIP 88399	2016-04-16	CC2	3.422 ± 0.001	141.7 ± 0.1	[a] $(1.32 \pm 2.97) \times 10^{-5}$	$(1.47 \pm 0.05) \times 10^{-5}$	$(1.64 \pm 0.05) \times 10^{-5}$	1.57×10^{-5}
HIP 88399	2016-04-16	CC3	4.240 ± 0.003	228.6 ± 0.1	[a] $(5.35 \pm 4.20) \times 10^{-7}$	$(5.10 \pm 4.65) \times 10^{-7}$	$(7.64 \pm 1.36) \times 10^{-7}$	5.78×10^{-7}
HIP 88399	2016-04-16	CC4	5.282 ± 0.003	322.6 ± 0.1	[a] $(5.11 \pm 6.00) \times 10^{-7}$	$(7.21 \pm 1.14) \times 10^{-7}$	$(8.02 \pm 1.25) \times 10^{-7}$	7.58×10^{-7}
HIP 88399	2016-04-16	CC5	6.561 ± 0.003	201.8 ± 0.1	–	$(1.10 \pm 2.19) \times 10^{-6}$	$(1.22 \pm 0.94) \times 10^{-6}$	1.18×10^{-6}
HIP 88399	2018-04-11	CC1	1.106 ± 0.001	217.9 ± 0.1	[a] $(1.61 \pm 1.27) \times 10^{-5}$	$(2.01 \pm 0.25) \times 10^{-5}$	$(2.22 \pm 0.15) \times 10^{-5}$	2.22×10^{-5}
HIP 88399	2018-04-11	CC2	3.289 ± 0.001	140.1 ± 0.1	[a] $(2.21 \pm 1.74) \times 10^{-5}$	$(2.65 \pm 2.19) \times 10^{-5}$	$(2.82 \pm 0.06) \times 10^{-5}$	2.88×10^{-5}
HIP 88399	2018-04-11	CC3	4.140 ± 0.003	230.3 ± 0.1	[a] $(8.79 \pm 6.90) \times 10^{-7}$	$(1.10 \pm 0.46) \times 10^{-6}$	$(9.70 \pm 1.31) \times 10^{-7}$	1.03×10^{-6}
HIP 88399	2018-04-11	CC4	5.411 ± 0.002	323.6 ± 0.1	[a] $(1.21 \pm 0.95) \times 10^{-6}$	$(1.31 \pm 0.25) \times 10^{-6}$	$(1.40 \pm 0.12) \times 10^{-6}$	1.38×10^{-6}
HIP 88399	2018-04-11	CC5	6.429 ± 0.002	202.5 ± 0.1	–	$(4.68 \pm 1.85) \times 10^{-7}$	$(2.83 \pm 0.33) \times 10^{-6}$	2.81×10^{-6}
HD 131399	2015-06-12	Ab	0.845 ± 0.004	195.6 ± 0.2	[d] $(7.94 \pm 2.10) \times 10^{-6}$	$(9.36 \pm 8.01) \times 10^{-6}$	$(1.05 \pm 0.17) \times 10^{-5}$	8.23×10^{-6}
HD 131399	2016-05-07	Ab	0.825 ± 0.003	194.2 ± 0.2	[d] $(7.66 \pm 2.21) \times 10^{-6}$	$(7.58 \pm 2.10) \times 10^{-6}$	$(7.92 \pm 1.37) \times 10^{-6}$	7.86×10^{-6}
HIP 65426	2017-02-09	b	0.834 ± 0.001	149.9 ± 0.1	[e] $(9.91 \pm 7.38) \times 10^{-5}$	$(6.20 \pm 0.46) \times 10^{-5}$	$(5.87 \pm 0.17) \times 10^{-5}$	6.17×10^{-5}
HIP 65426	2018-05-13	b	0.825 ± 0.001	149.9 ± 0.1	[a] $(8.47 \pm 2.24) \times 10^{-5}$ [f] $(8.39 \pm 2.22) \times 10^{-5}$	$(9.08 \pm 2.44) \times 10^{-5}$	$(8.98 \pm 0.09) \times 10^{-5}$	9.06×10^{-5}
HIP 72192	2015-06-11	CC1	4.632 ± 0.012	166.7 ± 0.1	–	$(2.94 \pm 2.31) \times 10^{-6}$	$(3.24 \pm 1.07) \times 10^{-6}$	2.88×10^{-6}
HIP 72192	2015-06-11	CC2	4.891 ± 0.006	342.8 ± 0.1	–	$(5.44 \pm 1.85) \times 10^{-6}$	$(6.07 \pm 1.10) \times 10^{-6}$	5.95×10^{-6}

Notes. Photometry estimations extracted from the literature has been obtained with TLOCI by Langlois et al. (2021) for [a], TLOCI by Desgrange et al. (2022) for [b], ANDROMEDA by Desgrange et al. (2022) for [c], PCA by Chauvin et al. (2017) for [e], and TLOCI by Cheetham et al. (2019) for [f]. Reported uncertainties correspond to three times the estimated standard deviation.

sensitivity is improved by a factor between 2 and 5 in average with respect to PACO for the whole range of angular separations. This gain in even more important compared to baseline methods as cADI or PCA. Interestingly, we showed experimentally that the proposed approach is able to (mostly) reach the fundamental detection sensitivity driven by the photon noise limit for angular separations above 0.5 arcsec. This corresponds to an optimal extraction of the sought signals in that regime of separations. The gain brought by the characterization stage of the proposed approach is more moderate but non-negligible. The absolute error of photometric estimation is reduced by a factor 2 with respect to PACO for sources of contrast up to 10^{-6} located above 0.5 arcsec. Nearer the star, the advantage is on average either to PACO or to the proposed approach, depending on the data set and of the source location. The gain brought by the characterization module of the proposed approach could allow to further constrain the physical properties of detected exoplanets, usually performed by fitting atmospheric models to the extracted photometry.

Based on the analysis of these results, several lessons can be drawn regarding the deployed methodology. As a major point, this work illustrates the feasibility to build a complex model (with millions of parameters) of the nuisance component through supervised deep learning from a single ADI data set of observations. The underlying detection model is learned, without overfitting, thanks to a custom data-augmentation strategy based on two key ingredients: (i) a random temporal shuffling of the individual images applied before injections of each set of synthetic training sources, and (ii) a pre-processing step encompassing a whitening procedure that removes most of the spatial correlations of the nuisance component. Prior to the injection of synthetic training sources, each training set is thus formed by a quasi-random realization of (mostly) uncorrelated Gaussian noise. This property prevents a leak between the train and the test sets as well as a potential memorization of the nuisance structures by the network during its training. Beyond that aspects, the pre-processing step allows to improve the stationarity of the data, thus preventing the high number of false alarms that can occur with

other approaches of the field based on supervised deep learning. Our results also emphasize that deriving a flux estimate is a more complex task than providing a qualitative result related to the presence or to the absence of a source with our hybrid modelling of the nuisance component. In particular, we illustrate numerically that the whitening process is detrimental for source characterization (hence, it is not applied) because it modifies both the shape and the amplitude of the exoplanetary signature. In the absence of the whitening procedure, a shallower architecture should be used to avoid overfitting.

We are currently working on the extension of the proposed algorithm for the joint processing of multispectral data sets such as the ones provided by the SPHERE-IFS instrument using the angular plus spectral differential imaging technique. Besides, we are currently investigating the three main limitations of the proposed approach: (i) the lack of control of the uncertainties, (ii) its task dependence which is not adapted to reconstruct spatially resolved objects like circumstellar discs, and (iii) the data dependence of the learning procedure which does not take benefits from multiple observations to build a more general and robust model of the nuisance component. Concerning the later point, building a model from multiple observations could be a promising step to reduce the remaining gap (by a factor 10–30) between the current detection performance and the theoretical ultimate detection sensitivity driven by the fundamental photon noise limit. Besides, we would like to incorporate within our deep models some meta-data (e.g. monitoring of the observing conditions and telemetry of the adaptive optics), with the aim to further improve their sensitivity and robustness.

ACKNOWLEDGEMENTS

We thank the anonymous referee for her/his careful reading of the manuscript as well as her/his insightful comments and suggestions.

This project is supported in part by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (COBREX; grant agreement no. 885593). The work of TB and JP was supported in part by the INRIA/NYU collaboration, the Louis Vuitton/ENS chair on artificial intelligence and the French government under management of Agence Nationale de la Recherche as part of the *Investissements d'avenir* program, reference ANR19-P3IA0001 (PRAIRIE 3IA Institute). The work of JM was supported in part by the ERC grant no. 714381 (SOLARIS project) and by ANR 3IA MIAI@Grenoble-Alpes (ANR-19-P3IA-0003).

This work was granted access to the HPC resources of IDRIS under the allocation 2022-AD011013643 made by GENCI.

OF, TB, JM, JP, ML, and AML conceived and designed the method as well as the analysis presented in this paper. OF and TB developed, tested, and implemented the algorithm. OF and ML selected the raw data. ML pre-reduced them through the SPHERE Data Centre. OF and TB performed the analysis of the data. OF, TB, JM, JP, ML, and AML wrote the manuscript.

DATA AVAILABILITY

The raw data used in this article are freely available on the ESO archive facility at http://archive.eso.org/eso/eso_archive_main.html. They were pre-reduced with the SPHERE Data Centre, jointly operated by OSUG/IPAG (Grenoble), PYTHEAS/LAM/CESAM (Marseille), OCA/Lagrange (Nice), Observatoire de Paris/LESIA (Paris), and Observatoire de Lyon/CRAL (Lyon, France). The resulting pre-processed data sets will be shared based on reasonable request to the corresponding author.

REFERENCES

Allard F., Guillot T., Ludwig H.-G., Hauschildt P. H., Schweitzer A., Alexander D. R., Ferguson J. W., 2003, in IAU Symp. p. 325
Allard F., Allard N. F., Homeier D., Kielkopf J., McCaughrean M. J., Spiegelman F., 2007, A&A, 474, L21
Amara A., Quanz S. P., 2012, MNRAS, 427, 948
Badrinarayanan V., Kendall A., Cipolla R., 2017, IEEE Trans. Pattern Anal. Mach. Intell., 39, 2481
Beuzit J.-L. et al., 2019, A&A, 631, A155
Boucaud A. et al., 2020, MNRAS, 491, 2481
Bowler B. P., 2016, PASP, 128, 102001
Burrows A. et al., 1997, ApJ, 491, 856
Cabayol L. et al., 2021, MNRAS, 506, 4048
Cantalloube F. et al., 2015, A&A, 582, A89
Cantalloube F. et al., 2020, in Adaptive Optics Systems VII. p. 1027
Carbillet M. et al., 2011, Exp. Astron., 30, 39
Castellá B. F. et al., 2016, in Adaptive Optics Systems V. p. 697
Chabrier G., Baraffe I., Allard F., Hauschildt P., 2000, ApJ, 542, 464
Chauvin G., Lagrange A.-M., Dumas C., Zuckerman B., Mouillet D., Song I., Beuzit J.-L., Lowrance P., 2004, A&A, 425, L29
Chauvin G. et al., 2005, A&A, 438, L29
Chauvin G. et al., 2017, A&A, 605, L9
Chauvin G. et al., 2018, A&A, 617, A76
Cheetham A. et al., 2019, A&A, 622, A80
Chen Y., Wiesel A., Eldar Y. C., Hero A. O., 2010, IEEE Trans. Signal Process., 58, 5016
Chomez A. et al., 2023, A&A, 675, A205
Conte E., Lops M., Ricci G., 1995, IEEE Trans. Aerosp. Electr. Syst., 31, 617
Currie T., Fukagawa M., Thalmann C., Matsumura S., Plavchan P., 2012a, ApJ, 755, L34
Currie T. et al., 2012b, ApJ, 760, L32
Daglayan H., Vary S., Cantalloube F., Absil P.-A., Absil O., 2022, preprint (arXiv:2210.10609)
Dahlqvist C.-H., Cantalloube F., Absil O., 2020, A&A, 633, A95
Dahlqvist C.-H., Louppe G., Absil O., 2021a, A&A, 646, A49
Dahlqvist C.-H., Cantalloube F., Absil O., 2021b, A&A, 656, A54
Delorme P. et al., 2017, in Annual meeting of the French Society of Astronomy and Astrophysics.
Desgrange C. et al., 2022, A&A, 664, A139
Desidera S. et al., 2021, A&A, 651, A70
Dohlen K., Saisse M., Origne A., Moreaux G., Fabron C., Zamkotsian F., Lanzoni P., Lemarquis F., 2008, in SPIE Astronomical Telescopes + Instrumentation. p. 701859
Fergus R., Hogg D. W., Oppenheimer R., Brenner D., Pueyo L., 2014, ApJ, 794, 161
Flasseur O., Denis L., Thiébaut É., Langlois M., 2018a, in IEEE International Conference on Image Processing. p. 2735
Flasseur O., Denis L., Thiébaut É., Langlois M., 2018b, A&A, 618, A138
Flasseur O., Denis L., Thiébaut É. M., Langlois M., 2018c, in SPIE Astronomical Telescopes + Instrumentation. p. 107032R
Flasseur O., Denis L., Thiébaut É., Langlois M., 2020a, A&A, 634, A2
Flasseur O., Denis L., Thiébaut É., Langlois M., 2020b, A&A, 637, A9
Flasseur O., Thé S., Denis L., Thiébaut É., Langlois M., 2021, A&A, 651, A62
Flasseur O., Bodrito T., Mairal J., Ponce J., Langlois M., Lagrange A.-M., 2022, in Adaptive Optics Systems VIII. p. 1154
Galicher R. et al., 2018, A&A, 615, A92
Gawlikowski J. et al., 2023, Artif. Intell. Rev., 56, 1513
Gebhard T. D., Bonse M. J., Quanz S. P., Schölkopf B., 2022, A&A, 666, A9
Gonzalez C. G., Absil O., Absil P.-A., Van Droogenbroeck M., Mawet D., Surdej J., 2016, A&A, 589, A54
Gonzalez C. A. G. et al., 2017, AJ, 154, 12
Gonzalez C., Absil O., Van Droogenbroeck M., 2018, A&A, 613, A71
Goodfellow I., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A., Bengio Y., 2014, Adv. Neural Inform. Process. Syst., 27
He K., Zhang X., Ren S., Sun J., 2015, in Proceedings of the IEEE International Conference on Computer Vision. p. 1026

He K., Zhang X., Ren S., Sun J., 2016, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. p. 770
Huertas-Company M., Lanusse F., 2023, Publ. Astron. Soc. Aust., 40, e001
Hüllermeier E., Waegeman W., 2021, Mach. Learn., 110, 457
Jensen-Clem R. et al., 2017, AJ, 155, 19
Jovanovic N. et al., 2015, PASP., 127, 890
Kay S. M., 1993, Fundamentals of Statistical Signal Processing: Estimation Theory. Prentice-Hall, Inc., Hoboken, NJ
Kendall M. G., Stuart A., Ord J. K., 1948, The Advanced Theory of Statistics. Vol. 1, JSTOR
Keppler M. et al., 2018, A&A, 617, A44
Kingma D. P., Ba J., 2014, preprint (arXiv:1412.6980)
Lafrenière D., Marois C., Doyon R., Nadeau D., Artigau E., 2007, ApJ, 660, 770
Lagrange A.-M. et al., 2009, A&A, 493, L21
Langlois M. et al., 2021, A&A, 651, A71
Ledoit O., Wolf M., 2004, J. Multivariate Anal., 88, 365
Macintosh B. et al., 2014, Proc. Natl. Acad. Sci., 111, 12661
Macintosh B. et al., 2015, Science, 350, 64
Marois C., Lafrenière D., Doyon R., Macintosh B., Nadeau D., 2006, ApJ, 641, 556
Marois C., Macintosh B., Barman T., Zuckerman B., Song I., Patience J., Lafrenière D., Doyon R., 2008, Science, 322, 1348
Marois C., Correia C., Véran J.-P., Currie T., 2013, Proc. Int. Astron. Union, 8, 48
Marois C., Correia C., Galicher R., Ingraham P., Macintosh B., Currie T., De Rosa R., 2014, in SPIE Astronomical Intrumentation + Telescopes. p. 91480U
Mawet D. et al., 2014, ApJ, 792, 97
Mesa D. et al., 2019, MNRAS, 488, 37
Milletari F., Navab N., Ahmadi S.-A., 2016, in 2016 Fourth International Conference on 3D Vision (3DV). p. 565
Morzinski K. M. et al., 2014, in Adaptive Optics Systems IV. p. 914804
Mugnier L. M., Cornia A., Sauvage J.-F., Rousset G., Fusco T., Védrenne N., 2009, J. Opt. Soc. Amer. A, 26, 1326
Nielsen E. L. et al., 2012, ApJ, 750, 53
Nielsen E. L. et al., 2017, AJ, 154, 218
Nielsen E. L. et al., 2019, AJ, 158, 13
Nocedal J., Wright S. J., 1999, Numerical Optimization. Springer, Berlin
Pairet B., Cantalloube F., Gomez Gonzalez C. A., Absil O., Jacques L., 2019, MNRAS, 487, 2262
Paszke A. et al., 2019, In Wallach H., Larochelle H., Beygelzimer A., d'Alché-Buc F., Fox E., Garnett R., eds, Advances in Neural Information Processing Systems.Vol. 32, Curran Associates, Inc., p. 8024
Pavlov A., Möller-Nilsson O., Feldt M., Henning T., Beuzit J.-L., Mouillet D., 2008, in SPIE Astronomical Telescopes + Instrumentation. p. 701939
Pueyo L., 2018, Handbook of Exoplanets. Springer, Berlin, p. 705
Rameau J. et al., 2013a, ApJ, 772, L15
Rameau J. et al., 2013b, ApJ, 779, L26
Reddi S. J., Kale S., Kumar S., 2019, preprint (arXiv:1904.09237)
Ronneberger O., Fischer P., Brox T., 2015, in International Conference on Medical Image Computing and Computer-assisted Intervention. p. 234
Ruffio J.-B. et al., 2017, ApJ, 842, 14
Samland M., Bouwman J., Hogg D., Brandner W., Henning T., Janson M., 2021, A&A, 646, A24
Santos N. C., 2008, New Astron. Rev., 52, 154
Schneider J., Dedieu C., Le Sidaner P., Savalle R., Zolotukhin I., 2011, A&A, 532, A79
Simonyan K., Zisserman A., 2014 preprint (arXiv preprint arXiv:1409.1556)
Soummer R., Pueyo L., Larkin J., 2012, ApJ, 755, L28
Sudre C. H., Li W., Vercauteren T., Ourselin S., Jorge Cardoso M., 2017, in Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer, Berlin, p. 240
Traub W. A., Oppenheimer B. R., 2010, Exoplanets. Univ. Arizona Press, Tucson, AZ, p. 111
Vigan A., Moutou C., Langlois M., Allard F., Boccaletti A., Carbillet M., Mouillet D., Smith I., 2010, MNRAS, 407, 71
Vigan A. et al., 2021, A&A, 651, A72
Wagner K., Apai D., Kasper M., Kratter K., McClure M., Robberto M., Beuzit J.-L., 2016, Science, 353, 673
Wahhaj Z. et al., 2015, A&A, 581, A24
Wahhaj Z. et al., 2021, A&A, 648, A26
Wainwright M. J., Simoncelli E. P., 2000, Adv. Neural Inform. Process. Syst., 12, 855
Wang L., Wang C., Sun Z., Chen S., 2020, IEEE Access, 8, 167939
Wertz O., Absil O., González C. G., Milli J., Girard J. H., Mawet D., Pueyo L., 2017, A&A, 598, A83
Yalniz I. Z., Jégou H., Chen K., Paluri M., Mahajan D., 2019, preprint (arXiv:1905.00546)
Yip K. H. et al., 2019, in Joint European Conference on Machine Learning and Knowledge Discovery in Databases. p. 322

SUPPORTING INFORMATION

Supplementary data are available at *MNRAS* online.

Figure S1: Detection maps obtained with the selected algorithms (see Section 4.1 of the main paper). Dataset: HD 95086 (2018-01-05).
Figure S2: Zoom near the host star. Dataset: HD 95086 (2018-01-05).
Figure S3: Same caption than Fig. 1. for dataset HD 95086 (2021-03-11).
Figure S4: Same caption than Fig. 2. for dataset HD 95086 (2021-03-11).
Figure S5: Same caption than Fig. 1. for dataset HIP 88399 (2015-05-10).
Figure S6: Same caption than Fig. 2. for dataset HIP 88399 (2015-05-10).
Figure S7: Same caption than Fig. 1. for dataset HIP 88399 (2016-04-16).
Figure S8: Same caption than Fig. 2. for dataset HIP 88399 (2016-04-16).
Figure S9: Same caption than Fig. 1. for dataset HIP 88399 (2018-04-11).
Figure S10: Same caption than Fig. 2. for dataset HIP 88399 (2018-04-11).
Figure S11: Same caption than Fig. 1. for dataset HD 131399 (2015-06-12).
Figure S12: Same caption than Fig. 2. for dataset HD 131399 (2015-06-12).
Figure S13: Same caption than Fig. 1. for dataset HD 131399 (2016-05-07).
Figure S14: Same caption than Fig. 2. for dataset HD 131399 (2016-05-07).
Figure S15: Same caption than Fig. 1. for dataset HIP 65426 (2017-02-09).
Figure S16: Same caption than Fig. 2. for dataset HIP 65426 (2017-02-09).
Figure S17: Same caption than Fig. 1. for dataset HIP 65426 (2018-05-13).
Figure S18: Same caption than Fig. 2. for dataset HIP 65426 (2018-05-13).
Figure S19: Same caption than Fig. 1. for dataset HIP 72192 (2015-06-11).
Figure S20: Same caption than Fig. 2. for dataset HIP 72192 (2015-06-11).
Figure S21: Contrast as a function of the angular separation. For the 2021-03-11 epoch of HD 95086, we experienced during training, validation and inference a few false alarms with the proposed approach in an area localized very near the star. This area is marked

by a gray rectangular, and the contrast is not statistically grounded in the corresponding range of angular separations.
Figure S22: Continuation of Fig. 21.
Figure S23: TPR (in per cent) of the proposed `deep PACO` algorithm for synthetic sources missed by `PACO` (i.e. for S/N ≤ 5) as a function of the angular separation (on the x-axis) and of the `PACO`'s S/N of detection (on the y-axis).
Figure S24: Continuation of Fig. 23.
Figure S25: Mean throughput ($\widehat{\alpha}/\alpha$) on the estimated photometry of injected synthetic sources as a function of the their contrast and of their angular separation.
Table S1: AUC for ROCs giving the TPR as a function of the FDR. The experiments were conducted by considering only known real sources.
Table S2: AUC for ROCs giving the TPR as a function of the FDR. The experiments were conducted by resorting to massive injections of synthetic sources, and known real sources were excluded of this study.

Please note: Oxford University Press is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

APPENDIX A: DISCUSSION ABOUT THE STATISTICAL MODEL AND THE WHITENING OF THE OBSERVATIONS

A1 Refinement of the statistical model

Concerning the statistical model of the nuisance component, the multivariate Gaussian assumption described in Section 2.1.1 is a convenient approximation leading to closed-form expressions for the underlying estimators. However, this formulation neglects all temporal fluctuations of the data. In Flasseur et al. (2020a, b), we consider a refinement of this model using a multivariate Gaussian scaled mixture (GSM, Conte, Lops & Ricci 1995; Wainwright & Simoncelli 2000). It amounts to model the distribution of patch $\boldsymbol{f}_{n,t}$ centered around pixel n at time t by a temporally weighted multivariate Gaussian $\mathcal{N}(\boldsymbol{m}_n, \sigma_{n,t}^2\,\mathbf{C}_n)$. Under this model, the sample estimates $\{\widehat{\boldsymbol{m}}_n; \widehat{\sigma}_{n,t}^2; \widehat{\mathbf{S}}_n\}$ of the local mean, of the temporal scaling factors, and of the covariance coming from the maximum likelihood are obtained with a fixed-point iterative method as follows (Flasseur et al. 2020a):

$$\begin{cases} \widehat{\boldsymbol{m}}_n = \frac{1}{\sum_{t=1}^{T} 1/\hat{\sigma}_{n,t}^2} \cdot \sum_{t=1}^{T} \frac{1}{\hat{\sigma}_{n,t}^2} \mathbf{E}_{n,t}\,\boldsymbol{r} \;\in \mathbb{R}^K, \\ \widehat{\sigma}_{n,t}^2 = \frac{1}{K} (\mathbf{E}_{n,t}\,\boldsymbol{r} - \widehat{\boldsymbol{m}}_n)^\top \widehat{\mathbf{S}}_n^{-1} (\mathbf{E}_{n,t}\,\boldsymbol{r} - \widehat{\boldsymbol{m}}_n) \;\in \mathbb{R}_+, \\ \widehat{\mathbf{S}}_n = \frac{1}{T} \sum_{t=1}^{T} \frac{1}{\hat{\sigma}_{n,t}^2} (\mathbf{E}_{n,t}\,\boldsymbol{r} - \widehat{\boldsymbol{m}}_n)(\mathbf{E}_{n,t}\,\boldsymbol{r} - \widehat{\boldsymbol{m}}_n)^\top \;\in \mathbb{R}^{K\times K}. \end{cases} \tag{A1}$$

The regularized covariance $\widehat{\mathbf{C}}_n$ is estimated on the fly by shrinkage of $\widehat{\mathbf{S}}_n$ as defined in equation (5), with $Q = \left(\sum_{t=1}^{T} 1/\widehat{\sigma}_{n,t}^2\right)^2 / \left(\sum_{t=1}^{T} 1/\widehat{\sigma}_{n,t}^4\right)$ the equivalent number of patches involved in the computation of $\widehat{\mathbf{S}}_n$ in the presence of the weighting factors $\{\widehat{\sigma}_{n,t}^2\}_{t=1:T}$, see Flasseur et al. (2020a). The scaling factors $\{\widehat{\sigma}_{n,t}^2\}_{t=1:T}$ can be interpreted as the local variance of the the residual data (i.e. after centering and whitening). This approach allows to identify and to neutralize outliers, taking the form both of spatially interpolated defective pixels and of local areas displaying fluctuations on some temporal frames larger than on other ones. These properties transfer to the estimators of the statistics of the nuisance component with an improved robustness against bad data, thus leading to a better detection sensitivity and characterization accuracy.

We tested to integrate a GSM model in the pre-processing step of the proposed algorithm. Table A1 compares, with the same procedure as described in Section 2.2.3, the detection performance in terms of precision and recall of the proposed algorithm using either a multivariate Gaussian model (equations 2) or a GSM model (equations A1) in the pre-processing step. It illustrates that, as for `PACO`, the GSM model leads to an improved detection sensitivity, with an improved stability over data sets (i.e. smaller error bars). However, the gain is significantly smaller than the typical gain, reaching 10 per cent–15 per cent, obtained when substituting the multivariate Gaussian assumption with a GSM model in the `PACO` algorithm (Flasseur et al. 2020a). This observation can be attributed to the presence of the additional learning stage of our proposed method that, hopefully, partially correcting for the approximate fidelity (with respect to the observations) of the statistical model embedded in the pre-processing step.

Given the increased computational burden of the proposed detection approach with a GSM model by a factor between 10 and 30 (which represents the typical number of iterations needed to reach convergence of the estimators A1), we recommend, for practical reasons, to use the standard version of `deep PACO` embedding a multivariate Gaussian model, as defined in Section 2.1.1. This is also the choice we made for the presentation of the results in the main core of this paper. The alternative version of the algorithm embedding a GSM model can be reserved to refine, in a second step, the reduction of some data sets with ambiguous detections. When comparing with state-of-the-art detection algorithms, we use for `PACO` the version embedding a GSM model to compare fairly the proposed method against the best setting of existing methods.

Table A1. Comparison between a multivariate Gaussian model (equations 2) and a GSM model (equations A1) for the statistical modelling of the nuisance component in the pre-processing step of the proposed detection approach. Reported scores are AUC of F1R score (best when close to 1) as an overall measurement of the precision-recall trade-off of the underlying detector. Mean results and standard deviation are obtained on the 11 SPHERE-IRDIS data sets considered in this work, see Section 4 for the recording logs.

Ang. sep. (arcsec)	Gaussian model (default)	GSM model (variant)
[0; 2]	0.88 ± 0.03	0.90 ± 0.03
[2; 4]	0.90 ± 0.05	0.91 ± 0.04
[4; 6]	0.89 ± 0.04	0.91 ± 0.01
[6; 7]	0.88 ± 0.06	0.90 ± 0.03

Table A2. Comparison between non-normalized (equations 7) and normalized (equations A2) outputs produced by the pre-processing step of the proposed detection approach. Reported scores are AUC of F1R score (best when close to 1) as an overall measurement of the precision-recall trade-off of the underlying detector. Mean results and standard deviation are obtained on the 11 SPHERE-IRDIS data sets considered in this work, see Section 4 for the recording logs.

Ang. sep. (arcsec)	Not normalized (default)	Normalized (variant)
[0; 2]	0.88 ± 0.03	0.88 ± 0.04
[2; 4]	0.90 ± 0.05	0.92 ± 0.05
[4; 6]	0.89 ± 0.04	0.92 ± 0.03
[6; 7]	0.88 ± 0.06	0.91 ± 0.05

A2 Refinement of the whitening of the observations

The pre-processing step (centering and local whitening) described in Section 2.1.2 is computationally quite efficient since it involves, for non-overlapping square patches, only $\lfloor N/K \rceil$ ($\simeq 10^4$ for the SPHERE-IRDIS instrument) matrix multiplications of size $K \times K$. However, it has two drawbacks: (i) it leads to some spatial discontinuities between adjacent patches, and (ii) it does not account for the spatially non-stationary transformation induced by the whitening process on the off-axis PSF in terms of shape and of intensity. In practice, we observe that this approach sometimes lacks of robustness for observations recorded under medium to bad conditions, that is, where the unmixing between the nuisance component and the sought objects is even more difficult. Limitation (i) can be addressed by considering overlapping patches (e.g. with a patch stride of one pixel). Limitation (ii) can be addressed by adding an output term explicitly accounting for the shape and intensity transformation induced by the whitening process. In this context, the pre-processed images $\widetilde{\boldsymbol{r}}$ in $\mathbb{R}^{N\times 2}$ are now formed by the concatenation of two images $\widetilde{\boldsymbol{a}}$ in $\mathbb{R}^N$ and $\widetilde{\boldsymbol{b}}$ in $\mathbb{R}^N$ defined by:

$$\begin{cases} \widetilde{b}_{n'} = \dfrac{\left[\sum\limits_{n\in\mathbb{P}} \mathbf{E}_n^\top \boldsymbol{h}^\top \widehat{\mathbf{C}}_n^{-1}(\boldsymbol{r}_n - \widehat{\boldsymbol{m}}_n)\right]_{n'}}{q_{n'}}, \forall n' \in [\![1;N]\!], \\ \widetilde{a}_{n'} = \dfrac{\left[\sum\limits_{n\in\mathbb{P}} \mathbf{E}_n^\top \boldsymbol{h}^\top \widehat{\mathbf{C}}_n^{-1} \boldsymbol{h}\right]_{n'}}{q_{n'}}, \forall n' \in [\![1;N]\!], \end{cases} \tag{A2}$$

with $q_{n'}$ the number of patches averaged at each location n' of the field of view, as defined in equation (6). For overlapping square patches[12] with a unit patch stride, $q_{n'}$ is equal to K almost everywhere, excepted on the borders of the field of view where it progressively tends to zero. The term $\boldsymbol{b}$ can be interpreted as the correlation between the whitened off-axis PSF and the centred plus whitened observations, while $\boldsymbol{a}$ is a normalization term representing the autocorrelation of the whitened off-axis PSF. It can be noted that for each pixel n' of the field of view, the ratio $\widetilde{b}_{n'}/\sqrt{\widetilde{a}_{n'}}$ (respectively, the ratio $\widetilde{b}_{n'}/\widetilde{a}_{n'}$) corresponds to the S/N of detection (respectively, to the source flux) that would be estimated by PACO at pixel n' (Flasseur et al. 2018b). To prevent the deep model (built from the outputs of the pre-processing step) to learn only the mapping $\widetilde{\boldsymbol{r}} \to \widetilde{\boldsymbol{b}}/\sqrt{\widetilde{\boldsymbol{a}}}$, we resort to a residual learning procedure. It consists in evaluating the loss function (see Section 2.2.3) jointly on the detection map produced by the proposed algorithm and also on the PACO S/N map $\widetilde{\boldsymbol{b}}/\sqrt{\widetilde{\boldsymbol{a}}}$ computed on the fly. This strategy explicit rewards the deep model to perform better than PACO.

Table A2 compares, with the same procedure as described in Section 2.2.3, the detection performance in terms of precision and recall of the proposed algorithm using either the whitening procedure described in Section 2.1.2, and the whitening procedure described in this appendix. It shows that for the 11 SPHERE-IRDIS data sets we study in details in this work, the variant approach accounting for the transformation induced by the whitening process on the off-axis PSF leads only to a slight improvement of the overall detection performance. This is due to the fact that none of the 11 considered data sets was recorded under bad observing conditions. Fig. A1 gives a qualitative comparison between the two approaches for a data set of HR 8799 (see Section 4 for the recording logs) impacted by the wind-driven halo effect. The variant procedure described in this appendix allows to avoid numerous evident false alarms occurring when the normalization of the pre-processed frames is omitted.

[12]Given the typical circular shape of speckles, we also considered circular overlapping patches. Square and circular patches lead to very comparable detection performance.

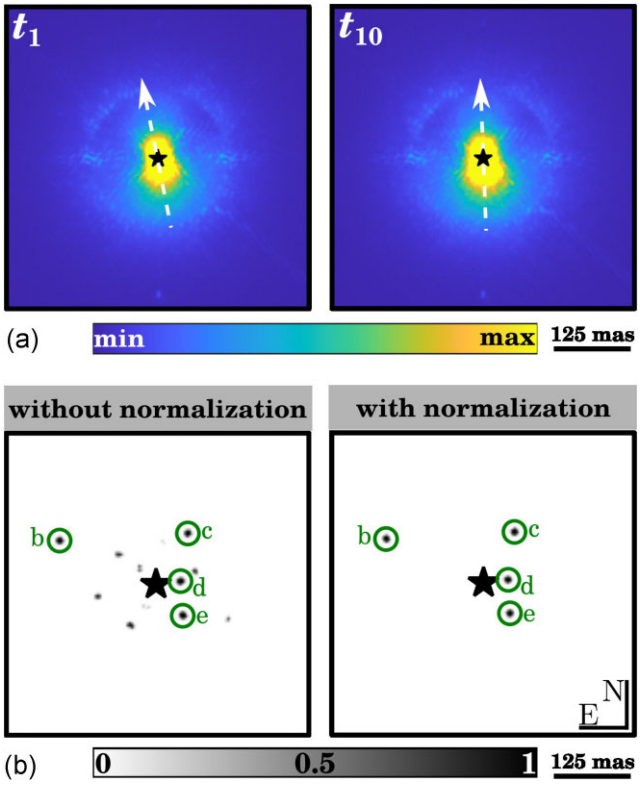

Figure A1. (a) Illustration of two data frames impacted by a wind-driven halo near the star. The direction of main elongation of the halo is symbolized by the arrow. (b) Detection map obtained on the corresponding data set with the proposed detection algorithm embedding the whitening procedure either defined in Section 2.1.2 (i.e. without normalization by the whitened off-axis PSF) or defined in Appendix A2 (i.e. with normalization by the whitened off-axis PSF). The four known exoplanets (HR 8799 b, c, d, e) are circled. Data set: HR 8799 (2015 July 04), see Section 4 for the observation logs.

Compared to the whitening procedure described in Section 2.1.2, the variant described in this appendix has a computational burden increased by a factor $4 \times K$, that is, by typically a factor between 200 and 500 for the VLT/SPHERE instrument. For practical reasons, we recommend to use the standard whitening procedure by default, as defined in Section 2.1.2. This is also the choice we made for the presentation of the results in the main core of this paper. The alternative version of the algorithm accounting for the whitening of the off-axis PSF can be reserved to refine, in a second step, the reduction of some data sets for which obvious and numerous false alarms are experienced at inference time.

APPENDIX B: MAIN SETTINGS OF THE DETECTION AND CHARACTERIZATION STAGES OF THE PROPOSED ALGORITHM

Table B1 gives a quick-look summary of the main settings for the proposed algorithm, both for the detection and the characterization modules. Fields are classified in three categories: *pre-processing*, *generation of the training set*, and *deep learning model*.

Table B1. Summary of the main settings used for the detection and characterization modules of the proposed algorithm.

Parameters	Detection	Characterization
	▶ Pre-processing	
Input	Observations $\boldsymbol{r} \in \mathbb{R}^{N\times T}$	Observations $\boldsymbol{r} \in \mathbb{R}^{N\times T}$
Strategy for known sources (training only)	Temporal shuffling	Source masking + temporal shuffling
Temporal centring	Yes	Yes
Spatial whitening	Yes	No
Whitening patch shape	square	–
Whitening patch area K	$\in [\![7^2; 12^2]\!]$ pixels (automatic)	–
Whitening patch stride	$\lfloor\sqrt{K}\rceil$ pixels (default) or 1 pixel (variants of Appendix A)	–
Whitening statistical model	Multivariate Gaussian (default) or GSM (variant of Appendix A1)	–
Whitening output quantity	$\widehat{\mathbf{L}}_n^\top(\boldsymbol{r}_n - \widehat{\boldsymbol{m}}_n)$ s.t. $\widehat{\mathbf{L}}_n\widehat{\mathbf{L}}_n^\top = \widehat{\mathbf{C}}_n^{-1}$, $\forall n \in \mathbb{P}$ (default) or concat. $\left(\left[\sum_{n\in\mathbb{P}} \mathbf{E}_n^\top \boldsymbol{h}^\top \widehat{\mathbf{C}}_n^{-1}(\boldsymbol{r}_n - \widehat{\boldsymbol{m}}_n)\right]_{n'}; \left[\sum_{n\in\mathbb{P}} \mathbf{E}_n^\top \boldsymbol{h}^\top \widehat{\mathbf{C}}_n^{-1}\boldsymbol{h}\right]_{n'}\right)$, $\forall n' \in [\![1; N]\!]$ (variant of Appendix A2)	–
Parallactic derotation	Yes	Yes
Temporal collapsing	No	Yes (after injections if any)
Implantation	GPUs or CPUs (parallelized)	GPUs or CPUs (parallelized)
Output	Pre-processed observations $\widetilde{\boldsymbol{r}} \in \mathbb{R}^{N\times T}$	Pre-processed observations $\widetilde{\boldsymbol{r}} \in \mathbb{R}^{N\times T}$
	▶ Generation of the training set	
Input	Pre-processed observations $\widetilde{\boldsymbol{r}} \in \mathbb{R}^{N\times T}$	Pre-processed observations $\widetilde{\boldsymbol{r}} \in \mathbb{R}^{N\times T}$
Pre-processing update on injection arcs	Yes	Yes
Data augmentation	Yes	Yes
Total number P of sources	$\in [\![25\,000; 50\,000]\!]$	$\simeq 40\,000$
Number S of training sets	$\in [\![500; 1\,000]\!]$	$\simeq 8000$
Number $P^{[s]}$ of sources per set s	$\in [\![1, 10]\!]$	$\in [\![1; 10]\!]$
Location ϕ of sources	Uniform in polar system	Uniform in Cartesian system
Contrast α of sources	Uniform in $\left[3\widehat{\sigma}_{\phi_p}^{\text{PACO}}; 12\widehat{\sigma}_{\phi_p}^{\text{PACO}}\right]$	Uniform in $[1\times10^{-6}; 3\times10^{-5}]$ (default)
Cropping patch area J	–	31^2 pixels
Implantation	CPUs (parallelized)	CPUs (parallelized)
Output	Sets of pre-processed observations with injections $\{\check{\boldsymbol{r}}^{[s]} \in \mathbb{R}^{N\times T}\}_{s=1:S}$	Sets of pre-processed patches with injections $\{\check{\boldsymbol{p}}^{[p]} \in \mathbb{R}^{J}\}_{p=1:P}$
	▶ Deep learning model	
Input	A set of pre-processed observations with injections $\check{\boldsymbol{r}}^{[s]} \in \mathbb{R}^{N\times T}$	A pre-processed patch with injection $\check{\boldsymbol{p}}^{[p]} \in \mathbb{R}^{J}$
Architecture	U-Net (Res-Net 18 backbone)	Custom VGG-like
Task	Pixelwise classification	Regression
Number of weights	$\simeq$ 11 millions	$\simeq$ 1.2 millions
Pre-trained weights	No	No
Optimization loss	Dice2 (overlap measure)	Absolute relative error
Optimizer	AMSGrad	Adam
Validation metric	F1R score (precision/recall trade-off)	Absolute relative error
Batch size	1 (set of pre-processed observations with injections)	1024 (pre-processed patches with injections)
Weight decay	10^{-5}	0
Initial learning rate	10^{-3}	10^{-3}
Learning rate scheduling	Yes (-10 per cent every 10 epochs)	Yes (-70 per cent every 50 epochs)
Number of epochs[a]	$\in [\![1; 100]\!]$ (on the fly)	300 (fixed)
Implantation	GPUs	GPUs
Output	Detection map $\widehat{\boldsymbol{y}} \in [0; 1]^M$	Photometry estimates $\widehat{\boldsymbol{\alpha}} \in \mathbb{R}_+$

Notes. [a]We recall that the training sets are generated on the fly for the detection stage, that is, the notion of epochs is used only to schedule the learning rate. For the characterization stage, the term epoch is used in a classical meaning so that the generated patches are seen multiple times (corresponding to the number of epochs) by the network.

This paper has been typeset from a TEX/LATEX file prepared by the author.