# DeepGhostBusters: Using Mask R-CNN to Detect and Mask Ghosting and Scattered-Light Artifacts from Optical Survey Images

Dimitrios Tanoglidis[a,b,*], Aleksandra Ćiprijanović[c], Alex Drlica-Wagner[c,b,a], Brian Nord[c,b,a], Michael H. L. S. Wang[c], Ariel Jacob Amsellem[b], Kathryn Downey[a], Sydney Jenkins[a,e], Diana Kafkes[c], Zhuoqi Zhang[d]

[a]*Department of Astronomy and Astrophysics, University of Chicago, Chicago, IL 60637, USA*
[b]*Kavli Institute for Cosmological Physics, University of Chicago, Chicago, IL 60637, USA*
[c]*Fermi National Accelerator Laboratory, P. O. Box 500, Batavia, IL 60510, USA*
[d]*University of Chicago, Chicago, IL 60637, USA*
[e]*Department of Physics, Massachusetts Institute of Technology, 77 Massachusetts Ave., Cambridge, MA 02139, USA*

**Abstract**

Wide-field astronomical surveys are often affected by the presence of undesirable reflections (often known as "ghosting artifacts" or "ghosts") and scattered-light artifacts. The identification and mitigation of these artifacts is important for rigorous astronomical analyses of faint and low-surface-brightness systems. However, the identification of ghosts and scattered-light artifacts is challenging due to a) the complex morphology of these features and b) the large data volume of current and near-future surveys. In this work, we use images from the Dark Energy Survey (DES) to train, validate, and test a deep neural network (Mask R-CNN) to detect and localize ghosts and scattered-light artifacts. We find that the ability of the Mask R-CNN model to identify affected regions is superior to that of conventional algorithms and traditional convolutional neural networks methods. We propose that a multi-step pipeline combining Mask R-CNN segmentation with a classical CNN classifier provides a powerful technique for the automated detection of ghosting and scattered-light artifacts in current and near-future surveys.

*Keywords:* Deep Learning, Object Detection, Image Artifacts

## 1. Introduction

Wide-field photometric surveys at optical and near-infrared wavelengths have provided a wealth of astronomical information that has enabled a better understanding of the processes that govern the growth and evolution of the Universe and its contents. Near-future surveys, such as the Vera C. Rubin Observatory's Legacy Survey of Space and Time (LSST; Ivezić et al., 2019)[1], will further expand our knowledge of the Universe by extending measurements to unprecedentedly faint astronomical systems. Such surveys will produce terabytes of data each night and measure tens of billions of stars and galaxies.

Images collected by optical/near-infrared surveys often contain imaging artifacts caused by scattered and reflected light (commonly known as "ghosting artifacts" or "ghosts") from bright astronomical sources. These image artifacts are an unavoidable feature of many optical systems. The effective mitigation of ghosts and scattered-light artifacts, and the spurious brightness variations they introduce, is important for the detection and precise measurement of faint astronomical systems. In particular, since many ghosts cover a large image area with relatively low surface brightness, they constitute a significant source of contamination in studies of the low-surface-brightness Universe, a major goal of current and upcoming surveys (e.g., Greco et al., 2018; Brough et al., 2020; Kaviraj, 2020; Tanoglidis et al., 2021b).

Modern wide-field telescopes and instruments greatly reduce the occurrence and intensity of

*Corresponding author; FERMILAB-PUB-21-374-AE
*Email address:* dtanoglidis@uchicago.edu (Dimitrios Tanoglidis)

[1] https://www.lsst.org/

ghosts and scattered-light artifacts by introducing light baffles, and high efficiency anti-reflective coatings on key optical surfaces. Strict requirements on the number and intensity of ghosts and scattered-light artifacts were achieved during the construction of the Dark Energy Camera (DECam; Abbott et al., 2009; Flaugher et al., 2015), which has enabled state-of-the-art cosmological analyses with the Dark Energy Survey (DES; DES Collaboration, 2005, 2016; DES Collaboration et al., 2018, 2021b).[2], Other smaller surveys have implemented novel optical designs to mitigate the presence of ghosts and scattered-light artifacts (Abraham and van Dokkum, 2014).

Despite these successful efforts, it is often impossible to completely remove ghosts and scattered-light artifacts. For example, the DES 3-year cosmology analyses masked ∼3% of the survey area around the brightest stars, and ∼10% of the survey area around fainter stars (Sevilla-Noarbe et al., 2021). Additional mitigation steps that go beyond the original survey design requirements are particularly important for studies of low-surface-brightness systems.

The large datasets produced by surveys like DES make the rejection of these residual artifacts by visual inspection infeasible. The situation will become even more intractable in upcoming surveys, like LSST, which will collect ∼ 20TB/night and ∼ 15PB of data over its nominal 10-year survey.[3] Furthermore, the deeper imaging of LSST will place even tighter requirements on low-surface-brightness artifacts (LSST Science Collaboration, 2009; Brough et al., 2020).

To mitigate residual ghosts and scattered-light artifacts, DES uses a predictive Ray-Tracing algorithm as the core of its detection process. This algorithm forward models the physical processes that lead to ghosting/scattered-light events (Kent, 2013), such as the configuration of the telescope and camera optics, and the positions and brightnesses of known stars obtained from catalogs external to the survey (for a more detailed description of the Ray-Tracing algorithm, see Kent 2013 and Sec. 2 of Chang et al. 2021). While the Ray-Tracing algorithm is largely successful in predicting the presence and location of artifacts in the images, this algorithm is also limited in predicting the amplitude of the ghost image by the accuracy of the optical model and the external star catalogs used.

Recently, Chang et al. (2021) demonstrated an alternative approach using a convolutional neural network (CNN; Lecun et al., 1998) to classify DES images containing ghosts and scattered-light artifacts. CNNs constitute a class of deep neural networks that are inspired by the visual cortex and optimized for computer vision problems. Since their invention, CNNs have found numerous applications in the field of astronomy, including galaxy morphology prediction (e.g., Dieleman et al., 2015; Cheng et al., 2021), star-galaxy separation (e.g., Kim and Brunner, 2017), identification of strongly lensed systems (e.g., Lanusse et al., 2018; Davies et al., 2019; Bom et al., 2019; Huang et al., 2020, 2021), classifying galaxy mergers (e.g., Ćiprijanović et al., 2021), and many other applications. The CNN developed by Chang et al. (2021) was able to predict whether an image contained ghosts or scattered-light artifacts with high-accuracy (∼ 96% in the training set, ∼ 86% in the test set), but did not identify the specific pixels of the image that were affected by the presence of artifacts. Since ghosts and scattered-light artifacts often affect a subregion of an image, flagging entire images rejects a significant amount of high-quality data.

In contrast to classic CNNs, object detection algorithms are designed to determine the location of objects in an image (e.g., place bounding boxes around objects or mask exact pixels that belong to objects). In this work, we study the use of a deep learning-based object detection algorithm, namely a Mask Region-Based Convolutional Neural Network (Mask R-CNN; He et al., 2017), to predict the location of ghosts and scattered-light artifacts in astronomical survey images. Mask R-CNNs have recently been demonstrated as an accurate tool to detect, classify, and deblend astronomical sources (stars and galaxies) in images (Burke et al., 2019).

Using 2000 manually annotated images, we train a Mask R-CNN model to identify artifacts in DES images. Comparing the results to those of the Ray-Tracing algorithm on ghost-containing images, we find that Mask R-CNN performs better in masking affected regions — indicated by the value of the $F1$ score (a combination of precision and recall). This demonstrates that deep learning-based object detection algorithms can be effective in helping to address a challenging problem in astronomical surveys without any *a priori* knowledge of the optical system used to generate the images.

[2] `https://www.darkenergysurvey.org/`

[3] `https://www.lsst.org/scientists/keynumbers`

This paper is organized as follows. In Sec. 2, we present the dataset, including the annotation process, used in this work. In Sec. 3, we describe the Mask R-CNN algorithm, implementation, and the training procedure. In Sec. 4 we present results from the Mask R-CNN model, including examples of predicted masks, custom and commonly used evaluation metrics, and we compare its performance to that of a conventional algorithm. We further summarize our results and their applications, and conclude in Sec. 5. The code and data related to this work are publicly available at the GitHub page of this project: `https://github.com/dtanoglidis/DeepGhostBusters`.

## 2. Data

In this section, we describe the datasets used for training and evaluating the performance of the Mask R-CNN algorithm for detecting ghosts and scattered-light artifacts. We briefly describe the DES imaging data, our manual annotation procedure, the creation of masks, and the agreement between the human annotators who performed these tasks.

### *2.1. Dark Energy Survey Data*

DES is an optical/near-infrared imaging survey that completed six years of observations in January 2019. The DES data cover $\sim$ 5000 deg$^2$ of the southern Galactic cap in five photometric filters, $grizY$, to a depth of $i \sim 24$ mag (DES Collaboration et al., 2021a). The observations were obtained with DECam, a 570-megapixel camera mounted on the 4m Blanco Telescope at the Cerro Tololo Inter-American Observatory (CTIO) in Chile (Flaugher et al., 2015). The focal plane of DECam consists of 62 $2048 \times 4096$-pixel red-sensitive scientific charge-coupled devices (CCDs), while its field-of-view covers 3 deg$^2$ with a central pixel scale of $0.263''$.

Our data come from the full six years of DES observations (DES Collaboration et al., 2021a). For the training, validation, and testing of the Mask R-CNN model, we use 2000 images that cover the full DECam focal plane and are known to contain ghosts and scattered-light artifacts. These are part of the positive sample used in Chang et al. (2021) to train a CNN classifier to distinguish between images with and without ghosts. This dataset was assembled by selecting images that the Ray-Tracing program identified as likely to contain ghosts, and subsequently visually inspecting them to correct for false detections.

As described in Chang et al. (2021), the image data were down-sampled images of the full DECam focal plane. Images were produced with the STIFF program (Bertin, 2012), assuming a power-law intensity transfer curve with index $\gamma = 2.2$. Minimum and maximum intensity values were set to the 0.005 and 0.98 percentiles of the pixel value distribution, respectively. The pixel values in each image were then normalized to a range whose minimum and maximum corresponded, respectively, to the first quartile $Q_1(x)$ and third quartile $Q_3(x)$ of the full distribution in the image, by multiplying each pixel value, $x_i$, by a factor $s_i = \frac{x_i - Q_1(x)}{Q_3(x) - Q_1(x)}$. Focal plane images were originally derived as $800 \times 723$-pixel, 8-bit grayscale images in Portable Network Graphics format, which were then downsampled to $400 \times 400$ pixels for use with the Mask R-CNN. The data from Chang et al. (2021) are publicly available.[4]

### *2.2. Annotation process*

Training the Mask R-CNN algorithm requires both images and ground-truth segmentation masks identifying objects of interest in each image. To create these masks, we used the VGG Image Annotator (VIA; Dutta and Zisserman (2019))[5], a simple manual annotation software for images, audio, and video. We split the 2000 images into batches of 100 images, and we randomly assigned each batch to one of eight authors for annotation.[6]

During manual annotation, we categorized the ghosting and scattered-light artifacts into three distinct morphological categories:

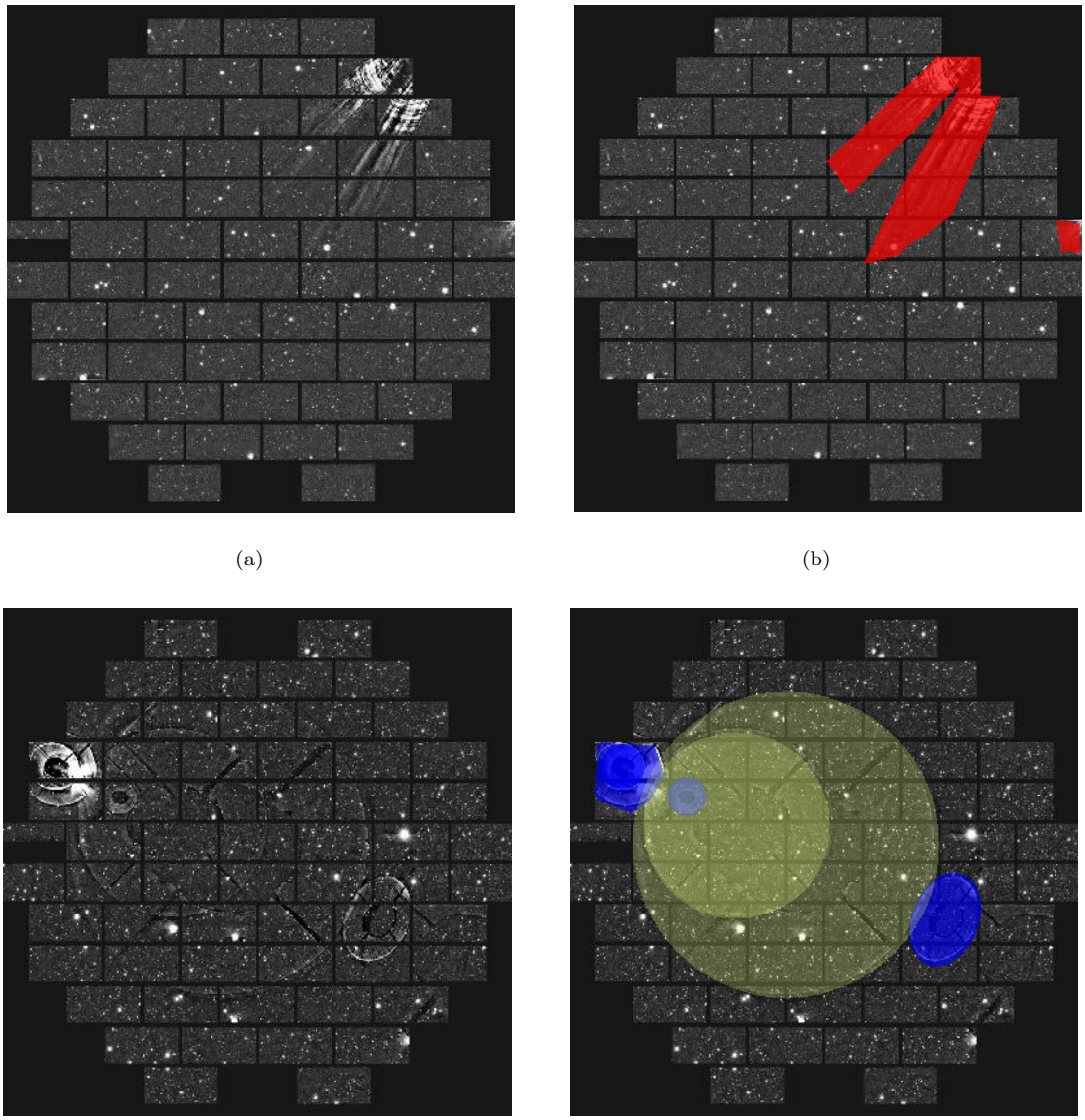
1. 'Rays': These are scattered-light artifacts originating from the light of off-axis stars scattering off of the DECam filter changer (Kent, 2013). They emanate from one of the edges of the image and span several CCDs. This is the most distinct artifact category and is not commonly confused with either of the other two categories.

2. 'Bright': These are high-surface-brightness ghosting artifacts that come from multiple reflections off the DECam focal plane and the C4

[4] `https://des.ncsa.illinois.edu/releases/other/paper-data`

[5] `https://www.robots.ox.ac.uk/~vgg/software/via/`

[6] Note that not every author annotated the same number of images; six of us annotated 200 images and two of us annotated 400 images.



(a) (b)

(c) (d)

Figure 1: Examples of full-focal-plane DECam images containing ghosts and scattered-light artifacts. The corresponding "ground truth" masks (right) were manually annotated. There are three categories of ghosting artifacts: image (a) contains a scattered-light artifact classified as 'Rays'; image (b) shows the masks for the 'Rays' in red; image (c) contains both 'Bright' and 'Faint' ghosts, and the corresponding masks in blue and yellow, respectively, are shown in image (d).

or C5 lenses (Kent, 2013). They are usually relatively small in size and circular or elliptical in shape. They have more distinct borders and are considerably brighter compared to the following category.

3. 'Faint': These are lower-surface-brightness ghosting artifacts that come from multiple reflections between the focal plane and the C3 lens or filter, or internal reflections off of the faces of the C3, C4, and C5 lenses (Kent, 2013). They are circular or elliptical in shape and are usually larger in size and significantly fainter than 'Bright' ghosts.

In Fig. 1, we present two examples of DECam images that contain ghosts and scattered-light artifacts, along with the annotated ground truth

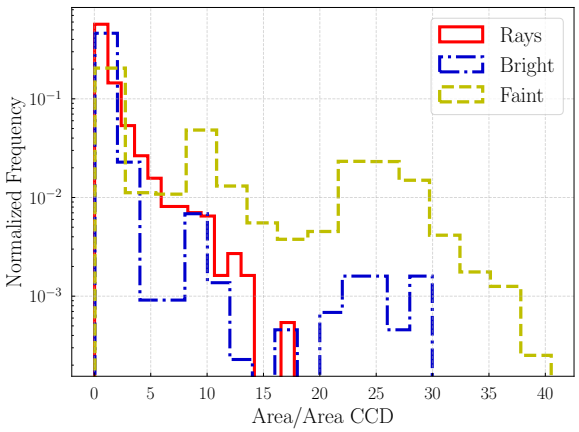



Figure 2: Histograms of the distribution in size (area) of the three artifact types presented in this work. The areas are quoted as multiples of the area of a single CCD.

masks. We trained the Mask R-CNN for these three distinct categories due to their significant morphological difference.

In total, our dataset contains 1566 'Rays', 2197 'Bright', and 2949 'Faint' artifact instances. In Fig. 2, we present the distribution in size (area) of these three ghost categories. The area of each ghosting artifact is presented as a fraction of the area of a single DECam CCD (area of artifacts in pixel over area of a CCD in pixels). Most 'Rays' have an area that covers fewer than 10 CCDs. 'Bright' ghosts are also relatively small in size, with a few spanning more than a couple of CCDs. On the other hand, 'Faint' ghosts are large in size, with a significant fraction of them covering an area of 20–30 CCDs. Many images contain multiple ghosts or scattered-light artifacts.

We note that the ghosting and scattered-light artifacts do not always have clear boundaries (especially those of type 'Rays') and that the distinction between 'Bright' and 'Faint' ghosts is not always well defined. For that reason we expect some disagreement between the human annotators in the extent and shape of the ground truth masks and in the assigned labels.

In Fig. 3, we overlay the masks generated by all eight annotators for the same two DECam images presented in Fig. 1. The colors correspond to the number of annotators that have labeled the region as containing an artifact; dark purple corresponds to fewer votes, while light yellow corresponds to more votes. We do not distinguish between the different artifact types in this image.

The right panel of Fig. 3 shows a significant variation in the masks created by the different annotators for the 'Rays'. The left panel shows generally good agreement between the different annotators for the most prominent ghosts in the image; however, there is a large area on the right of the image that is labeled by only two annotators. We discuss the agreement between the human annotators in more detail in Appendix A. In Section 4, we demonstrate that the Mask R-CNN is able to out-perform conventional algorithms even in the presence of the label noise introduced by disagreements in the existence, mask region, and classification of artifacts by individual annotators. Reduction in label noise from more uniform annotation could improve the performance of the algorithm in the future.

## 3. Methods

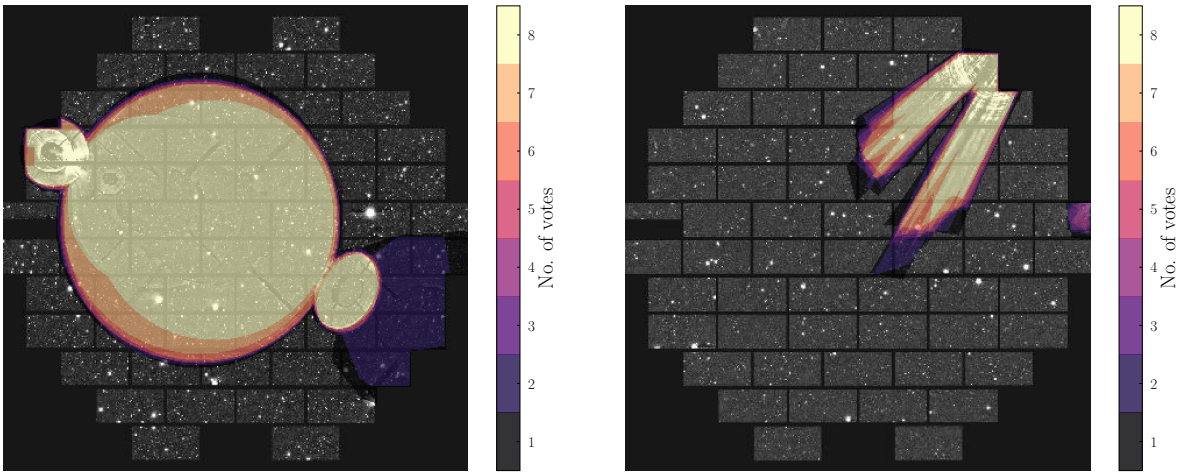
We use Mask R-CNN (He et al., 2017), a popular, state-of-the art instance segmentation algorithm, to detect and mask ghost and scattered-light artifacts.

Mask R-CNN is a powerful and complex algorithm, the latest in a series of object detection models, collectively known as the R-CNN family.[7] It builds upon many deep learning and computer vision techniques; we refer the reader to Weng (2017) for a detailed description of the R-CNN family.

Instance segmentation (e.g., for a review, Mueed Hafiz and Mohiuddin Bhat, 2020) combines the functions of object detection and image segmentation algorithms. Object detection (e.g., for a review, Zhao et al., 2018) is an active area of research in computer vision, with the goal of developing algorithms that can find the positions of objects within an image. Semantic segmentation (e.g., for a review, Minaee et al., 2020) on the other hand refers to the problem of pixel-level classification of different parts of an image into pre-defined categories. Instance segmentation is used to simultaneously detect objects in an image and to create a segmentation mask for each object.

A schematic description of the Mask R-CNN workflow is presented in Fig. 4. In the first stage of the model, the input images are fed into a pre-trained deep CNN — such as VGG (Simonyan and Zisserman, 2014) or ResNet (He et al., 2015) —

[7] Mask R-CNN is the latest in the R-CNN family for 2D object detection. Mesh R-CNN (Gkioxari et al., 2019) is a more recent addition to the family, and it is able to predict 3D shapes of the detected objects.




(a) (b)

Figure 3: Masks created by the eight different annotators (overlaid on top of each other) for the same two images presented in Fig. 1. The colors indicate the number of annotators that have labeled a given pixel as containing a ghost, from dark purple (one annotator) to light yellow (all the eight annotators).

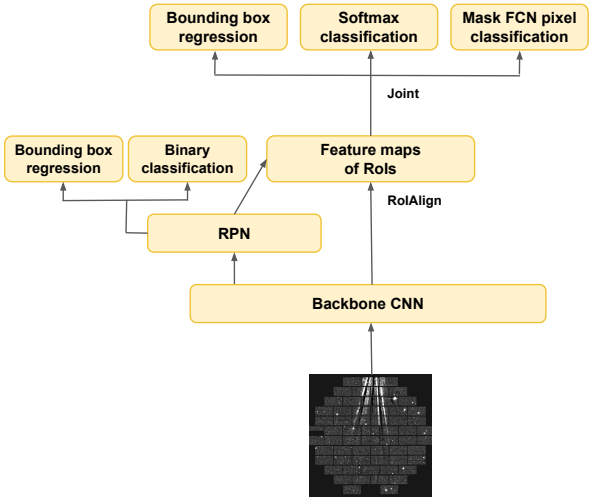



Figure 4: High-level schematic overview of the Mask R-CNN model. Figure adapted from Weng (2017).

also called the *backbone* network. The last, fully connected, classification layers of this network have been removed, and thus its output is a feature map. This feature map[8] is passed into the Region Proposal Network (RPN) to produce a limited number of Regions of Interest (RoIs) to be passed to the main network – i.e., candidate regions that are most likely to contain an object.

The RPN is a simple CNN that uses a sliding window to produce a number of *anchor boxes* – boxes of different scales and aspect ratios – at each position. When training the RPN network, two problems are considered — classification and regression. For classification, the algorithm considers the possibility that there is an object (without considering the particular class) that fits inside an anchor box. For regression, the best anchor box coordinates are predicted. The anchor boxes with the highest object-containing probability scores are passed as RoIs in the next step. The loss of the RPN network is composed of a binary classification loss, $L_{\text{RPN,cls}}$, and a bounding box regression loss, $L_{\text{RNP,bbox}}$, such that $L_{\text{RPN}} = L_{\text{RPN,cls}} + L_{\text{RNP,bbox}}$.

Each of the proposed RoIs has a different size.

[8] In practice, most Mask R-CNN implementations – like the one we are using in this work – use a Feature Pyramid Network (FPN; Lin et al., 2016) on top of the backbone. The FPN combines low-level features extracted from the initial stages of the backbone CNN with the high-level feature map output of the last layer. This improves the overall accuracy of the model, since it better represents object at multiple scales.

However, the fully connected networks used for prediction require inputs of the same size. For that reason, the RoIAlign method is used to perform a bilinear interpolation on the feature maps within the area of each RoI and output the interpolated values within a grid of specific size, giving fixed-size feature maps of the candidate regions.

Finally, these reshaped regions are passed to the last part of the Mask R-CNN that performs three tasks in parallel. A softmax classifier learns to predict the class of the object within the RoI; the output is one of the $K + 1$ classes, where $K$ are the different possible object types ($L_{\text{cls}}$ loss), plus one background class. A regressor learns the best bounding box coordinates ($L_{\text{bbox}}$ loss). Finally, the regions pass through a Fully Convolutional Network (FCN) that performs semantic segmentation ($L_{\text{mask}}$ loss), i.e. a per-pixel classification, that creates the masks. The total loss of this Mask R-CNN part is thus $L_{\text{tot}} = L_{\text{cls}} + L_{\text{bbox}} + L_{\text{mask}}$.

The *DeepGhostBusters* algorithm is the Mask R-CNN implementation by Abdulla (2017), trained on our manually annotated dataset of ghosting and scattered-light artifacts. This code is written in Python using the high-level `Keras`[9] library using a `TensorFlow`[10] backend. We use the default 101-layer deep residual network (ResNet-101; He et al. 2015) as the backbone convolutional neural network architecture.

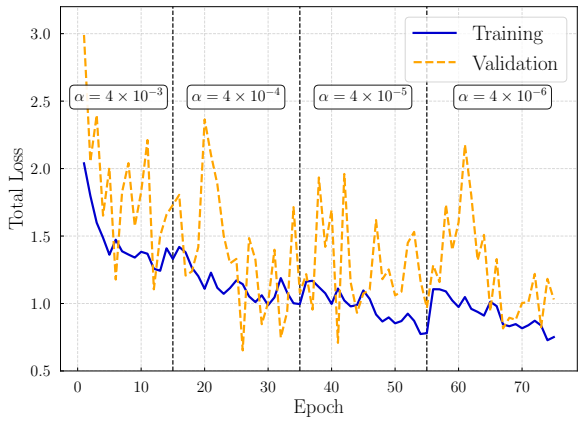



Figure 5: Total loss of the Mask R-CNN model as function of the training epoch. The training is performed using a progressively smaller learning rate, $\alpha$.

Before training, we randomly split the full dataset of 2000 images into a training set (1400 images), a validation set (300 images), and a test set (300 images). The annotation process was performed before this random split. Such a random split is generally important in machine learning problems for these three sets to be representative of the general population, but it becomes even more important here because different human annotators have different annotation styles. This could create significant systematic differences between the ground truth masks in the datasets if not properly randomized.

In computer vision problems where only a small training set is available, it is common to use *transfer learning* to improve results (for recent reviews, see Wang and Deng 2018 and Zhuang et al. 2019). Transfer learning is a process where the weights of a network that has already been trained for one detection task are used for a different, but related, task, usually with some further training. This speeds up the training process, reduces overfitting, and produces more accurate results. Here, we initialize the learning procedure (i.e., use transfer learning) using the weights learned from training Mask R-CNN on the Microsoft Common Objects in Context (MS COCO) dataset[11] (Lin et al., 2014), which consists of $\sim$ 330k images ($\sim$ 2.5M object instances) of 91 classes of common or everyday objects.

To reduce overfitting, we employ data augmentation (e.g., Shorten and Khoshgoftaar, 2019), by performing geometric transformations on the images and the masks. Specifically, we randomly apply zero to three of the following transformations:

- Rotation of the image and the masks by 270 degrees.
- Left-right mirroring/flip of the images and masks.
- Up-down mirroring/flip of the images and masks.

We re-train our model using stochastic gradient descent to update the model parameters. Similarly to what was proposed in Burke et al. (2019), the training is performed in different stages with progressively smaller learning rates, $\alpha$, at each stage. This allows for a deeper learning and finer tuning of the weights, while minimizing the risk of overfitting.

[9] `https://keras.io/`
[10] `https://www.tensorflow.org/`
[11] `https://cocodataset.org/#home`

Specifically, in the first stage (15 epochs), we retrain the top layers only and use a learning rate of $\alpha = 4 \times 10^{-3}$. Then, we train all the layers with decreasing learning rates: 20 epochs at $\alpha = 4 \times 10^{-4}$, 20 epochs at $\alpha = 4 \times 10^{-5}$, and 20 epochs at $\alpha = 4 \times 10^{-6}$. In total, we trained the model for 75 epochs, after which overfitting occurs. In all stages (training, validation, test) we ignore detections with less than 80% confidence (`DETECTION_MIN_CONFIDENCE = 0.8`). We utilized the 25 GB high-RAM Nvidia P100 GPUs available through the Google Colaboratory (Pro version). The training took $\sim$ 4 hours to complete. The inference time is $\sim 0.34$s per image to predict.

In Fig. 5, we present the total loss as a function of the training epoch for both training and validation sets. In Appendix B, we show the training history for the individual components of the total loss.

## 4. Results

We use an independent DECam test set to evaluate the performance of the *DeepGhostBusters* Mask R-CNN in detecting and masking ghost and scattered-light artifacts. We use both custom metrics appropriate for the problem at hand and metrics commonly used in the object detection literature. We also compare the performance of *DeepGhostBusters* with the conventional Ray-Tracing algorithm. Finally, we test the classification performance of *DeepGhostBusters* when it is presented with a dataset that also contains images that lack any ghosts or scattered-light artifacts.

### *4.1. Example Performance*

We first present the mask and class predictions of the *DeepGhostBusters* Mask R-CNN model on four example images (Fig. 6). The two top panels, (a) and (b), correspond to the same images whose ground truth masks were presented in Fig. 1. As in Fig. 1, the different colors represent the different ghosting artifact types: red for 'Rays', blue for 'Bright', and yellow for 'Faint'.

These examples demonstrate both the successes and failures of our model. For example, in panel (a) the model has successfully masked most of the central 'Faint' ghost, but it has also missed a significant part of its periphery, as well as the prominent ghost on the right of the image. Furthermore, although it has successfully deblended and separately masked the small 'Bright' ghost that is superimposed on the larger 'Faint' one, it has only partially masked the one on the left. Panel (b) presents a characteristic example of a false positive detection: predicting a mask for a 'Faint' ghost that is not there. The Mask R-CNN has predicted a mask that successfully covers most of the prominent 'Rays'-type artifact; it is also able to detect the smaller 'Rays' on the right. However, it has also erroneously masked a large central region (containing the edges of the rays) as a 'Faint' ghost. Panels (c) and (d) present mostly successful detections, although with some false negatives, as the undetected 'Faint' ghost on the top-left corner of panel (d). We next formally quantify and evaluate the performance of the Mask R-CNN model and compare it with that of the conventional Ray-Tracing algorithm.

### *4.2. CCD-based metrics*

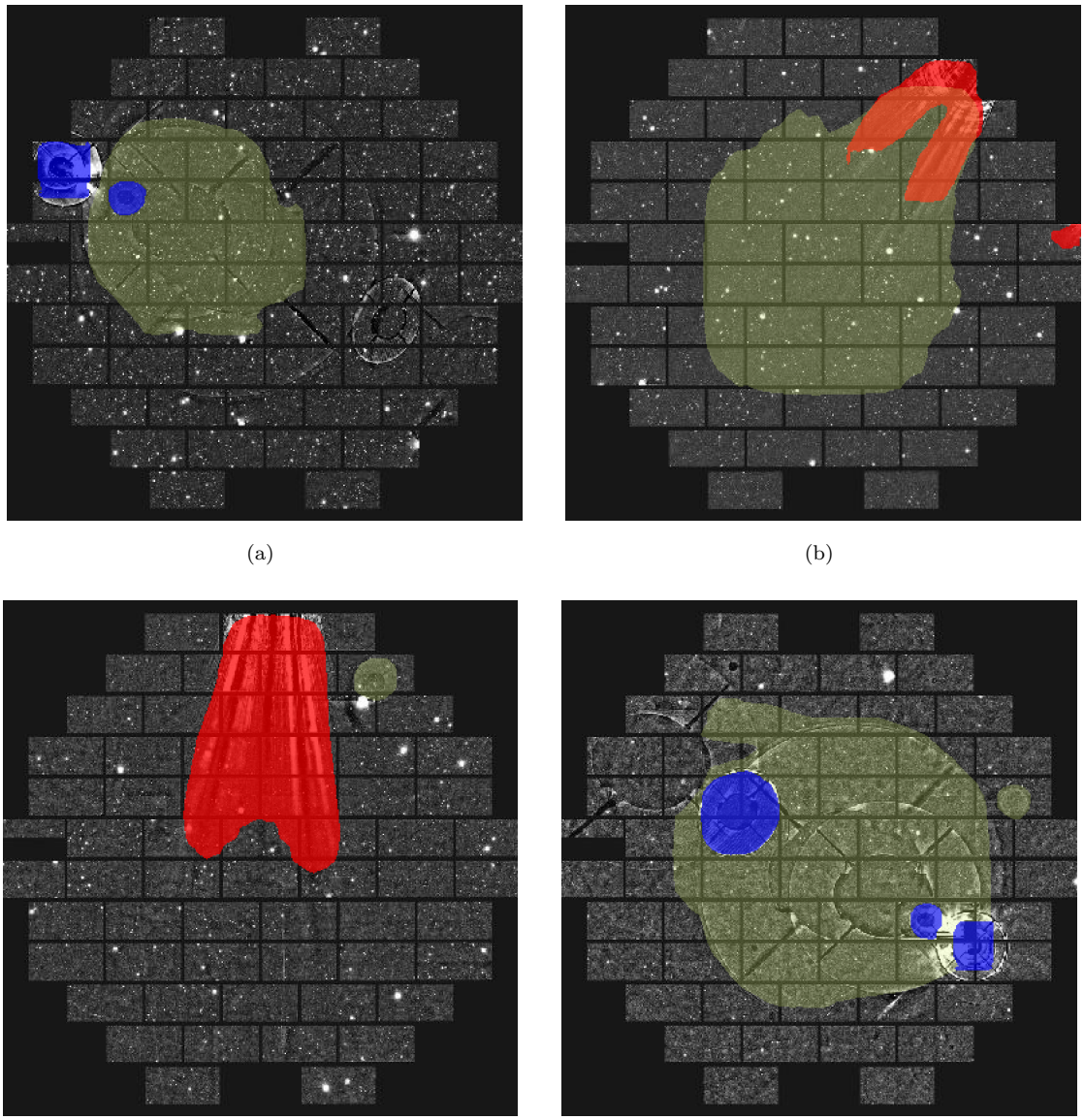
The DECam focal plane consists of 62 science CCDs. The conventional Ray-Tracing algorithm used by DES flags affected focal plane images on a CCD-by-CCD basis — i.e., if a CCD contains a ghost or scattered-light artifact, the entire CCD is removed from processing. To compare the performance of the Mask R-CNN to the conventional algorithm, we develop metrics that are based on whether a CCD contains a ghost or scattered-light artifact.

The resulting metrics depend on the size of individual artifacts. This is important for the problem at hand: for example, we care how well the algorithm can mask a larger ghost compared to a smaller one. At the same time, given the challenges of this problem (e.g., overlapping sources and borders that are not always well defined), assessing the performance at the CCD-level can be more robust than comparisons at the more granular pixel level.

We consider each image as a 1D array of length 62 with entries 0 and 1, where 0 corresponds to CCDs that do not contain a ghost, and 1 corresponds to those that do contain a ghost. For a batch of $M$ images containing $N = 62 \times M$ CCDs, we define the number of true positives ($N^{TP}$), true negatives ($N^{TN}$), false positives ($N^{FP}$), and false negatives ($N^{FN}$). Then, we define the CCD-based precision (purity) and recall (completeness) as:

$$\text{Precision}_{CCD} = \frac{N^{TP}}{N^{TP} + N^{FP}}, \tag{1}$$

$$\text{Recall}_{CCD} = \frac{N^{TP}}{N^{TP} + N^{FN}}. \tag{2}$$



(a) (b)

(c) (d)

Figure 6: Predicted masks on four example images that contain the three distinct artifact types — scattered-light 'Rays' (red), 'Bright' ghosts (blue), and 'Faint' ghosts (yellow). The top panels correspond to the images presented in Fig. 1.

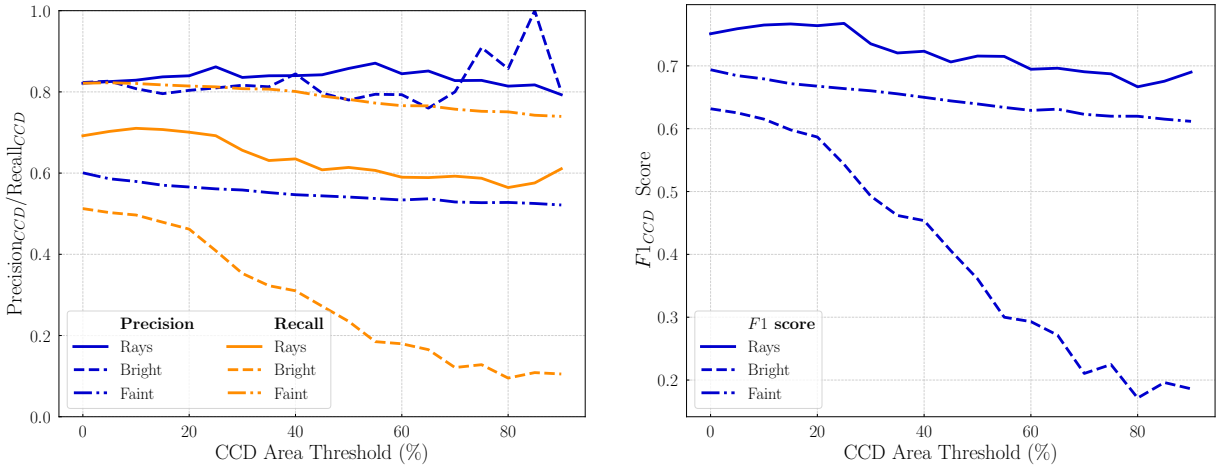
Based on the science case of interest, one may want to maximize either the precision or the recall. For example, for systematic studies of low-surface-brightness galaxies, high recall for ghosts and scattered-light artifacts may be preferred at the expense of some loss in precision.

One approach to assessing the trade-off between precision and recall is to define the $F1$ score, which is the harmonic mean of the precision and recall,

$$F1_{CCD} = 2\left(\frac{\text{Precision}_{CCD} \cdot \text{Recall}_{CCD}}{\text{Precision}_{CCD} + \text{Recall}_{CCD}}\right) \quad (3)$$

Note that we can use the above definitions for each type of artifact individually or for all artifact types combined.

The above metrics are based on the notion of a binary classification of CCDs as affected by ghosts or scattered-light artifacts. In reality, the ghosts and




(a) (b)

Figure 7: CCD-based (a) precision (blue) and recall (orange), and (b) $F1$ score as a function of the CCD area threshold (see main text) from the Mask R-CNN model and for the three ghosting artifact categories ('Rays', 'Bright', and 'Faint').

scattered-light artifacts will only cover some fraction of the CCD area. Thus, we define a threshold for the fraction of the CCD area that must be covered for the CCD to be classified as affected. In Appendix C we present examples of masked CCDs for two different area thresholds. Here, we study how the performance metrics change as a function of that threshold.

In panel (a) of Fig. 7, we present precision and recall as a function of the CCD area threshold for the three artifact categories individually. These metrics are related to the number of CCDs (as opposed to the number of artifacts) that were correctly or incorrectly classified. Therefore, the differences we observe between the artifact types depend on the different sizes of the artifacts. For example, as we have seen (Fig. 2), 'Faint' ghosts tend to cover $\sim 10-30$ CCDs, while 'Bright' ghosts are significantly smaller, covering $\sim 1-3$ CCDs. Thus, the classification and masking of a single large 'Faint' object has a greater effect on the metrics than the detection of two or three 'Bright' ghosts.

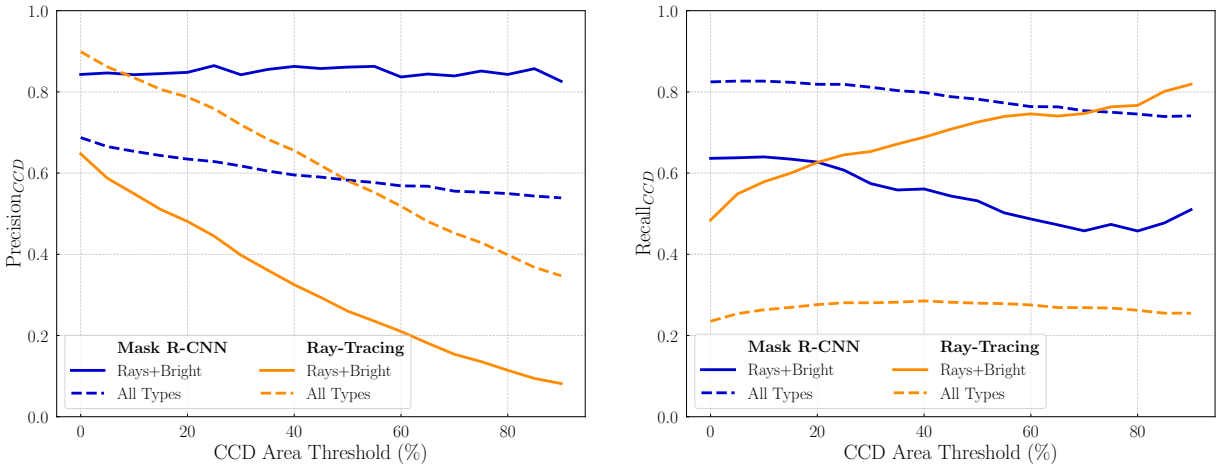
There are a few interesting trends to notice in this figure. First, for 'Rays' and 'Bright' ghosts, the precision is higher than the recall and almost constant as the area threshold changes. The high precision score ($\sim 80\%$) for these categories is easy to understand: these are the most distinct and prominent ghosts, and thus it is hard for a CCD with a 'Faint' ghost (or for a CCD without a ghost) to be mistaken as containing either of these types of artifacts.

Second, the recall score for 'Rays' is $\sim 70\%$ and constant as a function of the threshold. The recall score for 'Bright' ghosts greatly degrades with area threshold and it is generally low (less than 50%). 'Bright' ghosts are relatively small, only partially covering the CCDs that contain them; as we increase the area threshold, only a few such ghosts can pass it.

A third interesting point is that 'Faint' ghosts have higher recall than precision, in contrast to the two other categories. 'Faint' ghosts are usually large: even though some may go undetected, the largest cover many CCDs and are usually detected (at least partially), thus pushing the CCD-based recall (completeness) to higher values. On the other hand, some 'Bright' ghosts, especially those with a significant overlap with larger 'Faint' ghosts can be misclassified as 'Faint', leading to a lower precision.

In panel (b) of Fig. 7, we present the $F1$ score as a function of the CCD area threshold. The $F1$ score (see Eq. (3)) is useful as a way to compare the performance of the classifier for different ghost types using a single metric. As we can see in this figure, the Mask R-CNN performs best in finding CCDs containing 'Rays', while CCDs containing 'Faint' ghosts are identified with higher efficiency than CCDs containing 'Bright' ghosts.

In practice, we are interested in the ability of the *DeepGhostBusters* Mask R-CNN to detect combinations of ghosts and scattered-light artifacts. We present the CCD-based precision and recall as a




(a) (b)

Figure 8: CCD-based (a) precision and (b) recall of the Mask R-CNN model (blue lines) and the Ray-Tracing algorithm (orange lines). We consider both the combination of all types of artifacts (solid lines) and the combination of 'Rays'+'Bright' (dashed lines).

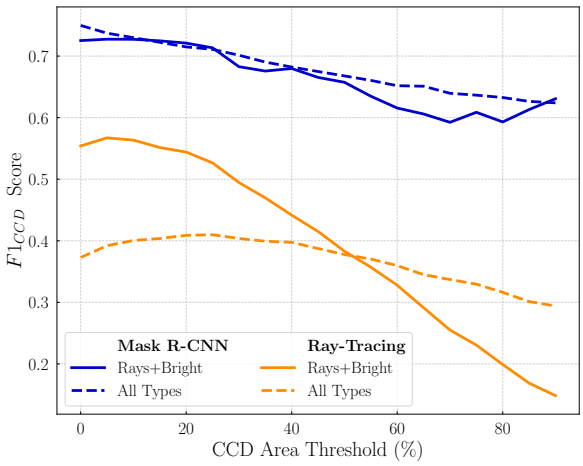



Figure 9: CCD-based $F1$ scores for the same models and ghost type combinations as in Fig. 8.

function of the area threshold in Fig. 8 (panels (a) and (b), respectively); we also present the $F1$ score in Fig. 9 for the two combinations, 'Rays'+'Bright' (solid blue lines) and 'Rays'+'Bright'+'Faint' (all ghost types, dashed blue line).

We chose this combination for two reasons: first, it allows a fairer comparison with the Ray-Tracing algorithm, which is not tuned for very low-surface-brightness ghosts (see next subsection); second, for a practical application, we may not need to reject CCDs containing very faint ghosts, because these have little influence on the surface brightness of real sources and can be effectively deblended.

### *4.3. Comparison with the Ray-Tracing algorithm*

Next, we compare the performance of our Mask R-CNN model in detecting ghost-containing CCDs to that of the Ray-Tracing algorithm. We note a few details of this comparison:

- The test dataset consists only of images known to contain at least one ghost or scattered-light artifact.
- When plotting metrics as a function of the CCD area threshold, this threshold is applied only to the ground-truth masks. This accounts for the fact that we only have predictions from the Ray-Tracing algorithm on a CCD-by-CCD basis.
- The available output from the Ray-Tracing algorithm does not distinguish between the different artifact categories. Furthermore, the Ray-Tracing algorithm applies a threshold to the predicted surface-brightness of artifacts, and thus is not optimized to detect 'Faint' ghosts. For that reason we exclude 'Faint' ghosts when evaluating metrics to compare performance between the Ray-Tracing and Mask R-CNN algorithms.

We plot the CCD-based precision and recall (Fig. 8) and $F1$ score (Fig. 9) resulting from the

Ray-Tracing algorithm (orange lines) and Mask R-CNN (blue lines), as a function of the ground truth threshold area. We consider two categories of artifacts selected based on the ground truth masks: all ghost types combined (solid lines) and the combination of 'Rays'+'Bright' ghosts (dashed line).

We first consider the limit of zero percent CCD area threshold: a single pixel of an artifact has to be in the CCD to be classified as ghost-containing. The Ray-Tracing algorithm achieves a high precision score, which, for the case when the combination of all ghost types is considered, is higher than that from the Mask R-CNN for the same case ($\sim 0.9$ vs. $\sim 0.7$). However, for the same case the recall is much lower ($\sim 0.8$ vs. $\sim 0.3$). In other words, Ray-Tracing produces results high in purity but low in completeness. When the combination of only 'Rays'+'Bright' ghosts is considered, both the precision and the recall from the *DeepGhostBusters* Mask R-CNN model are significantly higher than those from the Ray-Tracing algorithm.

Fig. 8 shows that precision decreases, while recall increases as a function of the CCD area threshold for both artifact combinations. As we increase the threshold, fewer CCDs are labeled as containing artifacts and thus the purity decreases while the completeness increases.

The $F1$ score, which combines precision and recall, demonstrates that the performance of the Mask R-CNN model is significantly higher than that of the Ray-Tracing algorithm for all area threshold values and for both artifact combinations (Fig. 9).

To facilitate the numerical comparison of the performance of the algorithms, we present in Table 1 the values of the different metrics for the two models, at a one pixel ($> 0\%$) CCD area threshold, for both algorithms. The results for both artifact category combinations ('Rays'+'Bright' and 'Rays'+'Bright'+'Faint') are presented.

### *4.4. Standard object detection evaluation metrics*

We now examine the Average Precision (AP; Everingham et al., 2010), a metric that is commonly used by the computer vision community to assess the performance of object detection algorithms. The AP is defined as the area under the Precision-Recall (PR) curve:

$$\mathrm{AP} = \int_0^1 p(r)dr, \tag{4}$$

where $p(r)$ is the precision, $p$, at recall level $r$. In practice, an 11-point interpolation method is used, and the AP score is calculated as:

$$\mathrm{AP} = \frac{1}{11} \sum_{r_i \in R} \tilde{p}(r_i), \tag{5}$$

where $\tilde{p}$ is the maximum precision at each recall bin and $R = \{0.0, 0.1, \ldots, 1.0\}$. Precision and recall are defined using the common formulae (Eqs. 1 and 2), but here the number of true positives, true negatives etc. refer to detections of individual artifacts and not single CCDs.

To define the detection of an artifact, we introduce the concept of the Intersection over Union (IoU; also known as the Jaccard index; Jaccard 1912), which quantifies the overlap between the masks of the ground truth and the prediction. As the name suggests, it is defined as the ratio of the area of the intersection of the predicted mask ($pm$) and the ground truth ($gt$) mask over the area of the union of the predicted and ground truth masks:

$$\mathrm{IoU} = \frac{\text{area of intersection}}{\text{area of union}} = \frac{area(gt \cap pm)}{area(gt \cup pm)}. \tag{6}$$

An IoU threshold is then used to determine if a predicted mask is a $TP, FP$, or $FN$. It is common to evaluate the AP score at different IoU levels, and we denote the AP at a IoU threshold $\beta$ as "AP@$\beta$".

By calculating the PR curves and the AP score at different IoU threshold and for the different artifact categories, we evaluate the performance of the Mask R-CNN model for different artifact categories. Furthermore, by determining how AP varies with increasing IoU, we evaluate the agreement between the true and predicted masks.

In Fig. 10, we present the PR curves and the corresponding AP scores for IoU thresholds in the range $0.5 - 0.9$ (with step size 0.05) for the three artifact types in panels (a)-(c), individually, and for all artifact types combined in panel (d). We find that 'Bright' ghosts are most easily detected by the Mask R-CNN, while 'Faint' ghosts are the most challenging to detect — in agreement with our expectations. Furthermore, for 'Rays', the AP decreases rapidly with increasing IoU threshold: the model struggles to accurately reproduce the ground truth masks for these artifacts. This is expected, because these artifacts do not have clear boundaries, as demonstrated by variation in the mask regions defined by the human annotators.

In that section, we have shown that the Mask R-CNN algorithm is superior to the Ray-Tracing

Table 1: CCD-based evaluation metrics (precision, recall, $F1$ score) for the Mask R-CNN and Ray-Tracing algorithms, at 0% CCD area threshold.

| **Metric** \ **Model** | Mask R-CNN | | Ray-Tracing | |
|---|---|---|---|---|
| | Rays+Bright | Rays+Bright+Faint | Rays+Bright | Rays+Bright+Faint |
| Precision | 84.3% | 68.7% | 64.7% | 89.9% |
| Recall | 63.6% | 82.5% | 48.4% | 23.5% |
| $F1$ score | 72.5 % | 75.0% | 55.4% | 37.3% |

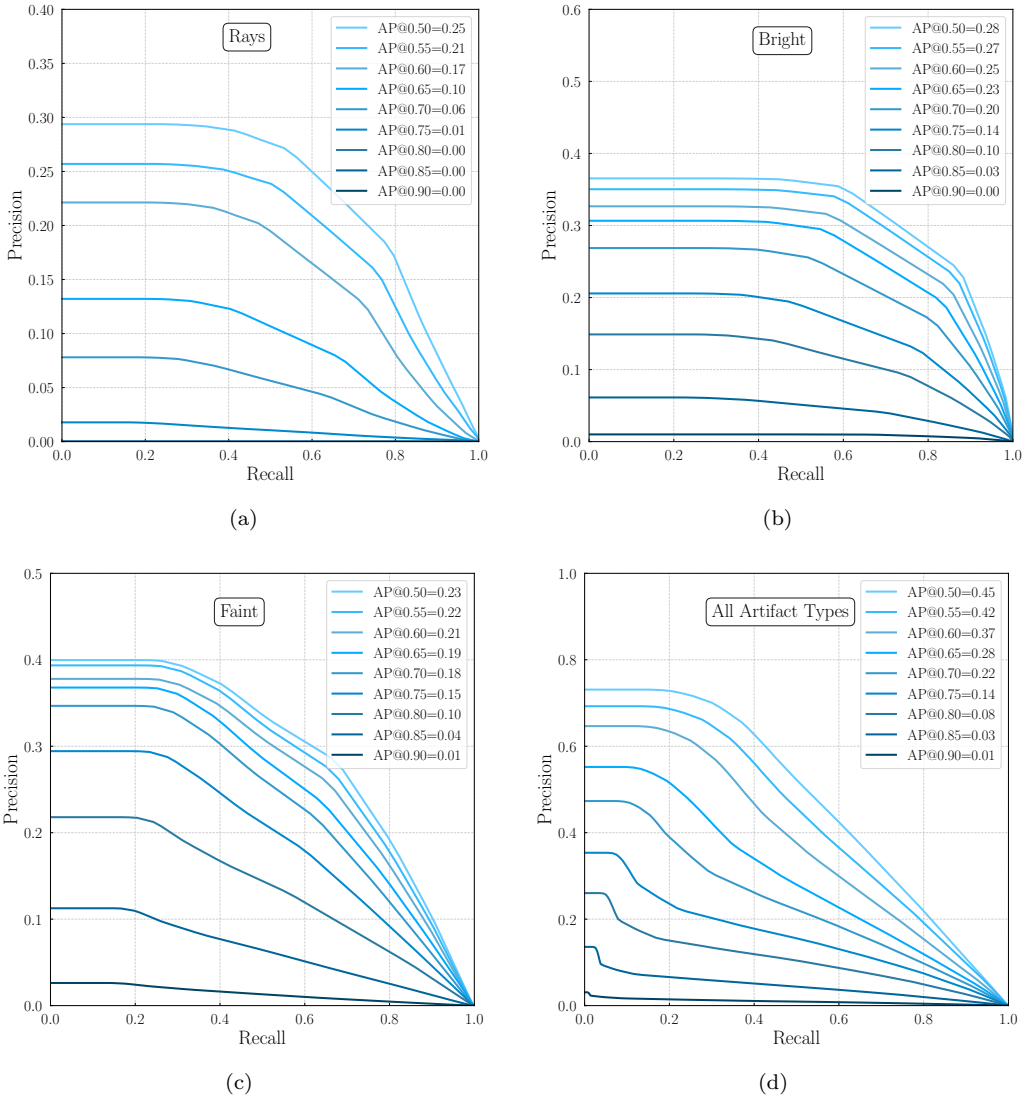



Figure 10: Precision-Recall curves and Average Precision scores at different IoU threshold values in the range $0.50 - 0.90$. We show these metrics for the different ghost types in this work ('Rays'-'Bright'-'Faint'), and for all ghost types, combined.

in detecting CCDs affected by ghosts or scattered-light artifacts.

### *4.5. Using Mask R-CNN to classify ghost-containing vs. ghost-free images*

So far, the images used for training and testing the performance of the Mask R-CNN model were known (by visual inspection) to contain at least one ghost or scattered-light artifact. However, most DECam images do not contain prominent ghost or scattered-light artifacts, and thus they systematically differ from those used to train and test the model. Such differences may result in a large number of false positive detections (e.g., real astronomical sources, especially large and bright objects) or systematically failing to detect ghosts in some images — for example, images that contain only very small or very faint ghosts.

To test the performance of the Mask R-CNN on images that do not contain ghosts, we use a set of 1792 images with an equal number of ghost-free and ghost-containing images. This set of images is independent of the 2000 images used to train, validate, and test the Mask R-CNN model. They constitute the test set used in Chang et al. (2021). For this dataset, the ground truth labels refer to the presence of a ghost in the image — not the number of ghosts or the regions affected by ghosts.

We run Mask R-CNN on this dataset: when the algorithm predicts the existence of even a single ghost or scattered-light artifact in the image, we assign a predicted label 'HAS GHOST' to that image. Otherwise the assigned predicted label 'CLEAN'. The confusion matrix resulting from this process is shown in Fig. 11. The accuracy is 79.7%, the precision is 77.3%, and the recall is 84.3%. Both the numbers of false positive and false negative cases are high: false positives occur at $\sim$ 22.7% of the total number of images classified as positives, and the false negatives occur at (Dimitrios)....

However, visual inspection of false positive examples and the predicted masks revealed that most contain objects or exhibit features similar to those found in ghost-containing images. These include bright streaks from artificial Earth-orbiting satellites (mimicking 'Rays'), low-surface-brightness emission from Galactic cirrus, images with poor data quality (due to cloud coverage that diffuses starlight), or large resolved stellar systems (e.g., dwarf galaxies and globular clusters). These are very similar to the cases of false positives returned by the CNN classifier in Chang et al. (2021). Similarly, most of the false negatives contain very small and faint ghosts (and usually each image contains only one such ghost) that could have been easily missed even by a human annotator.[12] Thus, we conclude that the false positives/negatives are qualitatively different from the true positives/negatives. and that – in practice – the Mask R-CNN is much better in classifying images that contain unusual and/or problematic areas, compared to what one would naively assume from the confusion matrix (Fig. 11).

We note that in practical applications of Mask R-CNN, we can reduce the number of false positives by first applying the CNN classifier presented in Chang et al. (2021), and then applying the Mask R-CNN only to those images that are identified as containing ghosts or scattered-light artifacts. The results of this process on the test dataset are presented in panel (b) of Fig. 11. We find that we are able to reduce the number of false positives to less that of the Mask R-CNN alone, but at the expense of increasing the number of false negatives. This combined model has an overall accuracy of 83.1%, precision of 87.3%, and recall of 75.6%. Because of this trade-off, the final decision of pre-processing with a CNN depends on the particular problem and whether we are willing to reject otherwise real astronomical objects (false positives) or to have residual ghost and scattered-light artifacts (false negatives).

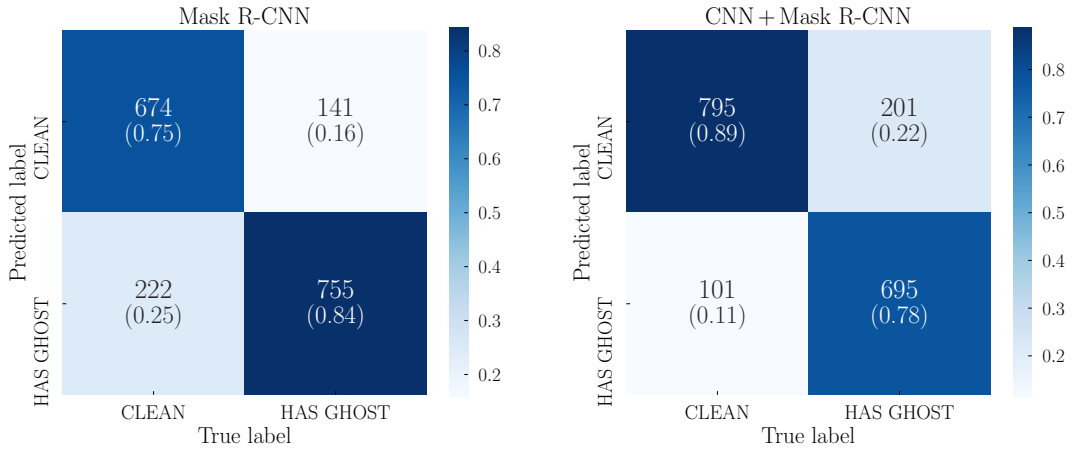
## 5. Summary and Conclusions

In this work, we applied a state-of-the art object detection and segmentation algorithm, Mask R-CNN, to the problem of finding and masking ghosts and scattered-light artifacts in astronomical images from DECam. The effective detection and mitigation of these artifacts is especially important for low-surface-brightness science, an important target of future surveys.[13] Given the sheer volume of data generated by current and upcoming surveys, automated methods must be used for the identification of these artifacts.

In this paper, we compared the performance of the Mask R-CNN algorithm to two previous approaches, each of which has benefits and limita-

[12] Examples of false positives and false negatives can be found in Appendix D.

[13] See, for example, `https://sites.google.com/view/lsstgsc/working-groups/low-surface-brightness-science`




(a) (b)

Figure 11: (a) Confusion matrix of predictions of the Mask R-CNN model on a dataset containing an even number of ghost containing and clean images. An image is predicted to 'have ghost' if even a single ghost is detected in that image by the Mask R-CNN model. (b) Confusion matrix of the predictions of the combined CNN + Mask R-CNN model (CNN model from Chang et al. (2021)). An image is said to 'have ghost' if and only if both the CNN and the Mask R-CNN models agree on that (otherwise the prediction is 'clean').

tions. First, the conventional Ray-Tracing algorithm currently used by DES identifies individual CCDs affected by ghosting or scattered-light artifacts. This is a predictive model that does not use the actual imaging data to detect artifacts. Thus, its performance is limited by the accuracy of the optical model and external catalogs of bright stars, and it fails to detect a significant number of artifacts. Second, we compared to a relatively standard CNN (Chang et al., 2021), which does not depend on modeling the optical processes that lead to the generation of artifacts or on external catalogs of bright astronomical objects. Furthermore, it separates "ghost-containing" from "clean" images with high accuracy. However, as a classifier, it does not identify the affected subregion(s) within the image: if used without further investigation, it can lead to the rejection of useful information from non–affected parts of the image.

The Mask R-CNN approach presented in this work has the benefits of a deep learning approach — i.e., it does not depend on physical modeling, except through that training data, themselves — that can predict the locations of ghosts and scattered-light artifacts, which can be used to create CCD- and pixel-level masks of the affected region of an image.

We compare the ability of Mask R-CNN in masking affected CCDs in *ghost-containing* images with that of the Ray-Tracing algorithm. We find that the Mask R-CNN model has superior performance, as measured by the $F1$ score, which is the harmonic mean of the precision (purity) and the recall (completeness). These results hold across different CCD area thresholds and for the two combinations of the morphological classes discussed in this work — 'Bright'+'Rays' and 'Bright'+ 'Rays'+'Faint'. At the threshold of one pixel ($> 0\%$), for example, and for the combination 'Rays+Bright' the $F1$ score of the Mask R-CNN model is 72.5% as opposed to 55.4% of the Ray-Tracing algorithm.

One weakness of our method is that it produces a large number of false positives when presented with images that do not contain ghosts or scattered-light artifacts — although many of these false positives contain other types of artifacts or bright astronomical objects. We show that, to mitigate this problem, a CNN classifier similar to that discussed in Chang et al. (2021) can be used as a pre-processing step before the Mask R-CNN is applied to images that are predicted to contain ghosts or scattered-light artifacts. This process reduces the number of false positives by a factor of two and increases the number of false negatives, and improves the accuracy.

The results presented here highlight the promise of object detection and segmentation methods in tackling the identification of ghosts and scattered-light artifacts. Since deep learning models that are trained on one data set can be adapted to a new

data set with many fewer examples through transfer learning, the *DeepGhostBusters* algorithm trained on DECam images can potentially be adapted and retrained to identify such artifacts in future surveys. Indeed, cross-survey transfer learning has already been shown to significantly reduce the need for large annotated datasets in deep learning-based classification cases (e.g., Domínguez Sánchez et al., 2019; Khan et al., 2019; Tanoglidis et al., 2021a). Additionally, these results indicate that such techniques are also promising for different, but related, problems, such as the the detection of artifacts from cosmic rays, satellite trails, etc. (e.g., Goldstein et al., 2015; Desai et al., 2016; Melchior et al., 2016; Zhang and Bloom, 2020; Román et al., 2020; Paillassa et al., 2020). Such automated techniques can facilitate the efficient separation of artifacts from scientifically useful data in upcoming surveys like LSST.

**Acknowledgements**

We would like to thank Colin Burke, Chihway Chang, Tom Diehl, Brenna Flaugher, and Steve Kent for useful discussions and suggestions. This paper has gone through internal review by the DES collaboration.

A. Ćiprijanović is partially supported by the High Velocity Artificial Intelligence grant as part of the Department of Energy High Energy Physics Computational HEP sessions program.

We acknowledge the Deep Skies Lab as a community of multi-domain experts and collaborators who've facilitated an environment of open discussion, idea-generation, and collaboration. This community was important for the development of this project.

This material is based upon work supported by the National Science Foundation under Grant No. AST-2006340. This work was supported by the University of Chicago and the Department of Energy under section H.44 of Department of Energy Contract No. DE-AC02-07CH11359 awarded to Fermi Research Alliance, LLC. This work was partially funded by Fermilab LDRD 2018-052.

This project used public archival data from the Dark Energy Survey (DES). Funding for the DES Projects has been provided by the U.S. Department of Energy, the U.S. National Science Foundation, the Ministry of Science and Education of Spain, the Science and Technology Facilities Council of the United Kingdom, the Higher Education Funding Council for England, the National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign, the Kavli Institute of Cosmological Physics at the University of Chicago, the Center for Cosmology and Astro-Particle Physics at the Ohio State University, the Mitchell Institute for Fundamental Physics and Astronomy at Texas A&M University, Financiadora de Estudos e Projetos, Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro, Conselho Nacional de Desenvolvimento Científico e Tecnológico and the Ministério da Ciência, Tecnologia e Inovação, the Deutsche Forschungsgemeinschaft and the Collaborating Institutions in the Dark Energy Survey.

The Collaborating Institutions are Argonne National Laboratory, the University of California at Santa Cruz, the University of Cambridge, Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas-Madrid, the University of Chicago, University College London, the DES-Brazil Consortium, the University of Edinburgh, the Eidgenössische Technische Hochschule (ETH) Zürich, Fermi National Accelerator Laboratory, the University of Illinois at Urbana-Champaign, the Institut de Ciències de l'Espai (IEEC/CSIC), the Institut de Física d'Altes Energies, Lawrence Berkeley National Laboratory, the Ludwig-Maximilians Universität München and the associated Excellence Cluster Universe, the University of Michigan, NSF's NOIRLab, the University of Nottingham, The Ohio State University, the University of Pennsylvania, the University of Portsmouth, SLAC National Accelerator Laboratory, Stanford University, the University of Sussex, Texas A&M University, and the OzDES Membership Consortium.

Based in part on observations at Cerro Tololo Inter-American Observatory at NSF's NOIRLab (NOIRLab Prop. ID 2012B-0001; PI: J. Frieman), which is managed by the Association of Universities for Research in Astronomy (AURA) under a cooperative agreement with the National Science Foundation.

The DES data management system is supported by the National Science Foundation under Grant Numbers AST-1138766 and AST-1536171. The DES participants from Spanish institutions are partially supported by MICINN under grants ESP2017-89838, PGC2018-094773, PGC2018-102021, SEV-2016-0588, SEV-2016-0597, and MDM-2015-0509, some of which include ERDF funds from the European Union. IFAE is partially

funded by the CERCA program of the Generalitat de Catalunya. Research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Program (FP7/2007-2013) including ERC grant agreements 240672, 291329, and 306478. We acknowledge support from the Brazilian Instituto Nacional de Ciência e Tecnologia (INCT) do e-Universo (CNPq grant 465376/2014-2).

This manuscript has been authored by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the U.S. Department of Energy, Office of Science, Office of High Energy Physics.

## Appendix A. Human annotator agreement

As mentioned in Sec. 2.2, human annotators do not always agree on the mask boundaries and the artifact types. A significant disagreement may affect the performance of the Mask R-CNN, so we study extent of the disagreement in more detail, which may suggest avenues for improvement of the annotation process.

All eight annotators were given a common subset of 50 images that were randomly drawn from the full dataset described in Sec. 2.1. When an annotator creates a mask for a specific artifact, they give a 'vote' to the region covered by that mask. A second annotator will create a different mask around the same object. The pixels where there is an overlap between the two masks will receive two votes in total while the non-overlapping parts only one. The same process continues for all the eight annotators. The same region may receive multiple different classifications (e.g., votes for both 'Bright' and 'Faint' ghosts).

In Fig. A.12, we present histograms of the distribution of the number of votes each pixel in the dataset received during the annotation process. We restrict it to pixels that have received at least one vote. We present the distributions for each artifact category separately in panels (a)-(c), and the case where we do not distinguish between different types in panel (d). A distribution that has a strong peak in the region of $\sim 8$ votes indicates that there is a very good agreement between the annotators.

The histogram for 'Rays' shows a strong bimodality, with many pixels receiving 8 votes , but also many pixels receive just 1–2 votes. These artifacts are distinct and bright, and hard to confuse with any one of the other two types. However, they do not have very clear boundaries, so, while annotators agree on the bulk of the pixels affected by a ghost, they do not agree on the extent/edges of the masks they create.

The histogram of votes for 'Bright' artifacts, panel (b), presents a peak at the low end (1–3 votes). This can be explained by the fact that there is significant confusion about the class of some large ghosts, which most annotators classify as 'Faint', while a few classify as 'Bright'. Since they are much larger compared to other typical 'Bright' ghosts, the distribution is dominated by the pixels belonging to these confusing artifacts.

Generally, there is a good agreement between the annotators when it comes to 'Faint' ghosts, with over 30% of the pixels having received the full eight votes. When not distinguishing between the different types of artifacts (panel (d)), we see very good agreement between the annotators in masking ghost-containing pixels, with $\sim 45\%$ of those pixels having received the maximum 8 votes, and an additional $\sim 25\%$ having received seven votes. Only $\sim 10\%$ of the pixels have received only one vote.

From the above discussion, we conclude that there is generally good agreement in the mask-creation process. Some confusion exists between 'Faint' and 'Bright' ghost types, because the distinction between the two is quite arbitrary. Some potential avenues for improvement are to consider these two categories as one, define more specific criteria for each class, or have multiple persons annotate the same images and assign each artifact to the class that receives the most votes.

## Appendix B. Training History

In Fig. 5, we presented the total loss as a function of the training epoch (training history). The total loss, $L_{\rm tot}$, is the sum of the classification, bounding box, and mask loss (see Sec. 3). We present the training histories for these losses individually in Figs. B.13, B.14, and B.15, respectively. As described in the main text, we train the model using progressively smaller learning rates for a finer tuning of the parameters. We stopped the process at 75 epochs due to overfitting thereafter.

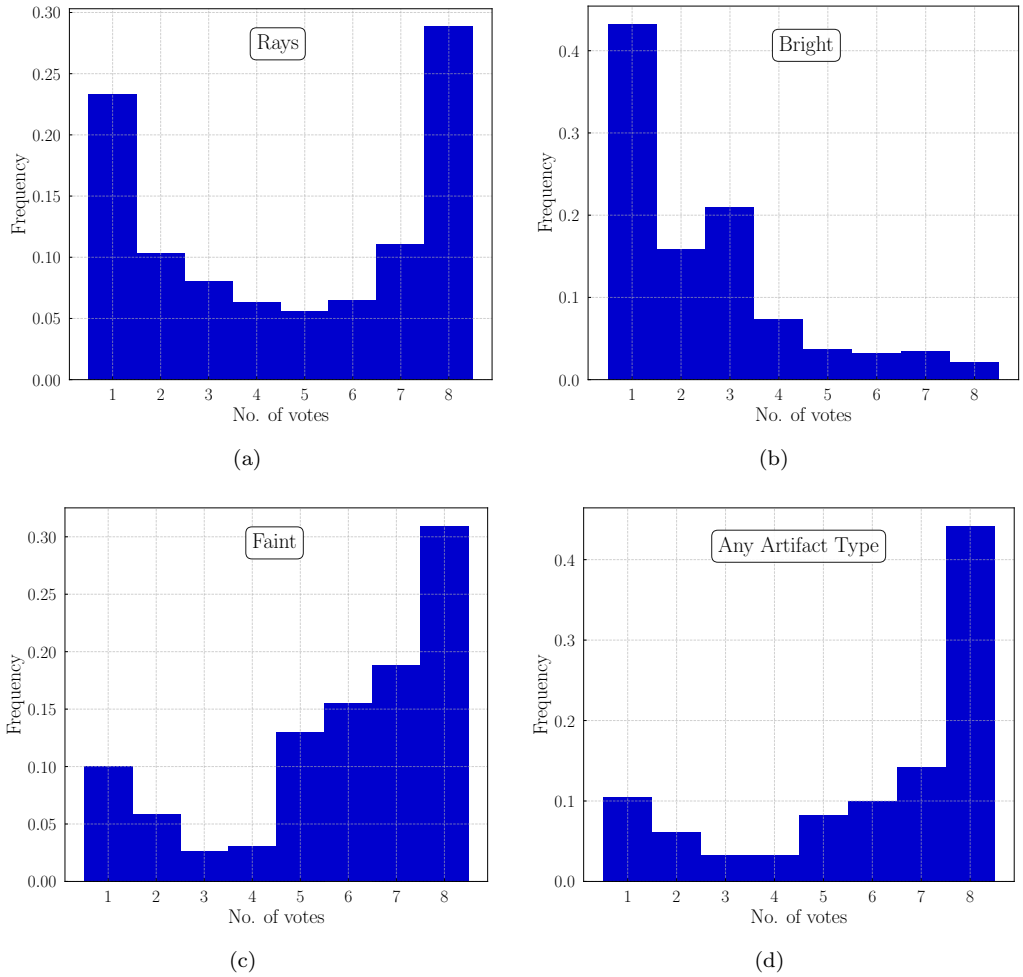



Figure A.12: Distribution of the number of votes each pixel in our dataset has received as containing a ghost, from the eight annotators. We include only pixels that have received at least one vote. We present the distributions for each ghost type separately (panels (a)-(c)) and without distinguishing between the different types (panel (d)).

## Appendix C. Masking CCDs

To help the reader better understand how the imposed area threshold affects the number of CCDs classified as ghost-containing (Sec. 4.2), in this Appendix we present the predicted artifact masks and the affected CCDs for two different threshold levels, for the same images presented in the top row of Fig. 6.

Specifically, in the panels (a) and (c) of Fig. C.16 we map (in blue) those CCDs that are classified as ghost-containing when even a single pixel of the predicted artifact mask lies within that CCD ($> 0\%$ threshold). In panels (b) and (d) we show, for the same images, the CCDs masked as ghost-containing when at least half of area of the CCD has to be covered by an artifact to be classified as such (50% threshold). To make the comparison easier, we overlay (yellow contours) the mask predictions of the Mask R-CNN model, without distinguishing between the different ghosting and scattered-light artifact types.

## Appendix D. False Positive and False Negative examples

Here we present examples of false positive and false negative classifications of ghosts and scattered-light artifacts from the Mask R-CNN method outlined in Sec. 4.5. Fig. D.17 presents examples of false positives (panel (a)) and the corresponding mask predictions of the Mask R-CNN model (panel (b)) for the same images. The color

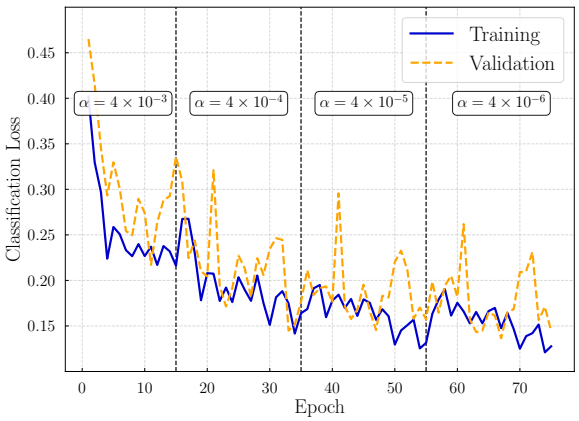



Figure B.13: Classification loss as a function of the training epoch.

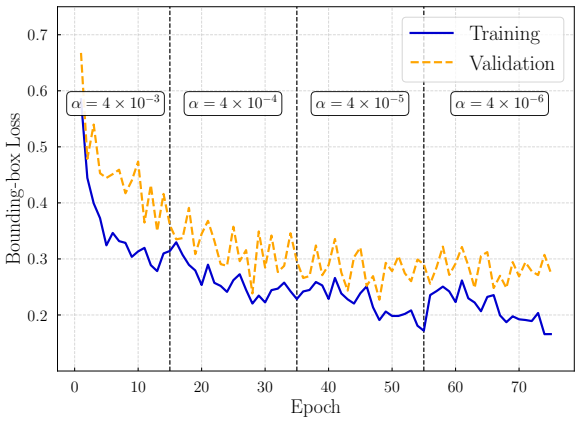



Figure B.14: Bounding box loss as a function of the training epoch.

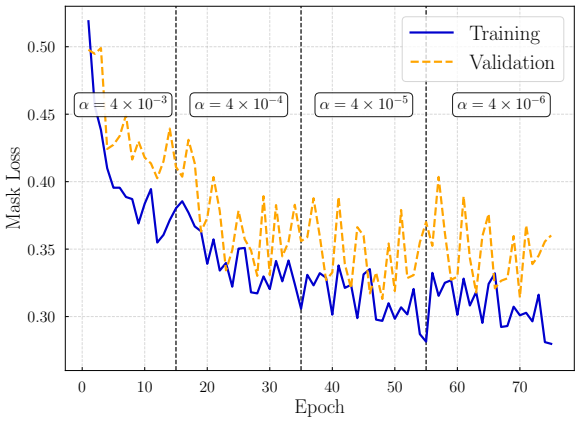



Figure B.15: Mask loss as a function of the training epoch.

scheme of the predicted masks follows that of the main text (see Fig. 6).

As discussed in the main text, Sec. 4.5, most of those images are qualitatively different from other ghost-free images and contain either other types of artifacts — for example, Earth-orbiting satellites ((2,2), (3,1)), airplane trails (1,4), structured cloud cover ((1,5), (3,2), (3,3)) or large galaxies ((2,2), (2,5)) and resolved stellar systems (4,1), where the tuplets signify rows and columns, respectively.

Fig. D.18 presents some examples of false negatives. These images contain ghosts (as confirmed by visual inspection), but they are actually very small or faint and hard to distinguish at the resolution presented here. Thus, it is not a surprise that these have been classified as "clean" by the mask R-CNN model, because they are different from the more prominent ghost-containing images that the network was trained on.

## References


Abbott, T.M.C., Annis, J., DePoy, D.L., Flaugher, B., Kent, S.M., Lin, H., Merritt, W., 2009. Dark energy camera specifications and technical requirements. URL: `https://www.noao.edu/meetings/decam/media/DECam_Technical_specifications.pdf`.

Abdulla, W., 2017. Mask r-cnn for object detection and instance segmentation on keras and tensorflow. `https://github.com/matterport/Mask_RCNN`.

Abraham, R.G., van Dokkum, P.G., 2014. Ultra-Low Surface Brightness Imaging with the Dragonfly Telephoto Array. PASP 126, 55. doi:`10.1086/674875`, arXiv:1401.5473.

Bertin, E., 2012. Displaying Digital Deep Sky Images, in: Ballester, P., Egret, D., Lorente, N.P.F. (Eds.), Astronomical Data Analysis Software and Systems XXI, p. 263.

Bom, C., Poh, J., Nord, B., Blanco-Valentin, M., Dias, L., 2019. Deep Learning in Wide-field Surveys: Fast Analysis of Strong Lenses in Ground-based Cosmic Experiments. arXiv e-prints , arXiv:1911.06341 arXiv:1911.06341.

Brough, S., Collins, C., Demarco, R., Ferguson, H.C., Galaz, G., Holwerda, B., Martinez-Lombilla, C., Mihos, C., Montes, M., 2020. The vera rubin observatory legacy survey of space and time and the low surface brightness universe. arXiv:2001.11067.

Burke, C.J., Aleo, P.D., Chen, Y.C., et al., 2019. Deblending and classifying astronomical sources with Mask R-CNN deep learning. MNRAS 490, 3952–3965. doi:`10.1093/mnras/stz2845`, arXiv:1908.02748.

Chang, C., Drlica-Wagner, A., Kent, S.M., Nord, B., Wang, D.M., Wang, M.H.L.S., 2021. A Machine Learning Approach to the Detection of Ghosting and Scattered Light Artifacts in Dark Energy Survey Images. arXiv e-prints , arXiv:2105.10524 arXiv:2105.10524.

Cheng, T.Y., Conselice, C.J., Aragón-Salamanca, A., et al., 2021. Galaxy Morphological Classification Catalogue of the Dark Energy Survey Year 3 data with Convolutional Neural Networks. arXiv e-prints , arXiv:2107.10210 arXiv:2107.10210.

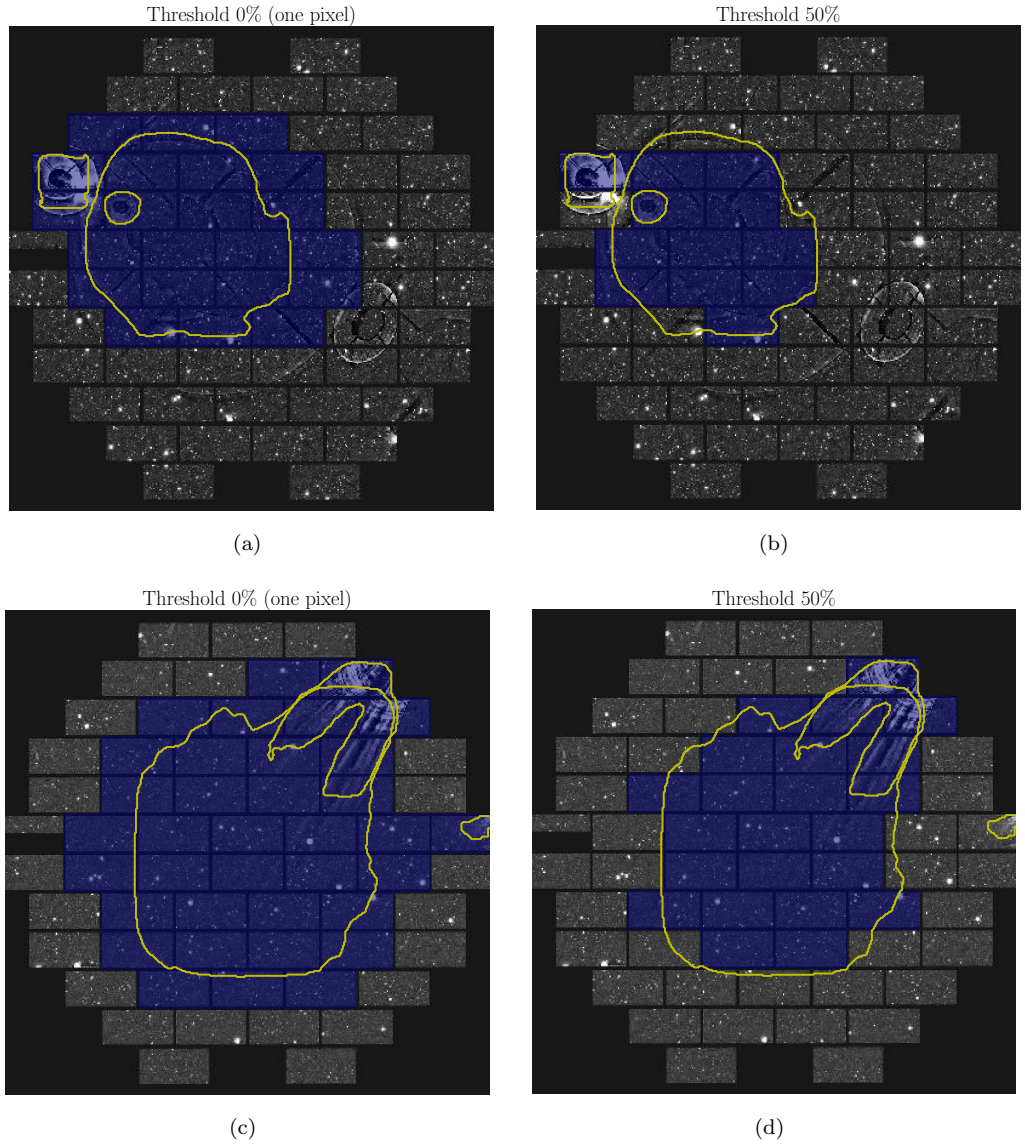



Figure C.16: CCDs masked as ghost-containing (in blue) when even a single pixel of the predicted ghost mask lies within the CCD (0% threshold, panels (a) and (c)), and when at least half of the CCD area has to be covered by the CCD (50% threshold, panels (b) and (d)). The yellow contours correspond to the mask predictions of the Mask R-CNN model (without distinguishing between the different types of artifacts).

Ćiprijanović, A., Kafkes, D., Downey, K., Jenkins, S., Perdue, G.N., Madireddy, S., Johnston, T., Snyder, G.F., Nord, B., 2021. DeepMerge II: Building Robust Deep Learning Algorithms for Merging Galaxy Identification Across Domains. arXiv e-prints , arXiv:2103.01373 arXiv:2103.01373.

Davies, A., Serjeant, S., Bromley, J.M., 2019. Using convolutional neural networks to identify gravitational lenses in astronomical images. MNRAS 487, 5263–5271. doi:10.1093/mnras/stz1288, arXiv:1905.04303.

DES Collaboration, 2005. The Dark Energy Survey. arXiv e-prints , astro–ph/0510346 arXiv:astro-ph/0510346.

DES Collaboration, 2016. The Dark Energy Survey: more than dark energy - an overview. MNRAS 460, 1270–1299. doi:10.1093/mnras/stw641, arXiv:1601.00329.

DES Collaboration, Abbott, T.M.C., Abdalla, F.B., Alarcon, A., et al., 2018. Dark Energy Survey year 1 results: Cosmological constraints from galaxy clustering and weak lensing. Phys. Rev. D 98, 043526. doi:10.1103/PhysRevD.98.043526, arXiv:1708.01530.

DES Collaboration, Abbott, T.M.C., Adamow, M., Aguena, M., et al., 2021a. The Dark Energy Survey Data Release 2. arXiv e-prints , arXiv:2101.05765 arXiv:2101.05765.

DES Collaboration, Abbott, T.M.C., Aguena, M., Alarcon, A., et al., 2021b. Dark Energy Survey Year 3 Results: Cosmological Constraints from Galaxy Clustering

and Weak Lensing. arXiv e-prints , arXiv:2105.13549 arXiv:2105.13549.

Desai, S., Mohr, J.J., Bertin, E., Kümmel, M., Wetzstein, M., 2016. Detection and removal of artifacts in astronomical images. Astronomy and Computing 16, 67–78. doi:10.1016/j.ascom.2016.04.002, arXiv:1601.07182.

Dieleman, S., Willett, K.W., Dambre, J., 2015. Rotation-invariant convolutional neural networks for galaxy morphology prediction. MNRAS 450, 1441–1459. doi:10.1093/mnras/stv632, arXiv:1503.07077.

Domínguez Sánchez, H., Huertas-Company, M., Bernardi, M., et al., 2019. Transfer learning for galaxy morphology from one survey to another. MNRAS 484, 93–100. doi:10.1093/mnras/sty3497, arXiv:1807.00807.

Dutta, A., Zisserman, A., 2019. The VIA annotation software for images, audio and video, in: Proceedings of the 27th ACM International Conference on Multimedia, ACM, New York, NY, USA. URL: https://doi.org/10.1145/3343031.3350535, doi:10.1145/3343031.3350535.

Everingham, M., Gool, L.V., Williams, C.K.I., Winn, J., Zisserman, A., 2010. The pascal visual object classes (voc) challenge.

Flaugher, B., Diehl, H.T., Honscheid, K., et al., 2015. The Dark Energy Camera. AJ 150, 150. doi:10.1088/0004-6256/150/5/150, arXiv:1504.02900.

Gkioxari, G., Malik, J., Johnson, J., 2019. Mesh R-CNN. arXiv e-prints , arXiv:1906.02739 arXiv:1906.02739.

Goldstein, D.A., D'Andrea, C.B., Fischer, J.A., et al., 2015. Automated Transient Identification in the Dark Energy Survey. AJ 150, 82. doi:10.1088/0004-6256/150/3/82, arXiv:1504.02936.

Greco, J.P., Greene, J.E., Strauss, M.A., et al., 2018. Illuminating Low Surface Brightness Galaxies with the Hyper Suprime-Cam Survey. ApJ 857, 104. doi:10.3847/1538-4357/aab842, arXiv:1709.04474.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask R-CNN. arXiv e-prints , arXiv:1703.06870 arXiv:1703.06870.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep Residual Learning for Image Recognition. arXiv e-prints , arXiv:1512.03385 arXiv:1512.03385.

Huang, X., Storfer, C., Gu, A., et al., 2021. Discovering New Strong Gravitational Lenses in the DESI Legacy Imaging Surveys. ApJ 909, 27. doi:10.3847/1538-4357/abd62b, arXiv:2005.04730.

Huang, X., Storfer, C., Ravi, V., Pilon, A., Domingo, M., Schlegel, D.J., Bailey, S., Dey, A., Gupta, R.R., Herrera, D., Juneau, S., Landriau, M., Lang, D., Meisner, A., Moustakas, J., Myers, A.D., Schlafly, E.F., Valdes, F., Weaver, B.A., Yang, J., Yèche, C., 2020. Finding Strong Gravitational Lenses in the DESI DECam Legacy Survey. ApJ 894, 78. doi:10.3847/1538-4357/ab7ffb, arXiv:1906.00970.

Ivezić, Ž., Kahn, S.M., Tyson, J.A., et al., 2019. LSST: From Science Drivers to Reference Design and Anticipated Data Products. ApJ 873, 111. doi:10.3847/1538-4357/ab042c, arXiv:0805.2366.

Jaccard, P., 1912. The distribution of the flora in the alpine zone.1. New Phytologist 11, 37–50. URL: https://nph.onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-8137.1912.tb05611.x, doi:https://doi.org/10.1111/j.1469-8137.1912.tb05611.x, arXiv:https://nph.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-8137.1912.tb05611.x.

Kaviraj, S., 2020. The low-surface-brightness Universe: a new frontier in the study of galaxy evolution. arXiv e-prints , arXiv:2001.01728 arXiv:2001.01728.

Kent, S.M. (DES), 2013. Ghost Images in DECam. doi:10.2172/1690257. FERMILAB-SLIDES-20-114-SCD.

Khan, A., Huerta, E., Wang, S., Gruendl, R., Jennings, E., Zheng, H., 2019. Deep learning at scale for the construction of galaxy catalogs in the dark energy survey. Physics Letters B 795, 248–258. URL: https://www.sciencedirect.com/science/article/pii/S0370269319303879, doi:https://doi.org/10.1016/j.physletb.2019.06.009.

Kim, E.J., Brunner, R.J., 2017. Star-galaxy classification using deep convolutional neural networks. MNRAS 464, 4463–4475. doi:10.1093/mnras/stw2672, arXiv:1608.04369.

Lanusse, F., Ma, Q., Li, N., et al., 2018. CMU DeepLens: deep learning for automatic image-based galaxy-galaxy strong lens finding. MNRAS 473, 3895–3906. doi:10.1093/mnras/stx1665, arXiv:1703.02642.

Lecun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE 86, 2278–2324. doi:10.1109/5.726791.

Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2016. Feature Pyramid Networks for Object Detection. arXiv e-prints , arXiv:1612.03144 arXiv:1612.03144.

Lin, T.Y., Maire, M., Belongie, S., et al., 2014. Microsoft COCO: Common Objects in Context. arXiv e-prints , arXiv:1405.0312 arXiv:1405.0312.

LSST Science Collaboration, 2009. LSST Science Book, Version 2.0. arXiv e-prints , arXiv:0912.0201 arXiv:0912.0201.

Melchior, P., Sheldon, E., Drlica-Wagner, A., et al., 2016. Crowdsourcing quality control for Dark Energy Survey images. Astronomy and Computing 16, 99–108. doi:10.1016/j.ascom.2016.04.003, arXiv:1511.03391.

Minaee, S., Boykov, Y., Porikli, F., et al., 2020. Image Segmentation Using Deep Learning: A Survey. arXiv e-prints , arXiv:2001.05566 arXiv:2001.05566.

Mueed Hafiz, A., Mohiuddin Bhat, G., 2020. A Survey on Instance Segmentation: State of the art. arXiv e-prints , arXiv:2007.00047 arXiv:2007.00047.

Paillassa, M., Bertin, E., Bouy, H., 2020. MAXIMASK and MAXITRACK: Two new tools for identifying contaminants in astronomical images using convolutional neural networks. A&A 634, A48. doi:10.1051/0004-6361/201936345, arXiv:1907.08298.

Román, J., Trujillo, I., Montes, M., 2020. Galactic cirri in deep optical imaging. A&A 644, A42. doi:10.1051/0004-6361/201936111, arXiv:1907.00978.

Sevilla-Noarbe, I., Bechtol, K., Carrasco Kind, M., et al., 2021. Dark Energy Survey Year 3 Results: Photometric Data Set for Cosmology. ApJS 254, 24. doi:10.3847/1538-4365/abeb66, arXiv:2011.03407.

Shorten, C., Khoshgoftaar, T., 2019. A survey on image data augmentation for deep learning. Journal of Big Data 6, 1–48.

Simonyan, K., Zisserman, A., 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv e-prints , arXiv:1409.1556 arXiv:1409.1556.

Tanoglidis, D., Ćiprijanović, A., Drlica-Wagner, A., 2021a. DeepShadows: Separating low surface brightness galaxies from artifacts using deep learning. Astronomy and Computing 35, 100469. doi:10.1016/j.ascom.2021.100469,

arXiv:2011.12437.

Tanoglidis, D., Drlica-Wagner, A., Wei, K., et al., 2021b. Shadows in the Dark: Low-surface-brightness Galaxies Discovered in the Dark Energy Survey. ApJS 252, 18. doi:10.3847/1538-4365/abca89, arXiv:2006.04294.

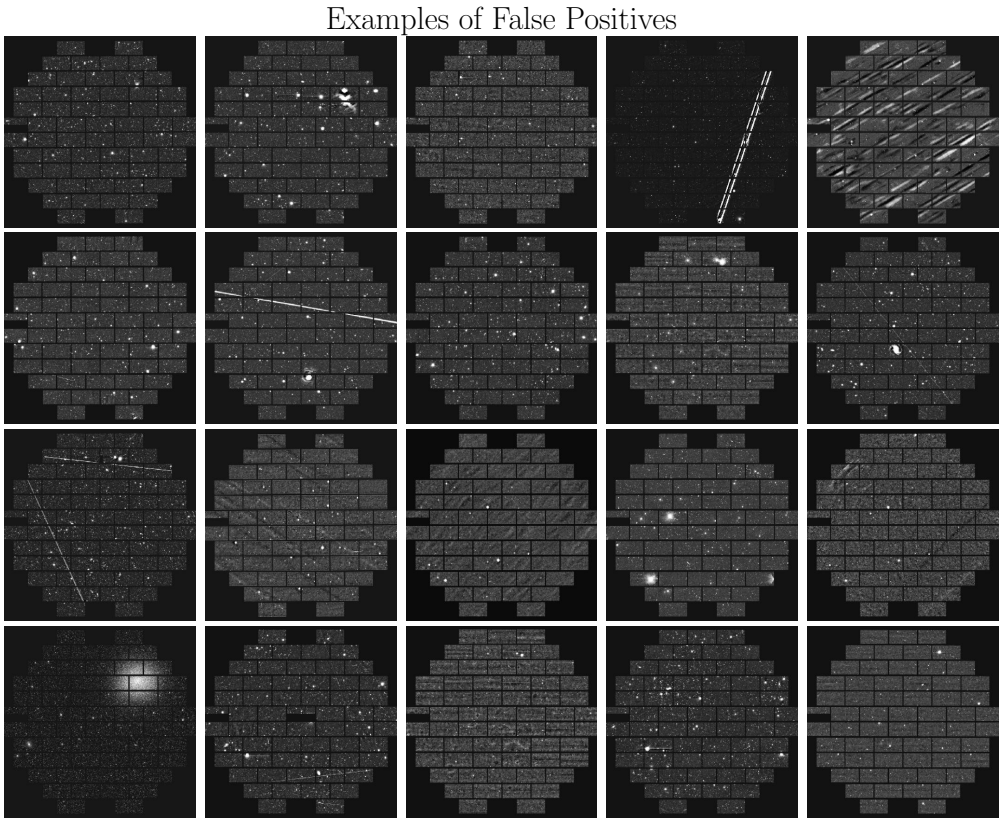
Wang, M., Deng, W., 2018. Deep Visual Domain Adaptation: A Survey. arXiv e-prints , arXiv:1802.03601 arXiv:1802.03601.

Weng, L., 2017. Object detection for dummies part 3: R-cnn family. lilianweng.github.io/lil-log URL: http://lilianweng.github.io/lil-log/2017/12/31/object-recognition-for-dummies-part-3.html.

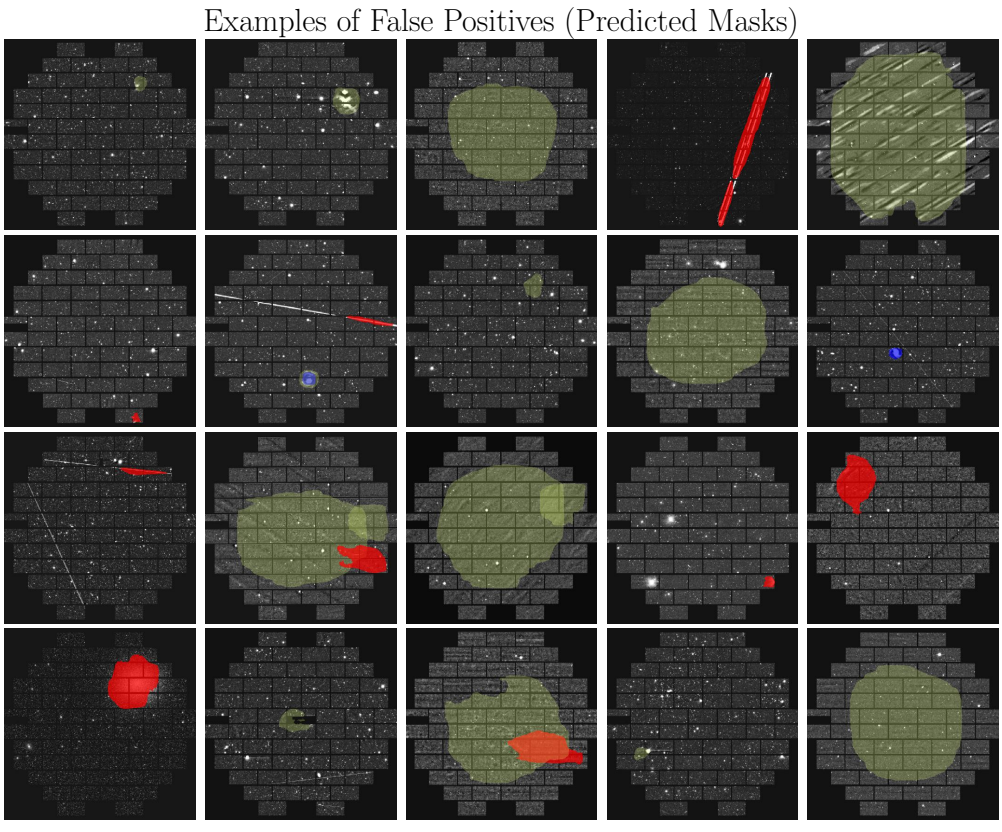
Zhang, K., Bloom, J.S., 2020. deepCR: Cosmic Ray Rejection with Deep Learning. ApJ 889, 24. doi:10.3847/1538-4357/ab3fa6, arXiv:1907.09500.

Zhao, Z.Q., Zheng, P., Xu, S.t., Wu, X., 2018. Object Detection with Deep Learning: A Review. arXiv e-prints , arXiv:1807.05511 arXiv:1807.05511.

Zhuang, F., Qi, Z., Duan, K., et al., 2019. A Comprehensive Survey on Transfer Learning. arXiv e-prints , arXiv:1911.02685 arXiv:1911.02685.




(a)




(b)

Figure D.17: (a) Example images classified as ghost-containing (false positives) and the corresponding predicted masks (lower panel, (b)).

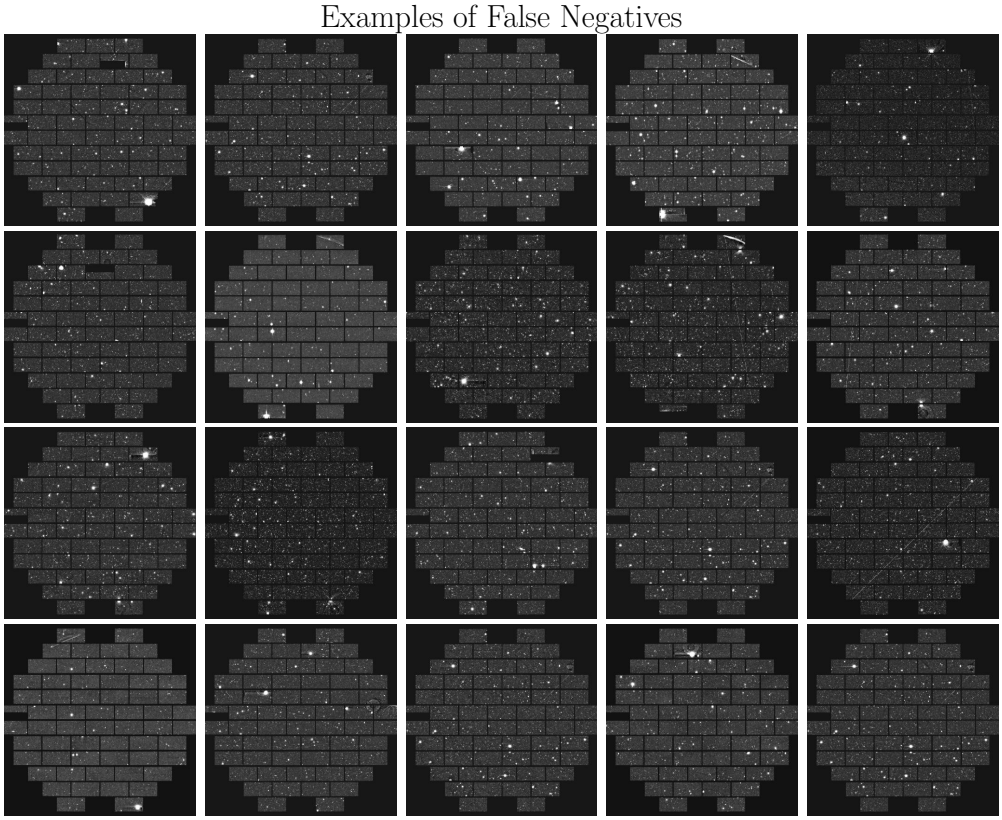



Figure D.18: Examples of false negatives, i.e. images that were classified as 'clean' by the Mask R-CNN model (no objects detected). In practice, the artifacts present in these images are very small and faint, and often go undetected by human annotators.