Draft version June 23, 2021
Typeset using LaTeX **twocolumn** style in AASTeX63

FERMILAB-PUB-21-268-AE
DES-2021-0638

Expediting DECam Multimessenger Counterpart Searches with Convolutional Neural Networks

A. Shandonay,[1] R. Morgan,[1,2] K. Bechtol,[1,3] C. R. Bom,[4,5] B. Nord,[6,7] A. Garcia,[8] B. Henghes,[9] K. Herner,[6] M. Tabbutt,[1] A. Palmese,[6,7] L. Santana-Silva,[10] M. Soares-Santos,[11] M. S. S. Gill,[12] and J. García-Bellido[13]

[1] *Physics Department, University of Wisconsin-Madison, 1150 University Avenue Madison, WI 53706, USA*
[2] *Legacy Survey of Space and Time Corporation Data Science Fellowship Program, USA*
[3] *Legacy Survey of Space and Time, 933 North Cherry Avenue, Tucson, AZ 85721, USA*
[4] *Centro Brasileiro de Pesquisas Físicas, Rua Dr. Xavier Sigaud 150, CEP 22290-180, Rio de Janeiro, RJ, Brazil*
[5] *Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, Rodovia Mário Covas, Cep 23810-000, Itaguaí, RJ, Brazil*
[6] *Fermi National Accelerator Laboratory, P. O. Box 500, Batavia, IL 60510, USA*
[7] *Kavli Institute for Cosmological Physics, University of Chicago, Chicago, IL 60637, USA*
[8] *Department of Astronomy, University of Michigan, Ann Arbor, MI 48109, USA*
[9] *Department of Physics & Astronomy, University College London, Gower Street, London, WC1 E 6BT, UK*
[10] *NAT-Universidade Cruzeiro do Sul / Universidade Cidade de São Paulo, Rua Galvão Bueno, 868, 01506-000, São Paulo, SP, Brazil*
[11] *Department of Physics, University of Michigan, Ann Arbor, MI 48109, USA*
[12] *SLAC National Accelerator Laboratory, Menlo Park, CA 94025, USA*
[13] *Instituto de Fisica Teorica UAM/CSIC, Universidad Autonoma de Madrid, 28049 Madrid, Spain*

ABSTRACT

Searches for counterparts to multimessenger events with optical imagers use difference imaging to detect new transient sources. However, even with existing artifact detection algorithms, this process simultaneously returns several classes of false positives: false detections from poor quality image subtractions, false detections from low signal-to-noise images, and detections of pre-existing variable sources. Currently, human visual inspection to remove remaining false positives is a central part of multimessenger follow-up observations, but when next generation gravitational wave and neutrino detectors come online and increase the rate of multimessenger events, the visual inspection process will be prohibitively expensive. We approach this problem with two convolutional neural networks operating on the difference imaging outputs. The first network focuses on removing false detections and demonstrates an accuracy above 95%. The second network focuses on sorting all real detections by the probability of being a transient source within a host galaxy and distinguishes between various classes of images that previously required additional human inspection. We find the number of images requiring human inspection will decrease by a factor of 1.5 using our approach alone and a factor of 3.6 using our approach in combination with existing algorithms, facilitating rapid multimessenger counterpart identification by the astronomical community.

Keywords: Optical astronomy – Machine learning – Transient sources

1. INTRODUCTION

Multimessenger astronomy utilizes the coordinated efforts of two or more types of detectors including electromagnetic, gravitational wave, and neutrino to gain increased understanding of astrophysical phenomena. Gravitational wave (GW) events detected by the Laser Interferometer Gravitational-Wave Observatory (LIGO) (Aasi et al. 2015) and Virgo (Acernese et al. 2014) or high-energy neutrinos detected by IceCube (Achterberg et al. 2006) and ANTARES (Ageron et al. 2011) may have electromagnetic counterparts that could offer insights for several fields of physics. Counterparts emitting in optical wavelengths have been and will continue to be detected using the Dark Energy Camera (DECam Flaugher et al. 2015) with difference imaging (Kessler et al. 2015). This technique compares a recent (referred to as "search") image of the area in the sky associated with a gravitational wave or neutrino to a previous (referred to as "template") image of the same area taken at earlier times. After matching the point spread functions (PSFs) to the search and template images, the result-

Corresponding author: Adam Shandonay and Robert Morgan
ashandonay@wisc.edu

robert.morgan@wisc.edu

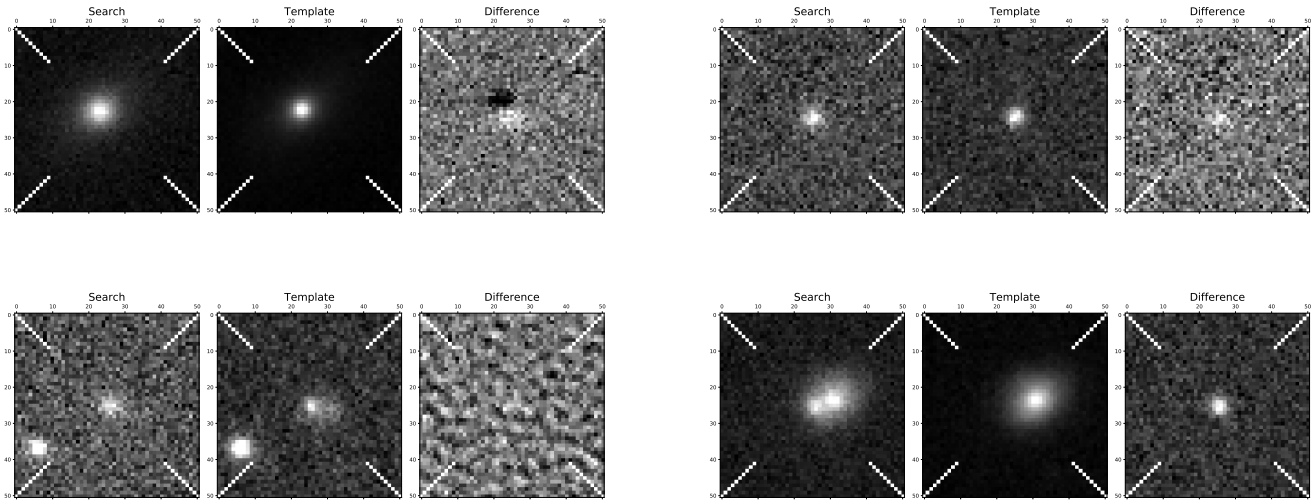

Figure 1. The upper left panel shows an example of a "bad subtraction" image. The dipole-like structure present in the difference image is an indicative feature of this type of artifact. The upper right panel shows an example of a "pre-existing point source" image. The lower left panel shows an example of a "no obvious transient" image. The lower right panel shows an example of a "transient + host" image where a smaller, point-like object is visible within a larger structure representing the host galaxy.

ing pixel-by-pixel subtraction of the search and template images is called a "difference image". The Dark Energy Survey Gravitational Wave (DESGW) team has developed a pipeline to efficiently perform difference imaging on large areas of sky (Herner et al. 2020).

When done correctly, the resulting difference image will contain transient objects that could potentially be the source of GWs or neutrinos. In practice, the rate of difference image detections is two orders of magnitude higher than the expected number of real transients (e.g. Morgan et al. 2019, 2020d). Many false detections are caused by moving objects such as asteroids and satellites that can be ruled out easily using multiple observations, while other detections that cannot be eliminated simply are known as "difference imaging artifacts". The most common example of an artifact is known as a "bad subtraction," where a slight misalignment or inaccurate determination of the point spread function (PSF) between the search and template images creates adjacent under-subtracted and over-subtracted regions in the difference image, the first of which can be interpreted as a real object by Source Extractor (Bertin, E. & Arnouts, S. 1996).

In the current DESGW pipeline, these bad subtractions are identified using a selection routine called `autoscan` (Goldstein et al. 2015) that assigns a score between 0 and 1 to each difference image with higher values corresponding to higher-quality detections. This score is determined by a random forest weighting of hand-engineered features. Bad subtractions scored below a threshold of 0.7 are generally ruled out, but a significant number of bad subtraction images are given scores above the threshold. Another type of false positive detection in counterpart searches is a pre-existing point source which is an object that was already visible in the template image, but produced a difference image due to changing brightness (e.g. variable Milky Way stars or astrophysical transients in the template image). Because these images contain real, astrophysical objects, they are often given high scores by artifact detection algorithms like `autoscan` even though they are of no interest in multimessenger astronomy.

The other common false positive that is not eliminated with `autoscan` is the marginal case of no obvious transient in the difference image. These are images which seem to contain a host galaxy and a new object may appear in the search image, however, the resulting difference image is inconclusive. In most cases, this class contains galaxies with small variability in their centers rather than supernovae or kilonovae producing an obvious transient. There are other less common types of false positives. For example, an asteroid detected in a previously empty patch of sky will appear to be a point-like source in the difference image and receive high `autoscan` scores because the lack of the presence of a host galaxy in the template image does not affect the scoring. There are also cases where realizations of Poisson noise produce groups of pixels that could resemble an object in the search image, and in some cases groups of under-fluctuations in the template image which create the appearance of an object after subtraction. We refer to this broad class as "noisy template". The true

Real Data Collection per Class			
Class	ArtifactSpy	GW200224	GW190814
Transient + Host Galaxy	388	214	38
No Obvious Transient	921	0	0
Bad Subtraction	9436	0	0
Other Artifact	731	0	0
Pre-existing Point Source	1050	0	0

Table 1. The distribution of counts for each class that we collected from ArtifactSpy and the supplemented counts of transient + host images from two follow-ups. These data (along with some simulations) were used for the algorithm development and CNN training.

positive case for multimessenger counterpart searches is when a distinguishable transient is visible only in the search image. The transient should exist within some host galaxy which will be present on both search and template images.

In a typical counterpart search with DECam, difference imaging produces $\sim 10,000$ detections per field of view, most of which are artifacts, but the size of the dataset prohibits excluding these artifacts by visual inspection. This problem could be remedied using machine learning, and specifically Convolutional Neural Networks (CNNs). In this analysis, a pipeline of CNNs and image processing routines are used to algorithmically remove all classes of false positives from consideration for being a transient with a host galaxy. The resulting output of our algorithm on a group of difference images is a sorted list of values associated with each image from 0 to 1 that represent the probability of being a transient + host galaxy. Additionally, we calculate a weighted combination of our method's probability and the `autoscan` score of the image. The output results can be selected based on a chosen probability threshold with high purity and completeness of the transient + host class.

2. METHODS

2.1. *Algorithm Summary*

The various classes of difference imaging detections present challenges for classification that demand a robust processing algorithm. Convolutional neural networks excel at image recognition (LeCun et al. 2015), making them suitable candidates for improving multimessenger counterpart searches. However, we can simplify the classification process by applying high fidelity selection criteria to images before they are passed to a CNN. We apply a series of preprocessing filters to a set of difference images which removes especially noisy images to facilitate classification and images containing objects that are already cataloged point sources are removed as well.

The first preprocessing step subtracts the median value of the image, corresponding to the background, from each pixel of the image and weights the result by a Gaussian realization of the PSF from each image to measure the flux. We remove the images with a PSF flux below a threshold. Next, the remaining images go through a second preprocessing step which calculates the signal to noise ratio (SNR) and removes images with SNR below a threshold. Both of these thresholds were determined by requiring a high completeness and purity of all classes not labeled other artifacts. The choice of these threshold cuts are discussed in section 3.1. Both of these steps aim to eliminate most of the so-called "noisy template" artifacts.

The third preprocessing step is primarily for removing pre-existing point sources that have already been cataloged. In multimessenger follow-up searchers, it is straightforward to match detections to a stellar catalog and filter out any associations, so we take a similar approach with the objects shown to the CNN. Any candidate objects with less than one arcsecond of separation from an already cataloged star are removed to mimic the selection criteria applied in real DECam follow-up observations. Finally, we remove any images that contain masking (from a nearby bright object) over the center of the difference image, since the flux would be measured inaccurately. The remaining dataset consists of low noise images that are not associated with cataloged objects.

The vast majority of the remaining dataset are artifacts that are more difficult to remove with simple filters. The most significant class of artifacts is the bad subtractions which are visually distinct from the other classes. A CNN is trained with the goal of exclusively identifying bad subtractions compared to the other classes. Singling out the bad subtractions ensures the highest level of accuracy when removing these artifacts from the dataset, as opposed to a multi-class classification scheme. The images classified as "Not Bad Subtraction" pass the first CNN and move along to another CNN that scores images with a probability of being a real transient + host galaxy. The probabilities from the second CNN are selected above a threshold with high purity and completeness of the transient + host class. We use these probabilities from our method and the `autoscan` scores to train a perceptron that applies weights to both metrics and

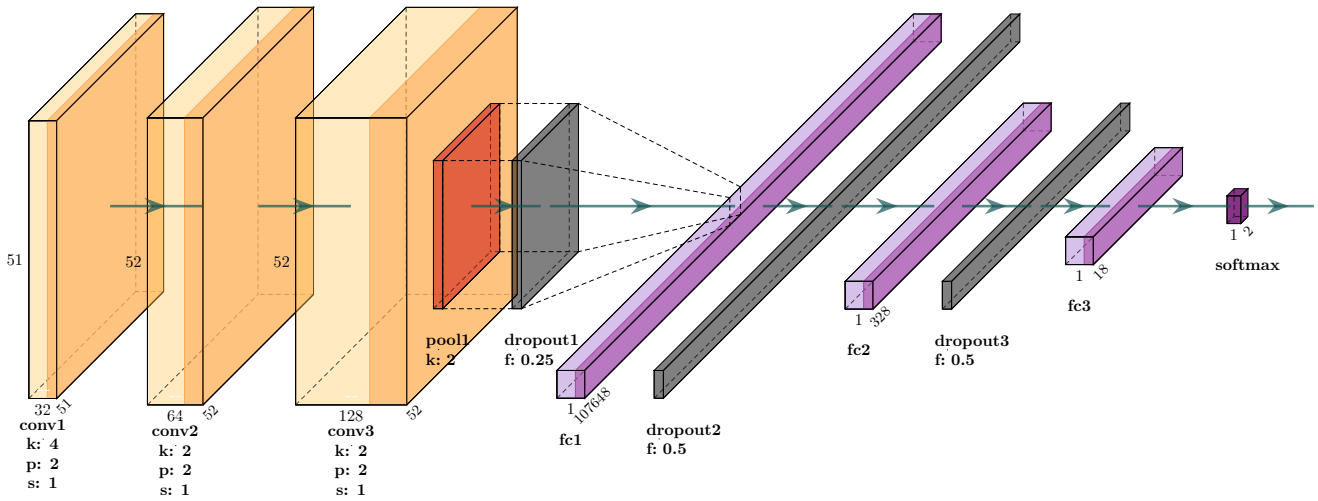

Figure 2. The architecture of the neural networks utilized in our image classification algorithm, which extracts features using three convolutional layers (orange blocks) and classifies the images using fully-connected layers (purple) that weight and aggregate the extracted features. The height and depth of the blocks correspond to the image dimensions while the width corresponds to the number of channels (convolutional operators) applied. The shading at the right edge of the blocks indicates a ReLU activation function. The various letters mean the following: k = kernel size, p = padding, s = stride, f = dropout probability. The figure was produced using `PlotNeuralNet` (Iqbal 2018).

produces a combined score. This combination method can also be used to filter images with high probability of being a transient + host class.

2.2. *Data Collection*

The algorithm development process required a large, diverse set of real images with accurate labels. We utilized over 12,000 images of randomly sampled objects detected from applying the DESGW Search and Discovery Pipeline to DES wide-field data (Herner 2019) for our training data. A team of six experts labeled the images corresponding to the five image types given in Table 1 using an interactive tool (`ArtifactSpy`; Morgan 2020) that cycled the images across the team to ensure precise labeling.

During the labeling process, it became clear that bad subtractions dominated the five image classes. To boost representation of the transient + host galaxy class, we obtained additional difference imaging data from three sources. First, we supplemented the dataset with a population of the DES wide-field difference imaging data that was given an `autoscan` score of at least 0.9. Second, we incorporated transient + host objects identified by human inspection during the DESGW follow-up observations of GW190814 (Morgan et al. 2020d) and GW200224 (Morgan et al. 2020a,b). The counts from each of these datasets are also given in table 1. Lastly, we simulated transient + host images using `deeplenstronomy` (Morgan et al. 2021).

We also obtained additional real follow-up observation data which we only used for testing the entire algorithm in Section 3.5. Specifically, difference imaging samples from DECam follow-up observations of neutrino counterpart searches IC171106A (Morgan et al. 2019), IC190331A, and IC201114A (Morgan et al. 2020c) as well as GW counterpart searches GW190728 (Soares-Santos et al. 2019) and GW190814. These counterpart searches represent a variety of observing conditions and optical bandpasses. We selected random samples of the difference images from these observations such that the size of the sample would produce a 68% confidence level sampling error equal to 1% of the entire population, indicating the sample size was large enough samples to be considered representative. These datasets were kept separate from each other to assess the performance of our algorithm on standalone follow-up observations.

2.3. *Network Design*

Convolutional Neural Networks are a particular type of Artificial Neural Network that convolve a kernel of hidden weights with the input features and produce feature maps that are used to classify images. This kind of approach is particularly well-suited for computer vision tasks performing as the state-of-art in image classification because they optimize their feature maps through automated learning and back propagation of errors. As a result, the input dataset can be diverse and still classified with high accuracy without the need for excessive amounts of training data.

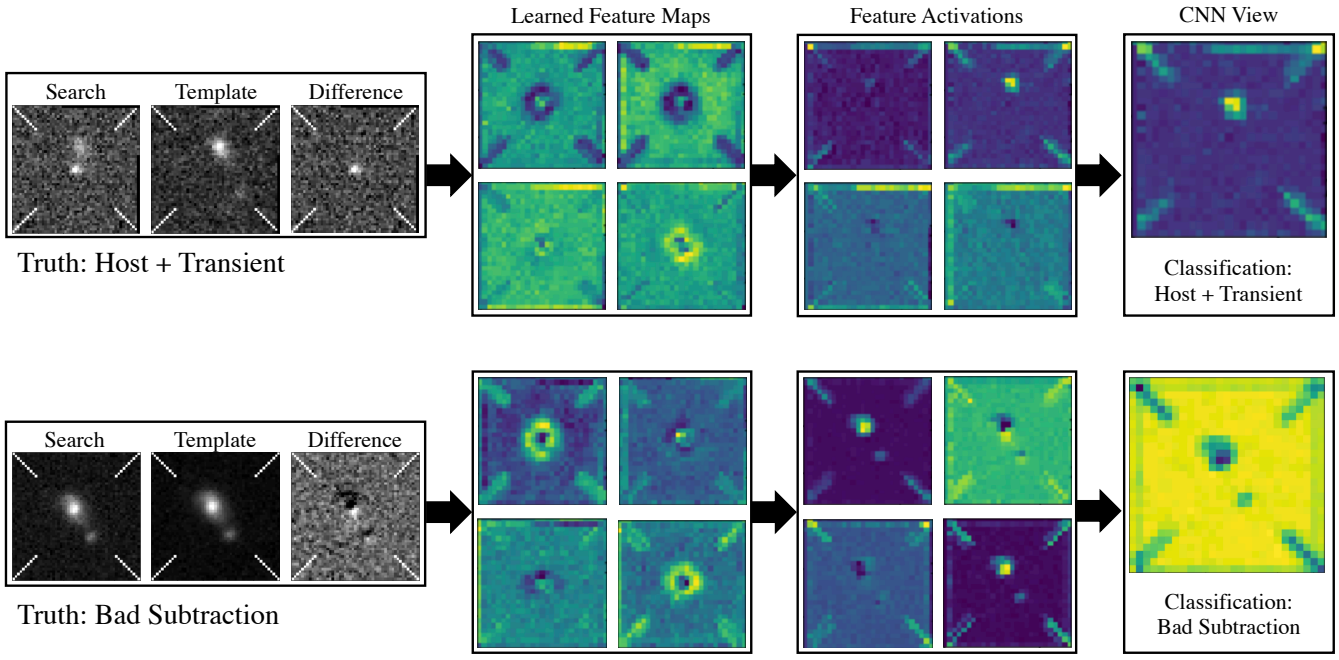

Figure 3. A visualization of how sets of difference images are interpreted by our networks. After training, the CNNs have learned what to look for in the images to make a classification. This information is stored in the gradients of the final convolutional layer of the network and is shown by the "Learned Feature Maps" column. After convolving the learned feature map with the image, the output is passed through a ReLU activation function to produce the class activation map shown as the "CNN View" column. The learned feature maps demonstrate the network is looking for point-like objects in the center of the images, elongated and off-center objects, and dipole-like regions. The bright spots in the "CNN View" of the host + transient image and inversely the dark spots for the bad subtraction indicate where those features are located.

The convolutional neural networks in the algorithm were developed using the `PyTorch` library (Paszke et al. 2019). Both networks have the same structure with three convolutional layers, a max pooling layer, three dropout layers, and three fully-connected layers. The network architecture has two main components: a feature extraction step where the convolutional layers locate edges and shapes within the images, and a classification step where the fully-connected layers weight the extracted features and reduce them to classifications. Figure 2 illustrates the flow of information through the layers of the network.

The number of `in_channels` in the first convolutional layer corresponds to the three images (search, template, and difference) contained in each difference imaging object detection. A kernel size of four pixels is used as the optimal value, as discussed in Section 2.4. The number of `out_channels` is set to 32 such that all of the features will be passed to the next layer. The convolutional layer is followed by a Rectified Linear Unit (ReLU) activation function (Glorot et al. 2011) that maps the values within each array of the `out_channels` to themselves when values are greater than zero and to zero otherwise. After the ReLU function, the 32 channels are input to the second convolutional layer, which applies a kernel of size two pixels and outputs 64 channels. Another ReLU activation function is applied, and the third layer uses a kernel size of two pixels and doubles the number of channels to 128. In each convolutional layer, the padding is set to two pixels so that the kernels can sample the edges of each image. The stride is set to one pixel so that each kernel moves over all pixels with the same weights, and small details are treated equally. After a third ReLU activation function is applied to the 128 `out_channels`, a max-pooling layer of kernel size two pixels reduces the size of the images by using the highest weights with each 2x2 square. Subsequently, a dropout layer with probability $p = 0.25$ is used to significantly decrease the network's ability to memorize images. This is achieved by ignoring each internal connection with a probability of p during training so that co-dependency between nodes does not develop.

The convolved images (or feature maps) are flattened by taking the 52×52 array and converting it to a 107648 length vector, such that full-connected layers can weight the features extracted by the convolutional layers. Then the first fully-connected layer reduces the number of features to 328 which is approximately the square root of the size of the input vector. Another dropout layer with probability $p = 0.5$ is applied to the data, and two more

fully connected layers with the same dropout layer in between are used to arrive at two features corresponding to the binary classification of both the first and second networks. Finally, a log softmax activation function is used to scale the scores given to each image such that the network is heavily penalized when it makes an incorrect prediction.

2.4. *Hyperparameter Optimization*

The convolutional layers of the neural networks can be modified by adjusting hyperparameters that determine how the CNN processes each image. All of the hyperparameters were chosen by varying the values and testing the CNN for performance improvements until the highest accuracy was achieved. The kernel size is the dimension of a matrix of weights that slides across the image and detects features by multiplying each matrix element by the corresponding pixel on the image. The optimal kernel size depends on the size of features in the image, so certain values will give better performance. A kernel size of 4 pixels was chosen for the first layer by using a set of simulated difference images to train and test the network with varying sizes and maximizing the accuracy of the results. The highest accuracies were obtained by using kernel sizes ranging from 1 to 8 pixels, corresponding to an upper limit of roughly 2 arcseconds (1 DECam pixel = 0.263 arcsec). The higher accuracy is due to the network's ability to detect the presence of host galaxies approximately 2 arcseconds in size. We therefore utilized a kernel size of four pixels as our largest kernel and decreased the kernel size in the deeper layers of the network to find smaller features such as point source transient objects.

2.5. *Data Cleaning*

All of the labels from `ArtifactSpy` were completely determined by human inspection. While this ensured that the CNNs would be tested against verifiable images, the subjective labels could not be expected to be 100% consistent. Even with a standard set of defining characteristics of each class, many types of images were difficult to always identify as a certain class, especially with a team of people applying labels. Preliminary tests of the CNN performance indicated a significant amount of misclassified bad subtractions. Closer inspection of the incorrect predictions showed that human error may have led to incorrect labels.

To inspect misclassified images, we employed a technique called Gradient-weighted Class Activation Mapping (Grad-CAM; Selvaraju et al. 2020) to highlight regions on the images with the highest contributions towards the CNN's classification decision. Convolutional layers retain spatial information that informs classifications in the fully-connected layers. Grad-CAM saves the gradients of each class with respect to the activation maps of the last convolutional layer to produce a localization map, essentially enabling us to inspect what the network sees right before the images are classified. The resulting heatmap of the difference images showed that the network was identifying the same features we had been using in our by-eye classifications; the final classification was only incorrect because the initial label was incorrect. The images determined to have incorrect initial labels were mostly subtle bad subtractions or transients with small amounts of host separation, making them hard for the team of labelers to agree on. With the Grad-CAM technique serving as a lens, we were able to identify when the neural network was making a correct, but wrong due to the initial label of an image, classification and corrected the initial labels in the difference imaging data. An example of this process is shown in Figure 3.

2.6. *Training*

After tuning the neural network hyperparameters for high accuracy using simulated data, we down-sampled our dataset such that the representation of each of the classes was approximately equal to prevent biased guessing by the neural network. To simplify the classifications of the first CNN, the training was performed using only the transient + host class to represent a negative result and the bad subtractions to represent a positive result. The final training of the first CNN consisted of 195 real transients + hosts with an additional 250 simulated images of transients as the positive class to expand the total amount of images used for training to approximately 440 per class. The simulated transients and additional bad subtractions increase the number of samples in our dataset to enhance training performance by preventing any possible overfitting by the CNN on a certain feature of each class. This is an essential trait for the CNN's ability to generalize to any set of images. The training accuracy reaches a value of 94% after 10 epochs as shown in the left panel of Figure 4.

The images that pass the first CNN by testing negative for being a bad subtraction are used to create the training set for the second CNN. The training was completed using the transient + host class to represent a positive result and the remaining classes represent a negative result. In this case, all five classes were used to train the network to account for the variety of not obvious transients that should be classified as negatives. The transient + host consisted of 200 real images with an additional 200 simulated images. The negative class of

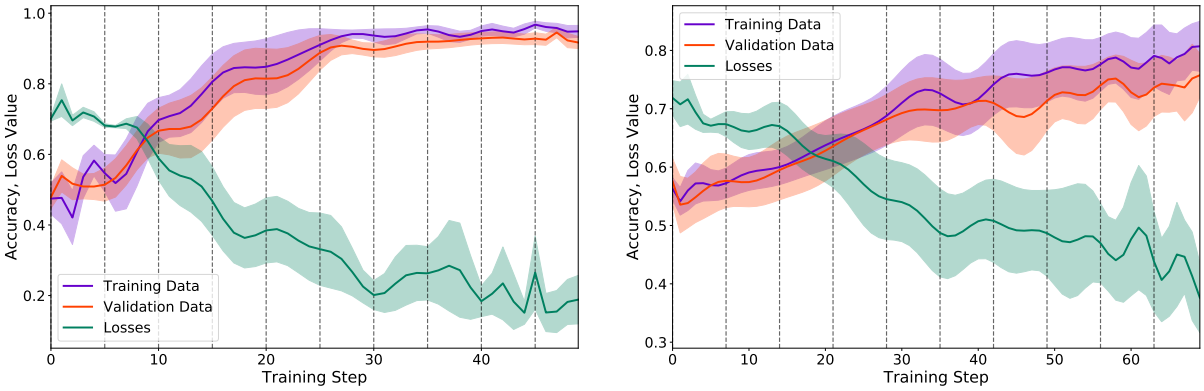

Figure 4. Training and validation accuracy from the first (left) and second (right) networks plotted using 10 k-folds with a batch size of 20. The standard deviation at each training step, corresponding to every tenth batch, is plotted using the shaded region around the mean value represented by the solid line. Additionally, the cross-entropy loss is plotted to show steadily decreasing values for both networks.

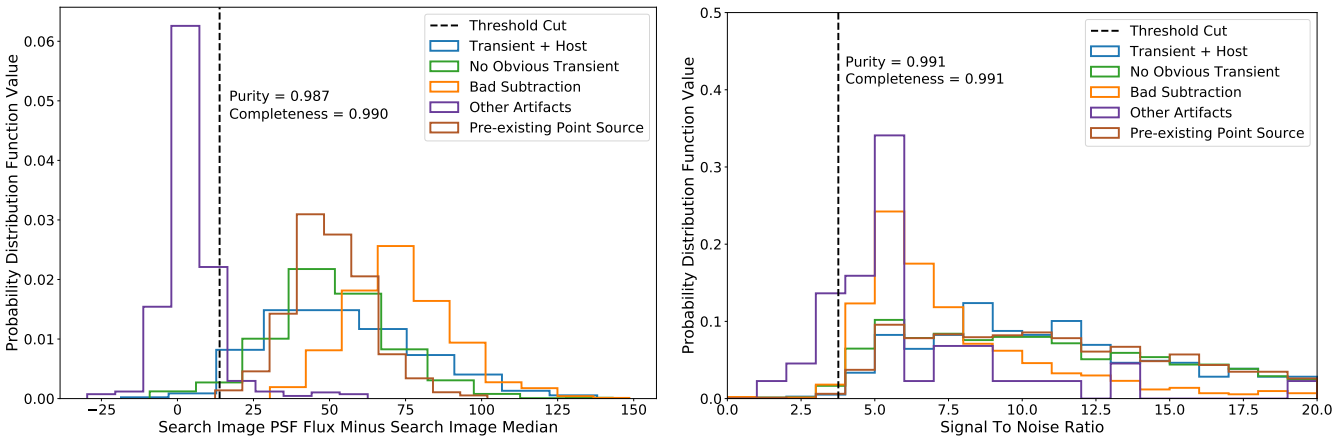

Figure 5. Left: Histogram of weighted flux values from first preprocessing step. Right: Histogram of signal to noise ratio values from second preprocessing step. The purity and completeness was calculated by identifying the other artifacts class as negative and all other classes as positive because we only intended to remove those images in this step.

images used for training consisted of 70 bad subtractions, 30 other artifacts, 250 pre-existed point sources, and 250 not obvious transients. The training accuracy reaches a value of 82% after 10 epochs as shown in the right panel of Figure 4.

3. RESULTS

3.1. *Preprocessing*

The effectiveness of the preprocessing filters are displayed in Figure 5 using histograms. The output from the first step of calculating a weighted value for the flux of each image after subtracting the background is shown in the left panel of Figure 5. The threshold was set by requiring a completeness value of 0.99 to ensure a high number of positive images, represented by all classes not defined as "Other Artifacts", are included. The resulting purity value was 0.987 at a threshold set to 13.86 such that nearly all passed images are true positives.

The output from the second step of calculating a signal to noise ratio of each image is shown in the right panel of Figure 5. The threshold was again set by requiring a completeness of 0.99 such that a high number of the positive images are included. The resulting purity value of 0.991 at a threshold set to 3.76 such that nearly all passed images are true positives. The final filter using catalog matching to eliminate already known stars removed 18% of the pre-existing point source class, 7% of the not obvious transient class, and less than 1% of the remaining classes. The star catalog matching step is included to mimic the procedures of real DECam follow-up observations, and is expected to perform near 100% accuracy.

3.2. *First Network*

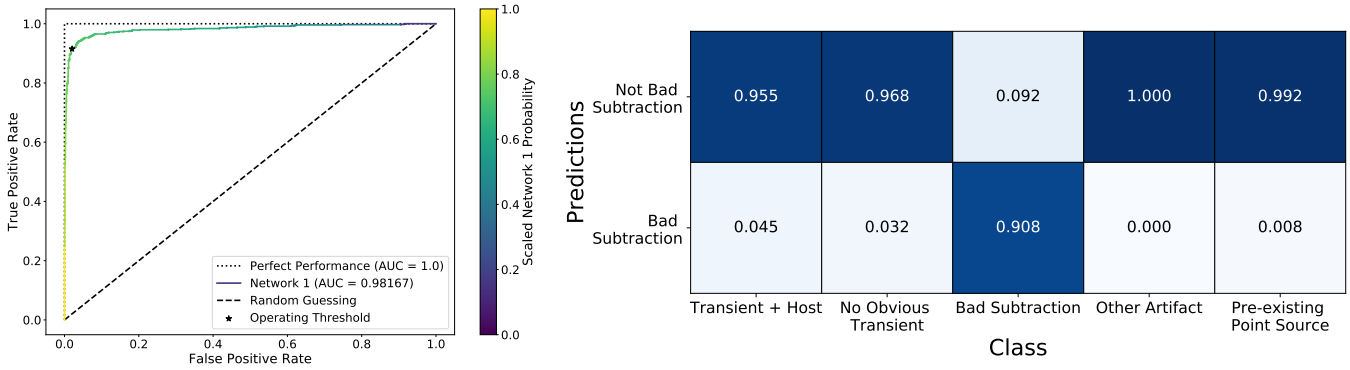

Figure 6. Left: ROC curve of the first CNN's output. Right: Confusion matrix of first CNN predictions.

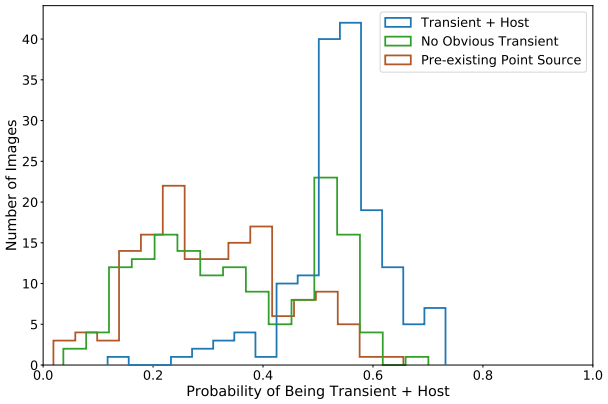

Figure 7. Final probabilities of being a transient + host galaxy for the remaining classes.

The first CNN in the algorithm is applied to the images that passed the preprocessing filters. The output validation probabilities were used to construct a Receiver Operating Characteristic (ROC) curve to evaluate the ideal operating threshold. The curve is plotted in Figure 6. The area under the curve (AUC) of the first network is 0.982 which is near perfect performance. The operating threshold for determining positive and negative results was calculated by maximizing the F1 score which is the harmonic mean of the purity and completeness. Using this threshold, the false positive rate was 3% and the true positive rate was 92%. This means that only 3% of the non-bad subtraction images were incorrectly called bad subtraction and will not be passed to the second CNN. A confusion matrix of the binary predictions using the chosen operating threshold is shown in Figure 6. All non-bad subtraction classes demonstrated an accuracy greater than 95% which was the intended result of using a straightforward binary classification.

3.3. *Second Network*

The final step in the image processing algorithm is applied to the images which were not identified as bad subtractions by the first CNN or removed as other artifacts by the preprocessing. The second network is used to sort the real detections, the images excluding bad subtractions and other artifacts, by their probability of being a real transient source with a host galaxy. The probabilities are plotted on a histogram for the three primary remaining classes in Figure 7. The majority of the transient + host class falls above the 0.5 probability threshold of being labeled correctly. Nearly all of the pre-existing point sources are correctly labeled below the 0.5 probability threshold indicating a clear distinction between these two classes. A large portion of the "No Obvious Transient" class falls at high probabilities which shows the network's ability to identify useful images that are not distinguishable with visual inspection.

3.4. *Efficacy of the Algorithm*

Multiple tests were conducted to examine whether the algorithm has the intended result of expediting multi-messenger counterpart searches. The first such test was comparing the output probabilities from the second network as shown in Figure 7 with the scores given to images by `autoscan`. `autoscan` is the current method for identifying difference image artifacts mainly focused on removing bad subtractions, but it was not designed to distinguish other types of images. The comparison is shown in Figure 8.

The biggest takeaway from this plot is that `autoscan`'s score shown in green has similar median values around 0.9 and similar upper and lower quartiles ranges, which means the not obvious transient class and pre-existing

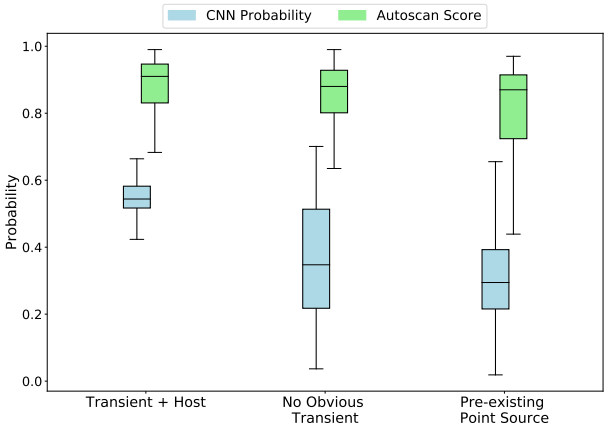

Figure 8. Final probabilities of being a transient + host galaxy compared to `autoscan` probabilities. The boxes extend from the lower to upper quartile of each group with the median marked by the line.

point source class are not being distinguished from the transient + host galaxy class. With our image classification algorithm, there is a significant difference in the probabilities assigned to transient + host compared to the other classes with higher median values by 0.15. This means the output probabilities can be used as indicators of real transients.

The selection function of our approach can also be determined using the purity, completeness, and false positive rate (FPR) of the output probabilities of the algorithm. The probabilities can vary depending on the properties of the images such as the transient apparent magnitude and signal to noise ratio (SNR). As the brightness of transients in images increases, the purity steadily increases to 93% and the FPR decreases to less than 5%. Similarly, for increasing SNR, the purity reaches values of 93% and a FPR of 10% as shown in Figure 9. These results demonstrate that our algorithm can effectively identify images with distinguishable features. The completeness does not change significantly because the transient + host class is consistently given probabilities above 0.5 at a rate of about 80% distributed over a large range of apparent magnitudes and SNR. For typical apparent magnitude and SNR ranges near the center of both plots from Figure 9, we expect purity above 80% for detections of real transients.

3.5. *Testing on Real Observations*

Ultimately, the goal of this processing algorithm is to decrease the number of images requiring visual inspection. We can measure an increase in efficiency by calculating how many images fall above a 0.5 threshold from the output of the 2nd CNN and comparing to the number of images with `autoscan` probability above 0.7. In reality, our approach is only meant to score the images based on their probability of being a host + transient, but we adopt a threshold to make comparisons to `autoscan`. Values above these thresholds are used as standard indicators of a potential transient + host galaxy. After seeing improvements using only our method to filter potential candidates, we applied a weighted combination of the `autoscan` score and our method to produce even better results in terms of the detections above a chosen threshold of 0.8.

The real datasets tested consist of samples of approximately 1,000 stamps from the total population of difference imaging detections. To boost the representation of the host + transient class, part of each sample was collected by sampling objects with an `autoscan` score above 0.7 and part of each sample was collected by sampling objects randomly. This sampling procedure produced a different distribution of `autoscan` scores in the samples than the full population, so we weight the reported detections to correct for that difference. Essentially, we place the population and sample `autoscan` score probability distribution functions (PDFs) in bins of 0.05 and determine the factor required to scale the value of the sample PDF to the population PDF in each bin. We also present the number of images in units of detections per square degree per night. Thus, all test datasets are approximately on the same footing and can be directly compared.

The results from these calculations are shown in Table 2. The number of detections per square degree per night above the thresholds for our method are generally lower in most cases compared to `autoscan`. The combined method of calculating a probability seems to improve on our method across all tested datasets. Focusing specifically on the transient + host class, the number of images below our algorithm's threshold (column 6) and `autoscan` (column 7) demonstrates that both methods do not completely capture the desired images. The significant decrease in the total number of detections using the combined method comes at the cost of fewer transient + host detections above the threshold. The fraction of images passing the thresholds that are transient + host for our algorithm (column 9) and `autoscan` shows the purity of the detections that require inspection. The combined method improves on the purity of detections for all of our tested datasets. Many of the incorrectly passed images are not obvious transient due to the marginal quality of the difference image, so a fairly low purity is expected.

4. DISCUSSION

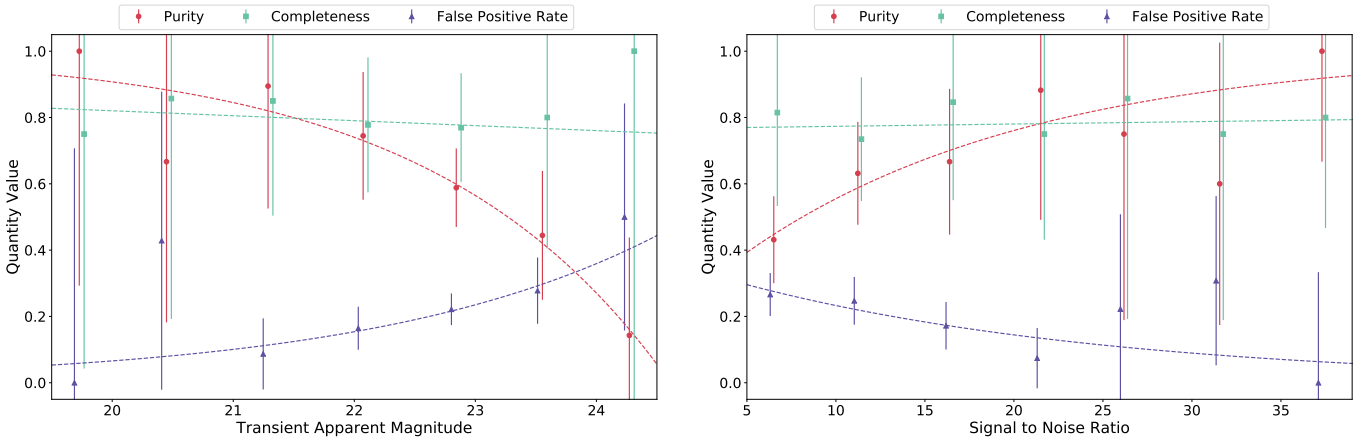

Figure 9. The purity, completeness, and false positive rate of the final output with probabilities greater than 0.5 considered a positive result. The points are determined using the mean value in each bin. The error bars become large on the edges of the apparent magnitude and signal to noise range due to the decreasing number of counts. Since images with SNR below 3.76 were removed during preprocessing, the SNR plot begins around 5.

Testing Results									
	Detections Above Threshold			Transient + Host Detections Below Threshold			Fraction Above Threshold that are Transient + Host Detections		
Dataset	`autoscan`	Our method	Combined	`autoscan`	Our method	Combined	`autoscan`	Our method	Combined
IC201114A	398.712	77.648	51.304	3.732	98.616	91.793	0.437	0.536	0.952
IC171106A	99.508	73.178	36.946	9.794	10.832	17.655	0.351	0.231	0.591
GW190814	152.368	86.072	43.047	7.664	6.335	12.460	0.080	0.150	0.170
IC190331A	25.056	682.883	6.725	2.536	3.454	11.754	0.012	0.014	0.013
GW190728	84.029	358.448	29.255	0.185	23.494	25.816	0.010	0.044	0.011

Table 2. The results from testing our method on real follow-up datasets compared to `autoscan` and the combined method. The values are given in units of detections per square degree per night. A detection was passed by each method if it was greater than an output probability threshold of 0.5 (our method), 0.7 (`autoscan`), and 0.8 (combined).

The development of our approach to do image processing using a convolutional neural network was motivated by the desire to improve the efficiency of multimessenger follow-ups. The efficacy of our algorithm and testing on real observations show improvements to the present method of image processing in the DESGW pipeline using `autoscan`. The improvements decrease the number of images requiring visual inspection by researchers which means the identification of real transient + host objects can be faster. Increased efficiency will be beneficial in the current era of multimessenger astronomy and even more so for future analyses using more advanced detectors with higher rates of events. The implementation of our algorithm will improve the processing technique by making it more robust for future data, in turn, it will be easier to find useful and interesting objects.

The improved efficiency is reflected in the purity of the passed images between our method and `autoscan` as shown in columns 6 and 7 of Table 2. With generally larger fractions of passed transient + host images, fewer images need to be disregarded in the search for real objects. `autoscan`'s lower purity is caused by overestimating the probability of being real with high values for other classes. The distribution of these high scores are shown in Figure 8 with similar median values for the transient + host, not obvious transient, and pre-existing point source classes. Our method has lower certainty for what constitutes a real transient + host with smaller probabilities, but notably higher values than the other two classes meaning the probabilities can be used as improved indicators of real transients. A threshold probability of 0.5 effectively achieves higher completeness and purity than an `autoscan` probability of 0.7.

Realtime follow-ups are dominated by false positives such as point sources and not obvious transients which slow down the identification of real optical counterparts in multimessenger searches. The purity of the transient + host class above the thresholds is shown in columns

8-10 of Table 2 with our method and the combined method outperforming `autoscan`. Candidate identification can happen approximately 1.5 times faster using our method alone and 3.6 times faster in combinations with `autoscan` because there is a decreased number of images requiring inspection with improved purity and completeness of the transient + host class. The benefits of higher efficiency will become even greater in the next era of gravitational wave detectors. We expect a huge increase in the rate of events requiring triggered follow-ups which means faster responses will make real detections easier to find. By integrating our tool into the DESGW pipeline, we are preparing DECam for the increased event rate which will lead to a higher probability of detecting the next multimessenger counterpart.

The deployment of our image processing algorithm will improve the efficiency and scalability of multimessenger counterpart searches with DECam. The output probabilities from these detection methods can be used to set thresholds that filter the vast majority of false detections while maintaining the transient + host detections. As a result, there will be a decreased number of images requiring visual inspection compared to `autoscan` and higher purity and completeness of datasets which demonstrates the efficacy of our method and the combined method. These improvements will become more important during the next generation of multimessenger astronomy when more advanced detectors have higher event rates.

5. CONCLUSION

We developed an algorithm with the goal of improving multimessenger counterpart searches. Presently, these follow-up studies are plagued by high rates of false positive detections, primarily in the form of pre-existing point source objects already visible in the template image and not obviously real transient + host galaxy images. The various classes of images have varying recognizable features that are visible in the pixel values of the images. This visual aspect of classification motivated the use of convolutional neural networks in our algorithm to identify the real transients + host images from false cases. Prior to reaching the CNNs, the images go through a series of preprocessing routines to simplify the classification and eliminate images that cause difficulties. Each CNN has the same structure of 3 convolutional layers followed by a series of activation functions that decrease the output to an array of two probabilities corresponding to the positive and negative cases. The CNNs differ by their training: the first network was trained to identify bad subtractions to be removed from the process, the second network was trained to identify transient + host images. The images classified as not bad subtractions were saved and passed along to the second network where the images were given a probability of being a real transient + host.

The preprocessing removed other artifacts with a completeness and purity of 0.99. The first and second network were trained to an accuracy of 94% and 82% respectively. The first network's ROC curve shown in the left panel of Figure 6 was calculated with an AUC of 0.982 and the operating threshold was used to find a 92% true positive rate. The second network's success is shown in Figure 7. The majority of the transient + host class received probabilities above a 0.5 threshold whereas the most common types of false positives (no obvious transient and pre-existing point source) fall below this value. The output probabilities of the second network compared to existing artifact detection software demonstrate our algorithm's ability to distinguish true positives from false positives. We tested the final product on five unseen real follow-up datasets and demonstrated improvements with a decreased number of images requiring visual inspection using our method. We also created a weighted combination of our method and `autoscan` that further improved our results and decreased the number of images requiring inspection by 3.6.

The results of the various tests and comparisons demonstrate the success of our algorithm at processing difference images such that real transient + host galaxies can be identified more efficiently. Such improvements over the current method built into the DESGW pipeline using `autoscan` will become especially beneficial during the next era of multimessenger astronomy when more advanced gravitational wave detectors go online. Decreasing the amount of required human inspection will expedite the search and the integration of our method into the DESGW pipeline will increase the probability of detecting a multimessenger counterpart.

ACKNOWLEDGMENTS

R. Morgan thanks the LSSTC Data Science Fellowship Program, which is funded by LSSTC, NSF Cybertraining Grant #1829740, the Brinson Foundation, and the Moore Foundation; his participation in the program has benefited this work.

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. 1744555. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

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, NFS'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) e-Universe (CNPq grant 465376/2014-2).

This paper has gone through internal review by the DES collaboration. 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.

Software: `ArtifactSpy` (Morgan 2020), `astropy` (Astropy Collaboration et al. 2013), `deeplenstronomy` (Morgan et al. 2021), `h5py` (Collette 2014), `lenstronomy` (Birrer & Amara 2018), `matplotlib` (Hunter 2007), `numpy` (Harris et al. 2020), `pandas` (McKinney et al. 2010), `PlotNeuralNet` (Iqbal 2018), `PyTorch` (Paszke et al. 2019), `Scikit-Learn` (Pedregosa et al. 2011)

REFERENCES

Aasi, J., Abadie, J., Abbott, B. P., et al. 2015, Classical and Quantum Gravity, 32, 115012

Acernese, F., Agathos, M., Agatsuma, K., et al. 2014, Classical and Quantum Gravity, 32, 024001

Achterberg, A., Ackermann, M., Adams, J., et al. 2006, Astroparticle Physics, 26, 155

Ageron, M., Aguilar, J., Al Samarai, I., et al. 2011, Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 656, 11

Astropy Collaboration, Robitaille, T. P., Tollerud, E. J., et al. 2013, A&A, 558, A33, arXiv:1307.6212

Bertin, E. & Arnouts, S. 1996, Astron. Astrophys. Suppl. Ser., 117, 393

Birrer, S. & Amara, A. 2018, Physics of the Dark Universe, 22, 189

Collette, A. 2014, Python and HDF5 (O'Reilly)

Flaugher, B., Diehl, H. T., Honscheid, K., et al. 2015, The Astronomical Journal, 150, 150

Glorot, X., Bordes, A., & Bengio, Y. 2011, in Proceedings of the fourteenth international conference on artificial intelligence and statistics, 315

Goldstein, D. A., D'Andrea, C. B., Fischer, J. A., et al. 2015, The Astronomical Journal, 150, 82

Harris, C. R., Millman, K. J., van der Walt, S. J., et al. 2020, Nature, 585, 357–362

Herner, K. 2019, in American Physical Society April Meeting, Denver, CO

Herner, K., Annis, J., Brout, D., et al. 2020, Astronomy and Computing, 33, 100425

Hunter, J. D. 2007, Computing in science and engineering, 9, 90

Iqbal, H. 2018, HarisIqbal88/PlotNeuralNet v1.0.0

Kessler, R., Marriner, J., Childress, M., et al. 2015, The Astronomical Journal, 150, 172

LeCun, Y., Bengio, Y., & Hinton, G. 2015, nature, 521, 436

McKinney, W. et al. 2010, in Proceedings of the 9th Python in Science Conference, Vol. 445, Austin, TX, 51

Morgan, R. 2020, ArtifactSpy

Morgan, R., Nord, B., Birrer, S., Lin, J. Y.-Y., & Poh, J. 2021, Journal of Open Source Software, 6, 2854

Morgan, R., Bechtol, K., Kessler, R., et al. 2019, The Astrophysical Journal, 883, 125

Morgan, R., Garcia, A., Soares-Santos, M., et al. 2020a, GCN Circ. 27227

Morgan, R., Palmese, A., Garcia, A., et al. 2020b, GCN Circ. 27366

Morgan, R., Garcia, A., Herner, K., et al. 2020c, GCN Circ. 28955

Morgan, R., Soares-Santos, M., Annis, J., et al. 2020d, The Astrophysical Journal, 901, 83

Paszke, A., Gross, S., Massa, F., et al. 2019, in Advances in Neural Information Processing Systems 32, ed. H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Curran Associates, Inc.), 8024

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, J. Mach. Learn. Res., 12, 2825–2830

Selvaraju, R. R., Cogswell, M., Das, A., et al. 2020, International Journal of Computer Vision, 128, 336

Soares-Santos, M., Herner, K., Garcia, A., et al. 2019, GCN Circ. 25302