Detection, Instance Segmentation, and Classification for Astronomical Surveys with Deep Learning (DeepDISC): Detectron2 Implementation and Demonstration with Hyper Suprime-Cam Data

Grant Merz,[1]⋆ Yichen Liu,[1] Colin J. Burke,[1] Patrick D. Aleo,[1] Xin Liu,[1,2,3] Matias Carrasco Kind,[1,2]
Volodymyr Kindratenko,[2,3,4,5] Yufeng Liu[6]

[1]*Department of Astronomy, University of Illinois at Urbana-Champaign, 1002 West Green Street, Urbana, IL 61801, USA*
[2]*National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, 1205 West Clark Street, Urbana, IL 61801, USA*
[3]*Center for Artificial Intelligence Innovation, University of Illinois at Urbana-Champaign, 1205 West Clark Street, Urbana, IL 61801, USA*
[4]*Department of Computer Science, University of Illinois at Urbana-Champaign, 201 North Goodwin Avenue, Urbana, IL 61801, USA*
[5]*Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, 306 North Wright Street, Urbana, IL 61801, USA*
[6]*Department of Physics, University of Illinois at Urbana-Champaign, 1110 West Green Street, Urbana, IL 61801, USA*

ABSTRACT

The next generation of wide-field deep astronomical surveys will deliver unprecedented amounts of images through the 2020s and beyond. As both the sensitivity and depth of observations increase, more blended sources will be detected. This reality can lead to measurement biases that contaminate key astronomical inferences. We implement new deep learning models available through Facebook AI Research's Detectron2 repository to perform the simultaneous tasks of object identification, deblending, and classification on large multi-band coadds from the Hyper Suprime-Cam (HSC). We use existing detection/deblending codes and classification methods to train a suite of deep neural networks, including state-of-the-art transformers. Once trained, we find that transformers outperform traditional convolutional neural networks and are more robust to different contrast scalings. Transformers are able to detect and deblend objects closely matching the ground truth, achieving a median bounding box Intersection over Union of 0.99. Using high quality class labels from the Hubble Space Telescope, we find that the best-performing networks can classify galaxies with near 100% completeness and purity across the whole test sample and classify stars above 60% completeness and 80% purity out to HSC i-band magnitudes of 25 mag. This framework can be extended to other upcoming deep surveys such as the Legacy Survey of Space and Time and those with the Roman Space Telescope to enable fast source detection and measurement. Our code, DEEPDISC, is publicly available at https://github.com/grantmerz/deepdisc.

Key words: techniques: image processing – methods: data analysis – galaxies: general – Sky Surveys

1 INTRODUCTION

The rise of machine learning/artificial intelligence has allowed for rapid advancement in many image analysis tasks to the benefit of researchers who wish to work with large sets of imaging data. This active field of study, known as computer vision, has led to developments in many disciplines including medical imaging (Zhou et al. 2021), urban planning (Ibrahim et al. 2020), autonomous systems (Pavel et al. 2022) and more.

Tasks such as image compression, inpainting, object classification and detection, and many others have been extensively studied. Astronomy is no exception, and many methods that utilize deep learning have been applied to simulations and real survey data for tasks such as object detection, star/galaxy classification, photometric redshift estimation, image generation, deblending and more (see Huertas-Company & Lanusse 2023 for a comprehensive review). Machine learning methods are already becoming instrumental in handling the large volume of data processed every day in survey pipelines (e.g., Bosch et al. 2018; Russeil et al. 2022; Tachibana & Miller 2018; Malanchev et al. 2021; Mahabal et al. 2019)

The next generation of astronomical surveys such as the upcoming Legacy Survey of Space and Time (LSST; Ivezić et al. 2019) at the Vera C. Rubin Observatory, the Wide-Field Imaging Survey at the Nancy Grace Roman Space Telescope (*Roman*; Spergel et al. 2013), and *Euclid* (Amiaux et al. 2012) will produce unprecedented amounts of imaging data throughout the 2020s and beyond. LSST will provide incredibly deep ground-based observations of the sky, revealing a map of the universe including objects as faint as ~25-27 mag at a 5σ detection for 10 year observing runs. Ground-based surveys such as the Hyper Suprime-Cam Subaru Strategic Program (HSC SSP; Aihara et al. 2018a) and the Dark Energy Survey (DES; Dark Energy Survey Collaboration et al. 2016) have already mapped large swaths of the sky and produced catalogs of tens of millions of objects, with HSC depths being comparable to LSST. The astronomical research community is now in an era that demands robust and efficient techniques to detect and analyze sources in images.

⋆ E-mail: gmerz3@illinois.edu

Current surveys such as HSC already report large fractions of blended (overlapping) objects. For instance, 58% of objects in the the shallowest field (Wide) of the HSC survey are blended, i.e., detected in a region of sky above the 5σ threshold (26.2 mag) containing multiple significant peaks in surface brightness. As depths increase, line-of-sight projections and physical mergers cause the overall number of blends to increase. This fraction rises to 66% for the Deep and 74% for the UltraDeep layers, which are comparable to LSST depths (Bosch et al. 2018). If blends are not identified, they will bias results from pipelines that assume object isolation. For example, Boucaud et al. (2020) show that the traditional detection/deblending methods can lead to a photometric error of >0.75 mag for ~12% of their sample of artificially blended galaxies from the Cosmic Assembly Near-infrared Deep Extragalactic Legacy survey (CANDELS Grogin et al. 2011; Koekemoer et al. 2011). Unrecognized blends can cause an increase in the noise of galaxy shear measurements by ~14% for deep observations (Dawson et al. 2016). Deblending, or source separation, has been recognized as a high priority in survey science, especially as LSST begins preparations for first light.

Despite rigorous efforts to deblend objects, the problem of deblending remains, and in some sense will always remain in astronomical studies. Deblending involves separating a mixture of signals in order to independently measure properties of each individual object. This an imaging problem analogous to the "cocktail party problem", in which an attempt is made to isolate individual voices from a mixture of conversations. However, since it is impossible to trace a photon back to an individual source, astronomical deblending is characterized as an under-constrained inverse problem. Deblending methods must rely on assumptions about source properties and models of signal mixing (Melchior et al. 2021).

A first step in deblending is object detection. Many codes have been developed for source detection and classification, including FOCAS (Jarvis & Tyson 1981), NEXT (Andreon et al. 2000) and SExtractor (Bertin & Arnouts 1996). SExtractor is widely used in survey pipelines including HSC (Bosch et al. 2018) and DES (Morganson et al. 2018), but can be sensitive to configuration parameters. While SExtractor also deblends by segmenting, or identifying pixels belonging to unique sources, modern deblenders have been developed such as Morpheus (Hausen & Robertson 2020) and Scarlet, (Melchior et al. 2018) with the latter implemented in HSC and LSST pipelines. With hopes for real-time object detection and deblending algorithms in surveys such as LSST, machine learning applications to crowded fields offer a promising avenue. The use of deep neural networks, or deep learning has seen particular success in image processing. In addition to efficiency and flexibility, neural networks may be able to overcome limitations of traditional peak-finding algorithms due to their fundamentally different detection mechanism.

There is a growing body of deep learning deblending methods in astronomy. Reiman & Göhre (2019) use a Generative Adversarial Network (GAN) to deblend small cutouts of Sloan Digital Sky Survey (SDSS Alam et al. 2015) galaxies from Galaxy Zoo (Lintott et al. 2011). Arcelin et al. (2021) use a variational autoencoder to deblend small cutouts of simulated LSST galaxies. Hemmati et al. (2022) use GANs to deblend images with HSC resolution and recover Hubble Space Telescope resolution. On larger scales, Bretonnière et al. (2021) use a probabilistic U-net model to deblend large simulated scenes of galaxies.

In addition to blending, another pressing issue with increased depth is the presence of many unresolved galaxies in the deep samples of smaller and fainter objects. This will prove difficult for star-galaxy classification schemes that rely on morphological features to distinguish between a point source star or a point source galaxy, although machine learning methods have been employed to combat this problem (Tachibana & Miller 2018; Miller & Hall 2021). Muyskens et al. (2022) use a Gaussian process classifier to perform star/galaxy classification on HSC images. This is an important area of study, as misclassifications can introduce biases in studies that require careful measurement of galaxy properties. For instance, it has been shown that stellar contamination can be a significant source of bias in galaxy clustering measurements (Ross et al. 2011). Precise constraints of cosmological models require a correction of this systematic bias in measurements of clustering at high photometric redshifts.

The broader field of computer vision has seen a large growth in object detection, classification, and semantic segmentation models. Object detection and classification consist of identifying the presence of an object in an image and categorizing it from a list of possible classes. Semantic segmentation involves identifying the portion of an image which belongs to a specific class, i.e. deblending. Put together, these tasks amount to *instance segmentation*. This pixel-level masking can be used to deblend objects by selecting the pixels associated with each individual object by class. The benchmark leader in deep learning instance segmentation models has been the Mask-RCNN framework (He et al. 2017).

The Mask R-CNN architecture was implemented in Burke et al. (2019) to detect, deblend, and classify large scenes of simulated stars and galaxies. Other architectures have been tested in astronomical contexts, including You Only Look Once (YOLO Bochkovskiy et al. 2020). He et al. (2021) use a combination of the instance segmentation model YOLOv4 and a separate classification network to perform source detection and classification on SDSS images, and González et al. (2018) use a YOLO model to detect and morphologically classify SDSS galaxies. However, these models do not perform segmentation.

The rapid pace of research has led to many new variations and methods that can outperform benchmark architectures. To the benefit of computer vision researchers, Facebook AI Research (FAIR) has compiled a library of next-gen object detection and segmentation models under the framework titled Detectron2 (Wu et al. 2019). This modular, fast, and well-documented library makes a fertile testing ground for astronomical survey data. In addition to a variety of architectures, pre-trained models are also provided. By leveraging *transfer learning*, i.e., the transfer of a neural network's knowledge from one domain to another, we can cut back on training time and costs with these pre-trained models. It is also possible to interface new models with Detectron2, e.g., Li et al. (2022); Cheng et al. (2022), taking advantage of its modular nature and flexibility[1].

In this work, we leverage the resources of the Detectron2 library by testing state-of-the-art instance segmentation models on large scenes, each containing hundreds of objects. We perform object detection, segmentation, and classification simultaneously on large multi-band HSC coadds. Many deep learning applications have been tested on simulated images, but methods applied to real data are often limited by a lack of ground truth. Here, we construct a methodology for using instance segmentation models on real astronomical data, and demonstrate the potential and challenges of this framework when applied to deep images. The HSC data is ideal for testing this framework, as it represents the state-of-the-art among wide/deep surveys, and is closest in quality to upcoming LSST data. By interfacing with Detectron2, we are able to test new models as the repository is updated. We compare models with different performance metrics,

---

[1] See https://github.com/facebookresearch/detectron2/tree/main/projects for a comprehensive list of projects.

and test how robust they are to contrast scalings that alter the dynamic range of the data, which will be important to consider for application to other datasets.

The major contributions of this work can be summarized as 1) Using instance segmentation models to deblend and classify objects in real images from HSC. This demonstrates the feasibility for future integration with wide/deep survey pipelines. We will show that the models can learn inherent features in the data that lead to classification performance gains above traditional morphological methods. 2) Comparing the performances of different models when the input data undergoes different contrast scalings. There is no standard method for scaling image data in astronomical studies that use deep neural networks, so we apply a variety of pre-processing scalings to the data for each model. Dynamic ranges can vary significantly across datasets, and raw data may not be ideal for feature extraction. We test sensitivity to contrast scalings to identify models that will be more easily adapted to different datasets. 3) Interfacing our pipeline with the DETECTRON2 framework to test state-of-the-art models. Of particular note are our tests using transformer-based architectures, an emerging framework in computer vision studies. We will show that these architectures are more robust and accurate than traditional convolutional neural networks in both deblending and classifying objects in large scenes.

This paper is organized as follows. In §2, we present an overview of DETECTRON2 in which we highlight the flexibility of its modular nature and describe the portion of the available deep learning models we implemented. In §3, we describe the curation of our datasets, production of ground truth labels, data preparation and our training procedure. In §4 we present the results of training our suite of models and assess performance with different metrics. §5, we discuss the differences in model capabilities, compare the performance of our pipeline to existing results, and discuss the benefits and drawbacks of our method. In §6, we contextualize our findings and conclude.

2 DETECTRON2 FRAMEWORK

We leverage the modular power of DETECTRON2 by implementing models with varying architectures. The pre-trained models we test in Detectron2's Model Zoo have a structure that follows the GeneralizedRCNN meta-architecture provided by the codebase. This architecture is a flexible overarching structure that allows for a variety of changes, provided they support the following components: (1) a per-image feature extraction backbone, (2) region-proposal generation, (3) per-region feature extraction/prediction. The schematic of this meta-architecture is shown in Figure 1.

The feature extraction backbone takes an input image and outputs "feature maps" by running the input through a neural network, often composed of convolutional layers. In our tests, we use ResNet backbones and transformer-based backbones. ResNets are convolutional neural networks that utilize *skip connections* that allow for deep architectures with many layers without suffering from the degrading accuracy problem known to plague deep neural networks (He et al. 2016). In this paper we explore a few different ResNet backbones: ResNet50, ResNet101 and ResNeXt. A ResNet50 network consists of 50 total layers, with two at the head or "stem" of the network and then four stages consisting of 3, 4, 6 and 3 convolutional layers, respectively. Each stage includes a skip connection. A ResNet101 network is similar to a ResNet50 setup, but with each stage consisting of 3, 4, 23 and 3 convolutional layers, respectively. Subsequent layers undergo a pooling operation that reduces the input resolution. We refer the reader to He et al. (2016) for details regarding these
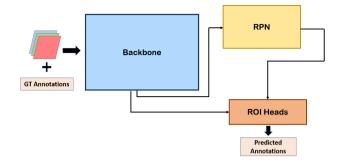

Figure 1. Generalized RCNN meta-architecture. A multi-channel image along with ground truth object annotations is fed to the backbone feature extractor. These features are passed to the RPN and ROI heads to predict object locations and annotations.

layers. ResNeXt layers work similar to ResNet layers, but include grouped convolutions which add an extra parallel set of transforms (Xie et al. 2017). We also test a network with deformable convolutions, in which the regularly spaced convolutional kernel is deformed by a pixel-wise offset that is learned by the network (Dai et al. 2017).

The stages of a ResNet backbone produce feature maps, representing higher level image aspects such as edges and corners. While one can simply take the feature map outputted by the last layer of the backbone, this can pose a challenge in detecting objects of different scales. This motivates the extraction of features at different backbones stages (and thus scale sizes). A hierarchical feature extractor known as a *feature pyramid network* (FPN Lin et al. 2017) has seen great success in object detection benchmarks. The FPN allows each feature map extracted by a ResNet stage to share information with other feature maps of different scales before ultimately passing on to the Region Proposal Network (RPN).

After the image features have been extracted, the next stage of Generalized-RCNN networks involves region proposal. This stage involves placing bounding boxes at points in the feature maps and sampling from the proposed boxes to curate a selection of possible objects. After this sampling has been done, bounding boxes are once again proposed and sent to the Region of Interest (ROI) heads, where they are compared to the ground truth annotations. The annotations consist of bounding box coordinates, segmentation masks, and other information such as class labels. Ultimately, many tasks can be done on the objects inside these regions of interest, including classification, and with the advent of Mask-RCNN frameworks, semantic segmentation. We do not include the details of the RPN and ROI heads, as these structures largely remain the same in our tests. We do test architectures with a cascade structure (Cai & Vasconcelos 2018) which involves iterating the RPN at successively higher detection thresholds to produce better guesses for object locations. For specifics, we refer the reader to Girshick (2015), He et al. (2017) and the DETECTRON2 codebase.

We train a suite of networks to allow for several comparisons. We use a shorthand to denote network configurations as follows.

• R101c4: A ResNet50 backbone that uses features from the last residual stage
• R101fpn: A ResNet101 backbone that uses a FPN
• R101dc5: A ResNet101 backbone that uses a FPN with the stride of the last block layer reduced by a factor of two and the dilation increased by a factor of two
• R50def: A ResNet50 backbone that uses a FPN and deformable convolutions

	median exposure (min)	seeing (")	depth (mag)
g	70	0.83	27.4
r	66	0.77	27.1
i	98	0.66	26.9

Table 1. Properties of the HSC Deep/UltraDeep images

- R50cas: A ResNet50 backbone that uses a cascaded FPN
- X101fpn: A ResNeXt101 backbone that uses a FPN

In addition to these ResNet based models, we also test transformer based architectures. A transformer is a encoder-decoder model that employs *self-attention*. Briefly, self-attention consists of applying linear operations to an encoded sequence to produce intermediate "query, key and value" tensors. A further series of linear operations and scalings are done to these intermediate tensors to produce an output sequence, and then a final linear operation is performed on the entire output sequence. Transformer models have exploded in popularity in the domain of natural language processing due to their scalability and generalizability on sequences, which translates well to language structure. Recently, transformers have been used in computer vision tasks such as image classification and object detection. These models been shown to be competitive with the dominant convolutional neural networks, and are seeing rapid advances in performance measures (Dosovitskiy et al. 2020; Caron et al. 2021; Oquab et al. 2023; Liu et al. 2021; Li et al. 2022). For example, MViTv2 utilizes multi-head pooling attention (MHPA Fan et al. 2021) to apply self-attention at different image scales, allowing for the detection of features of varying sizes. To obtain the input encoded sequences, an image is first divided into patches which are flattened and sent through a linear layer. MHPA is applied to the sequences to produce the image features. In an object detection context, these features are input to an FPN in the same way as features obtained from a ResNet in RCNN models. Another modern transformer model, the Swin Transformer (Liu et al. 2021), applies multi-head attention to image patches, but rather than a pooling operation, use patch merging to combine features of different image patches. Swin models also use shifted window attention to allow for efficient computation and information propagation across the image. We test both MViTv2 and Swin backbones in our implementation.

3 IMPLEMENTATION

3.1 HSC coadds

In this work, the data we use consist of multi-band image coadds of roughly 4000 pixels^2 from the Deep and Ultra-Deep fields of the Hyper Suprime Cam (HSC) Subaru Strategic Program (SSP; Aihara et al. 2018b) Data Release 3 (Aihara et al. 2022). The HSC SSP is a three-tiered imaging survey using the wide-field imaging camera HSC. The HSC instrument (Miyazaki et al. 2017) consists of a 1.77 deg^2 camera with a pixel scale of 0.168", attached to the prime focus of the Subaru 8.2 m telescope in Mauna Kea. The Deep+UltraDeep component of the HSC survey covers ~36 deg^2 of the sky in five broad optical bands (*grizy*; (Kawanomoto et al. 2018)) up to a full 5σ depth of ~27 mag (depending on the filter). Despite limitations (e.g., sky subtraction and crowded field issues), the HSC DR3 data provides the closest match among all currently available deep-wide surveys to the expected data quality of LSST wide fields. The Deep/Ultra-Deep field properties are listed in Table 1. We use the *g*, *r* and *i* bands.

Given the large depth of the survey, a significant portion of objects are blended in comparison to other ground-based surveys such as the

Dark Energy Survey (Dark Energy Survey Collaboration et al. 2016). For reference, 58% of objects in the the shallowest field (Wide) of the HSC survey are blended. While a significant challenge, this lends the HSC fields to be an excellent set of data for testing deblending algorithms, particularly those suited for crowded fields. The pipeline to produce the image coadds is described in detail in Bosch et al. (2018). There are two sets of sky-subtracted coadds. The first set consists of global sky-subtracted coadds. The second set also uses the global sky-subtracted images, but an additional local sky subtraction algorithm is applied. This is to remove the wings of bright objects, artifacts that can cause problems in object detection algorithms. However, this process creates a trade-off with removing flux from extended objects, and Aihara et al. (2018a) empirically find a local sky subtraction scale of 21.5 arcseconds to be a good balance. Ultimately, we use these local sky-subtracted images, as bright wings and artifacts can introduce problems of over-deblending or "shredding" and we want our "ground truth" detections to be as clean and accurate as possible. To further ensure a clean training set, we apply a few quality cuts to the sample. Some images suffer from missing data in one or more bands, especially at the edge of the imaging fields. We use the bitmasks provided in the coadd FITS files to exclude images with >30% of the pixels assigned a NO_DATA flag. Given that the neural network takes multi-band images, if one of the *g*, *r* or *i* band images is flagged in this way, we exclude the other bands as well. There remain some imaging artifacts and issues, such as saturated regions around bright stars, and we discuss how these affect network performance in Section 4.2.

3.2 Ground Truth Generation

We must provide ground-truth object locations and masks to the network to perform pixel-level segmentation. We utilize the multi-band deblending code SCARLET (Melchior et al. 2018) to produce a model for each individual source from which we create an object mask. SCARLET utilizes constrained matrix factorization to produce a spectral decomposition of an object. It is a non-parametric model that has been demonstrated to work well on individual galaxies and blended scenes. Before we run SCARLET, we extract an object catalog using SEP, the python wrapper for SEXTRACTOR. Then, each identified source is modelled and the "blend" or composition of sources is fit to the coadd image data. Once the final blend model is computed, the mask is determined by running SEP on each individual model source and setting a mask threshold of 5σ above the background. Both the SCARLET modelling and mask thresholding are done on the detection image, i.e., the sum over all bands. The run time of this process increases with the number objects in an image. In order to reduce run-time, we divide the 4k stitched coadd images into 16 images of ~1000×1000 pixels^2. While SCARLET on its own is a powerful deblender, the fits can take up to ~30 minutes depending on the number of objects in the image, which motivates the use of efficient neural networks. After this process is complete, we compile a training set of 1000 1k×1k pixels^2 images. The distribution of the number of sources per image is shown in Figure 2.

The trade-off in using real over simulated data is that in supervised tasks, there is a lack of predetermined labels. For the classification task, we produce object labels with a catalog match to the HSC DR3 catalogs. We convert each detected source center to RA and DEC coordinates and then run the MATCH_TO_CATALOG_SKY algorithm in astropy to find objects in the HSC catalog within 1 arcsecond. Then, we compare the *i*-band magnitude of the deblended source to the "cmodel" magnitude of the catalog objects and pick the object with the smallest magnitude difference. If no objects are within 1 arcsec-
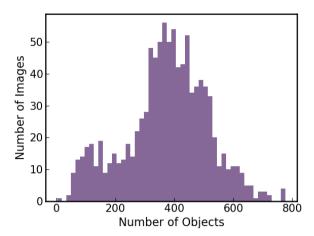
Figure 2. Histogram of the number of objects detected at $>5\sigma$ above the background for HSC images in the training set. The images are taken from both the Deep and UltraDeep fields.

ond or no objects have a magnitude difference smaller than 1, we discard the object from our labelled set. Once an object is matched, we use the HSC catalog "extendedness value" to determine classes, which is based on differences in PSF magnitudes and extended model magnitudes. While yielding high accuracy at bright magnitudes, this metric becomes unreliable for star classification around a limiting magnitude of 24 mag in the i band (Bosch et al. 2018). We additionally discard objects with NaN values in the DR3 catalog, as the class is indeterminate. We show an example image and the results of our labelling methodology in Figure 3, with color-coded classes.

3.3 Data Preparation

We employ three common methods for scaling the raw data from the coadd FITS files to RGB values. These are: a z-scale, a Lupton scale, and a high-contrast Lupton scale. The z-scale transformations are commonly employed in computer vision tasks and are given by

$$
\begin{aligned}
R &= A(i - \bar{I})/\sigma_I \\
G &= A(r - \bar{I})/\sigma_I \\
B &= A(g - \bar{I})/\sigma_I
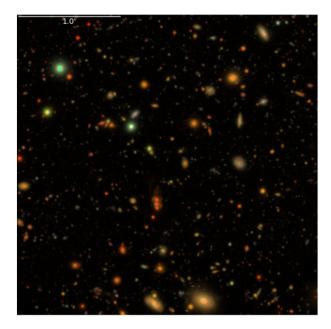\end{aligned} \tag{1}
$$

where $I = (i + r + g)/3$ with a mean $\bar{I}$ and standard deviation σ_I, R is pixel values in the red channel (and similarly for the green G and blue B channels using the r and g -bands respectively). We set $A = 10^3$ for the training and cast the images to 16-bit integers. In addition to z-scaling, we also apply a Lupton scaling from Lupton et al. (2004). This is an asinh scaling with

$$
\begin{aligned}
R &= i(\mathrm{asinh}(Q(I - \mathrm{minimum})/\mathrm{stretch})/Q \\
G &= r(\mathrm{asinh}(Q(I - \mathrm{minimum})/\mathrm{stretch})/Q \\
B &= g(\mathrm{asinh}(Q(I - \mathrm{minimum})/\mathrm{stretch})/Q.
\end{aligned} \tag{2}
$$

We use a stretch of 0.5 and $Q = 10$ and set the minimum to zero and cast the images to unsigned 8-bit integers. Lupton scaling brings out the fainter extended parts of galaxies while avoiding saturation in the bright central regions. These augmentations preserve the color information of objects to aid in classification. Lastly, we also use a high-contrast Lupton scaling, in which image brightness and contrast is doubled after applying the Lupton scaling. We test all of these scalings for each network architecture. In Figure 4, we show an
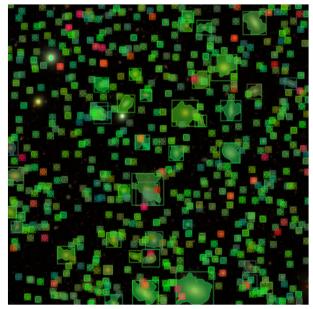

Figure 3. The ground truth masks and bounding boxes on an example image in the test set of our HSC Deep/UltraDeep field data. As this set is class-agnostic, we use white markings for every object. The image without overlaid masks/boxes in shown below for clarity. A Lupton contrast scaling is used in this visualization. Galaxies are colored green, and stars are colored red.

example image and a histogram of pixel values in i, r and g bands (corresponding to RGB colors)

We apply data augmentation to the training and test sets. Data augmentation has become a staple of many deep learning methods. It allows the network to "see" more information without needing to store extra images in memory. We employ spatial augmentations of random flips and 90° rotations. We do not employ blurring or noise addition, as the real data we train on is already convolved with a PSF and contains noise. For future generalizations of this framework to different datasets, then blur/noise augmentations may be useful, but
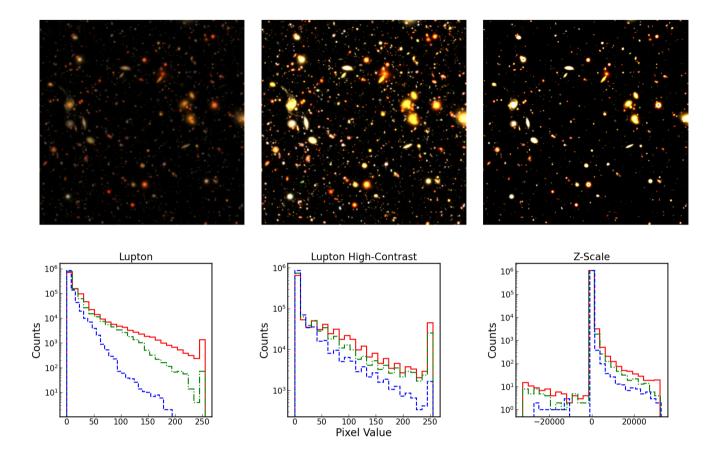
Figure 4. Top row: RGB images in the HSC DR3 dataset with different contrast scalings. The scalings are, from left to right: Lupton, Lupton high contrast, and z-scale. Bottom: Histograms of pixel values to the corresponding image in the top row. Red, green, and blue represent values in the *i*, *r*, and *g* filters, respectively.

for inference purposes on test data taken under the same conditions as the training data, spatial augmentations are sufficient. We also employ a random 50% crop on each image during training so that the data can fit into GPU memory. We considered applying all contrast scalings as a data augmentation, but did not find a significant improvement in network performance. However, this could be used in future work to reduce the training costs, as results were on par with networks trained with only one contrast scaling.

3.4 Training

Training is done using stochastic gradient descent to update the network weights by minimizing a loss function. The loss functions of these Mask-RCNN models is

$$L = L_{\text{cls}} + L_{\text{box}} + L_{\text{mask}} \tag{3}$$

where the classification loss L_{cls} is $-\log p_u$ or the log of the estimated probability of an object belonging to its true class u. Discrete probability distributions are calculated per class (plus a background class) for each ROI. L_{box} is a smoothed L1 loss calculated over the predicted and true bounding box coordinates as given in Girshick (2015). Finally, the mask loss L_{mask} is the per-pixel average binary cross-entropy loss between the ground truth and predicted masks.

All networks are pre-trained on either the MS-COCO (Lin et al. 2014) or ImageNet-1k (Deng et al. 2009) datasets of terrestrial images, and so we use transfer learning to apply these models to the our

astronomical datasets. Transfer learning is a technique in deep learning where networks can generalize knowledge of one task to complete a different but related task (See Tan et al. 2018 for an overview of deep transfer learning). It is often used when applying a pre-trained deep learning model to a different domain than the one seen during training. By using pre-trained weights as initial conditions, training is likely to converge faster and be less prone to over-fitting. We use weights provided by DETECTRON2 as the starting point for our training procedure. We then train the networks for 50 total epochs, i.e. the entire training set is seen 50 total times by the network. In order to facilitate the transfer of knowledge, we first freeze the feature extraction backbones of the models and only train the head layers in the ROI and RPN networks for 15 epochs. We use a learning rate of 0.001 for this step. Then, we unfreeze the feature extraction backbone and train the entire network for 35 epochs. We begin this step with a learning rate of 0.0001 and decrease by a factor of 10 every 10 epochs.

We use two NVIDIA Tesla V100 GPUs in HAL system (Kindratenko et al. 2020) to train on 1,000 images of size 500 pixels^2 paired with object annotations. When trained in parallel on each GPU, our models take roughly ~3 hours to complete. Transformer architectures tend to use more memory, and thus are trained on 4 GPUs for roughly 4 hours.

4 HSC RESULTS

After training, we evaluate network performance on the test set of HSC images. The test set is taken from the patches in the UltraDeep COSMOS (Scoville et al. 2007) field and consists of 95 images of 1000 pixels^2. No test set images were seen during training. A benefit of the instance segmentation models used in this work is their ability to infer on images of variable size. Thus, despite the need to crop images during training, we are still able to utilize the full size of the images in the test set.

We evaluate classification performance with precision and recall, given by

$$p = \frac{TP}{TP + FP}, \qquad (4)$$

$$r = \frac{TP}{TP + FN}. \qquad (5)$$

True positives (TP) are counted as a detection that has a confidence score outputted by the network above a certain threshold and additionally can be matched to a ground truth object by having an Intersection over Union (IOU) above another threshold. False negatives (FN) are those ground truth objects that do not have a corresponding detection. False Positives (FP) are those detections with a high confidence score but do not have a matching ground truth. The IOU is defined as

$$IOU = \frac{area(box_{predicted} \cap box_{truth})}{area(box_{predicted} \cup box_{truth})}. \qquad (6)$$

or the area of the intersection over the area of the union of the predicted and ground truth bounding boxes. Precision and recall are often broken down by class, or combined into one value, the AP score,

$$AP = \frac{1}{51} \sum_{r \in \{0, 0.02, ..., 1.0\}} p(r) \qquad (7)$$

where $p(r)$ is maximum the precision in a recall bin of width Δr. AP scores are computed for IOU thresholds of {0.5,0.55...0.95} and averaged.

AP scores on the HSC COSMOS test set are reported for all network configurations in Table 2. We report the per-class AP score for stars and galaxies separately, as well as the Small, Medium, and Large AP scores, defined by the object bounding box size of 0-32 pixels^2, 32-96 pixels^2 and >96 pixels^2, respectively. For galaxies and stars, AP score can vary significantly across network configurations. For ResNet-based architectures, AP for galaxies is consistently higher than stars, which may be due to the higher sample size of galaxies and morphological features that make galaxies easier to distinguish than compact stars. Among ResNet-based networks, a Lupton high-contrast scaling generally gives the highest galaxy AP score, while a z-scaling always gives the highest star AP score. It appears that these networks are very sensitive to the contrast scaling used, which is not desirable for application to other datasets with different dynamic ranges. However, transformer-based architectures perform more robustly with varying contrast scalings, and outperform ResNet architectures in almost all cases. For these networks, galaxy AP scores all lie within ∼50-52, showing a gain of about 5 over the highest performing ResNet configuration. Stellar AP scores for Lupton and z-scalings lie within ∼33-35, with high-contrast Lupton scalings performing worse by an AP of ∼8. Among the Small, Medium, and Large AP metrics, transformer-based networks also outperform ResNet-based networks, in some cases seeing massive

gains in AP score. The networks generally perform better on Small and Large object categories over Medium objects, again likely due to sample size.

Many studies of instance segmentation models use the MS-COCO or ImageNet-1k datasets as a benchmark to judge performance through the AP score. These data consist of terrestrial images with many object classes, so it can not necessarily be used as a comparison for our AP scores calculated on astronomical survey images with only 2 classes. However, to give a reader a sense of the range of typical values, the AP scores for models trained on terrestrial data typically range from ∼35-45 for convolutional backbones and push to ∼55 for transformer backbones (see the DETECTRON2 repo for results). For a more fair comparison, we look to Burke et al. (2019) in which instance segmentation models were tested on the simulated observations from the Dark Energy Camera (DECam Flaugher et al. 2015). The authors report an AP score for galaxies of 49.6 and score of 48.6 for stars, averaged to a combined score of 49.0. We also train our suite of models on the DECam dataset and report the results in Appendix A. More recently, He et al. (2021) use a combination of the instance segmentation model YOLOv4 (Bochkovskiy et al. 2020) and a separate classification network to perform source detection and classification on SDSS images. They report an AP score of 52.81 for their single-class detection network.

4.1 Incorrect Label Bias Mitigation

There is an inherent bias in our measure of AP scores due to incorrect object class labels. In measurements described above, we test the network abilities to infer classes based on labels generated from HSC catalogs. However, these labels are known to become unreliable, especially for stars, around *i*-band magnitudes of ∼24 mag (Bosch et al. 2018). We use HSC coadds in the COSMOS field for our test dataset, and attempt to mitigate this mislabelling bias by exploiting the overlap of this field with space-based observations using the Advanced Camera for Surveys (ACS) on the Hubble Space Telescope (HST). Because of the lack of atmospheric seeing, morphological classification of stars/galaxies using the HST COSMOS catalog data is much more precise for faint objects, and can be used as ground truth instead of HSC labels. This will test how much poor classification behaviour is due to label generation as opposed to limitations of the models. We generate HST labels by cross-matching detected sources to the catalog of Leauthaud et al. (2007) within 1 arcsecond. If there is no object within 1 arcsecond, we discard the object. There is not necessarily a one-to-one match of HSC versus HST labels, as we are cross-matching to different catalogs, but the number of objects per image remains roughly the same for either labelling scheme. We will refer to this as the HST COSMOS test set.

This small set is not sufficient to train a network, so instead of training on HST-labelled data, we take the models trained on HSC-labelled data and test their evaluation performance on the HST COSMOS test set. To highlight the differences in class label generation, in Figure 5 we show the number of stars and galaxies as a function of HSC *i*-band magnitude for the COSMOS set for both HSC and HST class labels. The unreliable quality of HSC labels at faint magnitudes is reflected in the increased counts of stars, especially the bump in stellar counts beginning at *i*∼25 mag. Also of note is the fewer amounts of star counts in the HSC COSMOS set at bright magnitudes. This is likely due to our HSC label generating procedure of discarding objects with NaN values in the HSC catalog. Bright stars are likely to have saturated pixels in their centers, causing these error flags to appear. With HST labels. we can test with a more astrophysically accurate baseline.

		ResNets						Transformers	
		R101C4	R101dc5	R101fpn	R50cas	R50def	X101fpn	MViTv2	Swin
Galaxies	Lupton	23.7	24.6	40.9	46.3	41.7	41.4	**51.7**	50.8
	LuptonHC	26.1	28.0	43.6	46.0	43.2	43.1	**50.9**	50.3
	zscale	22.9	30.7	40.2	39.6	21.8	34.1	**52.7**	52.5
Stars	Lupton	10.3	9.6	7.3	7.4	4.3	2.5	**34.1**	33.9
	LuptonHC	2.4	5.1	6.1	8.1	5.5	8.3	**28.0**	25.0
	zscale	15.6	10.5	17.9	25.5	12.7	17.2	**35.8**	33.9
Small	Lupton	17.6	18.0	26.1	28.0	24.6	23.7	**43.7**	43.1
	LuptonHC	14.8	17.2	25.9	27.7	25.4	26.9	**40.1**	38.4
	zscale	19.7	21.5	30.2	33.2	18.1	26.8	**44.8**	43.8
Medium	Lupton	8.7	11.9	14.4	11.5	13.7	11.7	**17.4**	16.1
	LuptonHC	7.8	11.1	13.4	12.7	10.3	12.6	**16.3**	15.5
	zscale	3.8	9.0	7.2	7.3	1.6	3.6	**15.1**	14.9
Large	Lupton	16.4	30.9	18.9	14.3	19.6	9.3	**43.1**	41.5
	LuptonHC	15.3	22.8	14.9	15.0	11.6	13.0	38.6	**39.7**
	zscale	0.7	3.6	3.8	5.2	0.1	0.9	**37.8**	37.0

Table 2. AP scores on COSMOS HSC set for all network configurations (larger is better). Galaxy and Star AP scores are calculated separately, whereas Small (0-32 pixels^2), Medium (32-96 pixels^2) and Large (>96 pixels^2) object AP scores are averaged across both classes. The best result for each row is emphasized in bold. The MViTv2 backbone gives the best results in all cases except for one.

Using this new test set, we present AP scores in Table 3. The results for galaxy/star AP scores are in line with the previous results on the HSC COSMOS test set. In all cases, transformer architectures outperform ResNet architectures and are more robust to different contrast scalings. AP scores for Small bounding box objects improves for all network configurations, Medium bounding box AP score roughly remains the same, and Large bounding box AP score worsens. The decrease in Large bounding-box AP scores is likely due to the initial label generation step with sep that over-deblends or "shreds" large extended galaxies and saturated regions around stars. With our HSC label generation, we exclude many of the shredded regions by enforcing the i-band Δ1 mag criterion and discarding labels matched to saturated catalog objects with NaN values. However, our HST label generation is solely based on a distance matching criterion, and so some of these shredded regions are included in the ground truth labels in the HST COSMOS test set. These spurious extra labels can lead to lower AP scores if the networks avoid shredding these regions at inference. In the next section, we examine metrics other than AP score that are less susceptible to this effect.

4.2 Missing and Extra Label Bias Mitigation

Since we have done the labelling ourselves using sep, scarlet and catalog matching to produce ground truth detections, masks and classes, traditional metrics of network performance may not be the best choice in characterizing efficacy. Consider the precision/recall and AP metric. An implicit assumption in these metrics is the completeness and purity of the ground truth labels. This assumption holds for large annotated sets of terrestrial images such as the MS-COCO set (Lin et al. 2014) commonly used as a benchmark in object detection/segmentation studies. It also holds for simulated datasets of astronomical images (Burke et al. 2019) as the ground truth object locations, masks, and classes are all known *a priori* when constructing the training and test set labels. However, real data of large astronomical scenes presents a challenge. Given that we must generate labels without a known underlying truth, any comparisons to this "ground truth" are really comparisons to the methods used to generate these labels. Issues in the label generating procedures will propagate to the performance metrics.

First, the ground truth detections are produced from running sep

using a detection threshold of 5σ above the background. This causes a lack of complete labels, as some objects are missed. We could lower this threshold, but then run the risk of further over-deblending extended/saturated objects. This leads to the second issue in that there will still remain some level of shredding that will cause spurious extra objects to appear in the ground truth set, i.e, a lack of pure labels. If the networks do not shred extended/saturated objects as much as sep, (which is a desirable feature of the networks) then the AP metric will be *lower* due to less spurious network detections than the ground truth. Finally, the object detection mechanisms of the neural networks used in this work are fundamentally different from the peak-finding detection used in sep.

These issues lead to cases in which the neural networks detect objects that are not labelled in our ground truth catalog, despite being actual objects, or cases in which the networks do not detect unphysical objects that are in the ground truth. Any metric that considers true/false detections is subject to this effect. We do not wish to count these cases of fake true/false positives, as this would lead to a reduction in performance metrics that does not reflect network classification/detection accuracy, but rather the limitations of our label generation. Therefore, we construct a set of metrics similar to the canonical precision and recall, but slightly alter our definitions of positive and negative detections. We use equations 4 and 5, but we limit our metrics to the set of objects D that are matched to a ground truth detection. The set of matched detections D is determined by selecting the inferred bounding box with the highest IOU to a ground truth bounding box, above a threshold of 0.5. Then for a given class C, true positives are the objects in D that are correctly classified, false positives are objects that are incorrectly assigned class C, and false negatives are matched objects with a ground truth class C that the network assigns to a different class. With these metrics, precision and recall measure purely the classification power of the network, without bias from missing labels or extra false labels. If we assume that the network's ability to classify remains consistent for objects outside of the matched set, we can generalize these metrics to overall classification performance.

We combine precision and recall into one metric to judge classification power, the F1 score, which is given by the harmonic mean

		ResNets						Transformers	
		R101C4	R101dc5	R101fpn	R50cas	R50def	X101fpn	MViTv2	Swin
Galaxies	Lupton	25.9	26.8	42.9	49.4	43.5	42.8	51.8	**52.4**
	LuptonHC	27.4	30.0	46.2	50.2	46.7	44.3	51.5	**51.6**
	zscale	25.5	32.5	42.7	41.5	23.0	35.6	52.2	**52.9**
Stars	Lupton	16.2	15.0	10.9	10.9	7.1	3.8	52.9	**53.7**
	LuptonHC	4.2	7.9	11.2	14.2	9.4	13.9	**42.1**	37.7
	zscale	28.3	19.1	29.3	41.6	23.8	29.0	**53.9**	52.6
Small	Lupton	22.0	22.1	29.3	31.4	27.0	25.2	54.0	**54.7**
	LuptonHC	16.4	19.9	30.0	33.3	29.4	30.7	**48.2**	46.0
	zscale	28.0	27.1	37.8	42.9	24.8	34.1	**54.7**	54.3
Medium	Lupton	8.3	11.7	13.8	11.0	13.1	11.1	**16.3**	15.2
	LuptonHC	7.5	10.8	12.7	12.2	9.9	12.0	**15.4**	14.6
	zscale	3.7	8.5	7.3	7.4	1.7	3.6	**14.1**	**14.1**
Large	Lupton	6.2	11.1	7.2	5.9	7.2	3.6	**15.1**	15.0
	LuptonHC	5.4	7.9	5.3	4.8	4.4	4.8	13.7	**14.0**
	zscale	0.3	1.2	1.3	1.9	0.1	0.2	**13.6**	13.5

Table 3. Same as Table 2, but with the COSMOS HST test set.

		ResNets						Transformers	
		R101C4	R101dc5	R101fpn	R50cas	R50def	X101fpn	MViTv2	Swin
Galaxies	Lupton	0.96	0.98	0.98	0.98	0.98	0.98	**0.99**	**0.99**
	LuptonHC	0.97	0.98	0.98	0.98	0.98	0.98	**0.99**	**0.99**
	zscale	0.98	0.98	0.98	0.99	0.97	0.98	**0.99**	**0.99**
Stars	Lupton	0.46	0.47	0.33	0.33	0.21	0.15	**0.88**	**0.88**
	LuptonHC	0.23	0.33	0.32	0.40	0.29	0.37	**0.80**	0.75
	zscale	0.69	0.57	0.61	0.76	0.60	0.64	**0.87**	**0.87**

Table 4. F1 scores for star and galaxy classes in the HST COSMOS test set, computed for all network configurations. Transformer networks outperform convolutional networks in all cases, especially for stars.

between precision and recall,

$$\text{F1} = 2 \times \frac{p * r}{p + r}. \tag{8}$$

The F1 score balances the trade-off between precision and recall, with a value close to unity being desirable. We report the F1 scores for the networks on the HST COSMOS test set in Table 4. The best performing configuration among ResNet architectures is the R50cas network with a z-scale scaling. A Swin network with a Lupton scaling achieves the highest overall galaxy and star F1 scores, although the MViTv2 architecture remains competitive. Nearly all transformer networks configurations perform better on star/galaxy classification than ResNet-based networks. Classification power of transformer-based networks is again more robust to contrast scalings than ResNet-based networks.

To examine network performance on faint objects, we show precision and recall as a function of i band magnitude for the HST COSMOS test set in Figure 6. Galaxy recall maintains a value close to one for all objects regardless of magnitude, with some fluctuations of a few percent for some models. Galaxy precision dips for some models at bright magnitudes, which may be due to compact galaxies with bright cores resembling stars. However, these dips are more likely due to inherent limitations of the models rather than label generation, as transformer architectures produce high galaxy precision and recall across magnitude bins compared to ResNet architectures. Most ResNet architectures suffer with stellar recall, with many showing poor performance even at bright magnitudes. Stellar precision reaches near unity at bright magnitudes for all architectures, but many networks configurations begin to drop in performance around i band magnitudes of 21 mag. The best performing networks maintain

a stellar precision above 0.8 out to ~25 mag in the i band. The transformer models we trained are able to achieve a 99.6 percent galaxy recall, 99.2 percent galaxy precision, 85.4 percent stellar recall and 91.5 percent star precision on our HST COSMOS test set, averaged over the whole magnitude range. For comparison, He et al. (2021) perform deep neural network object detection and classification of stars, galaxies, and quasars in large SDSS images. With their sample of objects that covers an r band magnitude range of 14-25 mag, they report a galaxy recall of 95.1 percent, galaxy precision of 95.8 percent, stellar recall of 84.6 percent and stellar precision of 94.5 percent.

4.3 Deblending

In order to quantify deblending performance of the networks, we compute IOU scores for matched objects. The process is similar to the matching done in computing classification precision/recall. We first set a detection confidence threshold of 0.5 and then compute the bounding box IOUs for all detected and ground truth objects. For each ground truth object, we take the corresponding detected object with the highest IOU above a threshold of 0.5. We employ this threshold to avoid the biases discussed in Section 4.2. An IOU of one indicates a perfect match between the ground truth box and the inferred box. In addition to bounding box IOU, we also compute the segmentation mask IOU, which follows from Equation 6, but uses the area of the true and predicted segmentation masks. We report the median IOU for all matched objects in Table 5, and show the distributions in Figure and 7. Transformer-based networks generally produce a higher bounding box IOU than ResNet-based networks, although
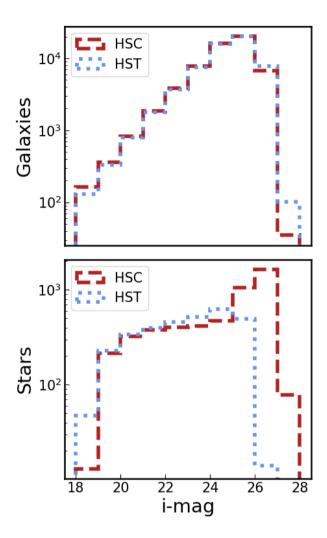
Figure 5. Galaxy and star counts for our COSMOS set, with labels generated from HSC and HST catalogs. The extra counts of HSC stars at faint magnitudes is due to galaxy contamination when classification is based on the extendedness metric. The low sample of bright HSC stars follows from our catalog matching procedure of excluding objects with NaN values.

the R50cas, R101fpn and X101fpn networks remain competitive. Segmentation mask IOUs are lower than bounding box IOUs in all cases. This indicates that while the networks are able to identify overall object sizes quite well, the finer details of object shapes within the bounding boxes are not as well inferred.

The median IOUs measure the ability of the network to detect and segment objects, but it does not fully capture the deblending power of the networks. We examine the cases of a few close blends to get a sense of the ability of the networks to distinguish large overlapping objects. We demonstrate the deblending capabilities of the different networks in Figure 8. In very crowded scenes, the networks are able to distinguish the individual sources, and even pick up objects that are not present in the labelled set, which may present an advantage for studies of low surface-brightness galaxies. As discussed in Section 4.2, this is likely due to the difference in object detection abilities of the Region Proposal Networks compared to peak-finding methods, and highlights that the models are not limited by the training data, but are able to extrapolate beyond it. It is also possible to alter inference hyperparameters such as IOU or detection confi-

dence thresholds, which could allow for more or less detections or overlap between detections. In Figure 9 we demonstrate the effect of lowering the confidence threshold hyperparameter, allowing for more low-confidence detections. While not equivalent, this is similar to lowering the detection threshold in peak-finding algorithms. There are cases in which deblending is poor, and these are typically very large galaxies with one or more very large and very close companions. In such instances, it may be better to use a different contrast scaling. In Figure 10, a Lupton contrast scaling prevents the network from deblending multiple large sources. With the same IOU/confidence score thresholds, a z-scaling works to better isolate the two sources. This is likely due to much larger dynamic range of our z-scaling, which allows for less smearing of the sources and more distinguishing power in this case. Overall, there does not seem to be a one-size-fits-all network configuration for the cases of very large and very close blends. Training on more data would likely improve the ability to detect and segment these objects.

5 DISCUSSION

The effectiveness of instance segmentation models has been proven in many domains, boosted by the ability of networks to work "out-of-the-box" and without much fine-tuning. It has been shown that an object detection model based on the Mask R-CNN framework performs well in the classification and detection/segmentation of simulated astronomical survey images (Burke et al. 2019). In this work, we have trained and tested a broad range of state-of-the-art instance segmentation models on real data taken from the HSC SSP Data Release 3 to push the direction of deep learning based galaxy detection, classification, and deblending towards real applications. Network training and evaluation performance is limited by the efficacy of our label generation methodology, a task not easily formulated when the ground truth is not completely known. This limitation also affects the choices of metrics we use to measure network performance. Often, classification and detection power are combined into the AP score, used throughout instance segmentation literature. However, this may not the best choice of metric for comparisons, as it implicitly assumes the completeness and correctness of the ground truth labels. To attempt to mitigate the effects of incorrect labels on performance metrics, we construct a test set of objects with class labels determined from more accurate space-based HST observations. However, since the AP metric artificially suffers from the detections of "false positives" that are true objects simply missing from the labelled set and/or the presence of spurious ground truth detections, we further attempt to mitigate this bias by constraining performance metrics to detected objects that have a matched ground truth label.

We find that all networks perform well at classifying galaxies, even out to the faintest in the sample. Despite the wide variety of colors, sizes, and morphologies in the real imaging data, our models can identify these objects. Stellar classification is worse, likely due to the smaller sample size in the training and test set. Transformer based networks generally outperform ResNet based networks in classification power of both stars and galaxies. They also appear to be more robust classifiers as magnitudes become fainter. Transformer based models maintain near 100% completeness (recall) and purity (precision) of galaxy selection across the whole sample and above 60% completeness and 80% purity of stars out to i-band magnitudes of 25 mag. These models are able to outperform the extendedness classifier used in the HSC catalogs, which depending on cuts yields near 100% galaxy purity, roughly 90% galaxy completeness, stellar completeness slightly above 50% and stellar purity slightly above
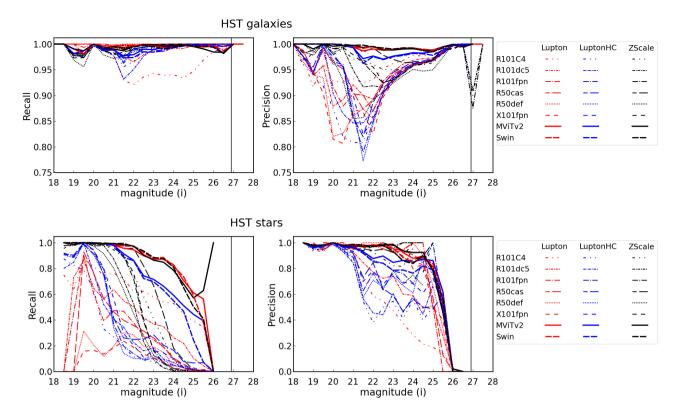
Figure 6. Top: Galaxy precision/recall metrics as a function of object magnitude in the HST i-band. The colors correspond to individual backbone architectures and are shown in the legend. Linestyles represent different network architectures following the legend, and colors indicate which contrast scaling was used (red for Lupton, blue for LuptonHC and black for z-scale). The black vertical line indicates the Deep/UltraDeep i-band 5σ magnitude of 26.9 mag. The y-axis is truncated to better show the differences across the models. Bottom: Stellar precision/recall metrics as a function of object magnitude in the HST i-band.

	ResNets						Transformers	
	R101C4	R101dc5	R101fpn	R50cas	R50def	X101fpn	MViTv2	Swin
Lup	0.75 (0.61)	0.78 (0.57)	0.93 (0.63)	**0.94** (0.62)	0.93 (0.64)	0.93 (0.64)	**0.94** (0.64)	**0.94** (0.64)
LupHC	0.76 (0.61)	0.79 (0.58)	0.93 (0.64)	**0.94** (0.64)	0.93 (0.64)	0.93 (0.64)	**0.94** (0.64)	**0.94** (0.64)
Zscale	0.78 (0.61)	0.81 (0.59)	0.92 (0.62)	0.93 (0.63)	0.82 (0.65)	0.91 (0.64)	**0.94** (0.65)	**0.94** (0.65)

Table 5. Median bounding box IOUs for matched objects in the COSMOS HST test. The best bounding box IOU for each row is emphasized in bold. Also shown in parentheses are the median segmentation mask IOUs. An IOU above 0.5 is considered to be a good match, and a score of 1.0 is a perfect overlap of ground truth and inference.

40% at i-band magnitudes of 25 mag (Bosch et al. 2018). The performance increase of our models is especially noteworthy because they are able to surpass the HSC class labelling despite being trained with it. Transformer models are also more robust to different contrast scalings than traditional convolutional neural networks, indicating that they may be more applicable to a wide range of images across surveys with different dynamic ranges.

The detection/deblending capabilities are measured by the median bounding box IOUs of the networks. Again, transformer based networks generally outperform convolutional ResNet based networks. The improved performance of transformer networks over convolutional based ones may be attributable to the ability of different attention heads to encode information at different image scale sizes (Dosovitskiy et al. 2020), allowing for more overall global information propagation than CNNs. While a convolutional neural network is able to learn spatial features through sliding a kernel across an image, a transformer learns features over the entire input at once, removing any limitations due to kernel sizes. It is possible that the

transformer backbones are implicitly utilizing large scale features in the images such as the spatial clustering of objects, background noise or seeing and using these bulk properties to inform the network.

We examine a few cases of close blends to qualitatively see how the networks distinguish objects. There are cases in which the networks do not detect close objects, but these can sometimes be mitigated by altering the confidence and NMS IOU threshold hyperparmeters (which can be done after training). In other cases, using a different contrast scaling helps to isolate closely blended objects.

There is room to improve both classification and segmentation of these models in future work. One possibility is constructing a larger training set with more accurate labels. With better and larger samples of stars/galaxies, networks may perform better on classification. The more close blends of large galaxies seen during training, the more likely the networks will be able to distinguish these scenes. There could be more fine-tuning of hyperparmeters done to the architectures before training, rather than running them out-of-the-box. Additionally, the use of more photometric information could help
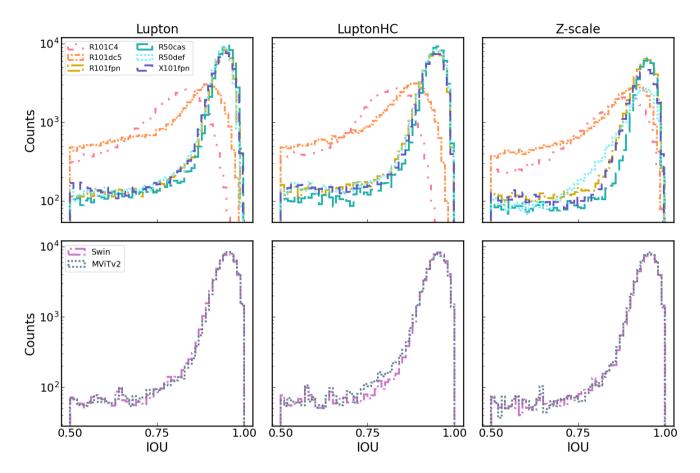
Figure 7. Bounding box IOUs of each detected object that is matched a a ground truth object. Rows show the results for different transformer backbones. Top: results for ResNet backbones. Bottom: results for transformer backbones. The left column represents Lupton scaling, the middle Lupton high-contrast and the right z-scaling.

in all tasks. We use the *i*, *r* and *g* bands on the HSC instrument in this work, corresponding to RGB color images, but could further investigate the performance if we include the *z* and *y* bands.

It is possible that these networks need to be trained longer, or that the fundamentally different properties of astronomical images over terrestrial ones limits the abilities of these architectures in extracting useful features for classification. Despite our attempts to mitigate measurement biases arising from label generation, classification remains a challenge for these models at faint magnitudes. A machine learning model has already been used to classify HSC data using photometry information with better accuracy than morphological methods, but relies on the upstream task of detection (Bosch et al. 2018). The instance segmentation models presented in this work are able to identify and assign classes after training using only an image as input.

6 CONCLUSIONS

It is already a necessary consequence of the current epoch of astronomical research for machine learning algorithms to parse through massive sets of images. A first step in catalog construction is detecting these objects from imaging data. Advancements in the broader computer vision community have given rise to a large ecosystem of models that perform many necessary tasks at once, including detection, segmentation, and classification. While tried and tested on

terrestrial data and shown to work on simulated astronomical data, the application on real survey images remains a work in progress. Many methods rely on the object detection stage to produce measurements of individual objects. In this work, we employ a variety of instance segmentation models available through DETECTRON2 to perform the detection task as well as deblending and object classification simultaneously on images taken from the HSC-SSP Data Release 3. We carefully construct ground truth labels with existing frameworks and catalog matching, and caution that real data gives no straightforward way of producing labels. We find that the best networks perform well at classifying the faintest galaxies in the sample, and perform better than traditional methods at classifying stars up to *i*-band magnitudes of ~25 mag. We find that even if trained on less accurate class labels, the neural networks still pick up on useful features that allow inference of the true underlying class. We expect more data with accurate labels to improve performance. The best performing models are able to detect and deblend by matching ground truth object locations and bounding boxes. Transformer networks appear to be a promising avenue of exploration in further studies.

There are many other areas for future study. While we tested a variety of models, there are many within DETECTRON2 that we did not implement. Some architectures are quite large and require significant resources to train. For example, we attempted to implement ViT backbones (Dosovitskiy et al. 2020) among our set of transformer-based architectures, but were limited by the available GPU memory.
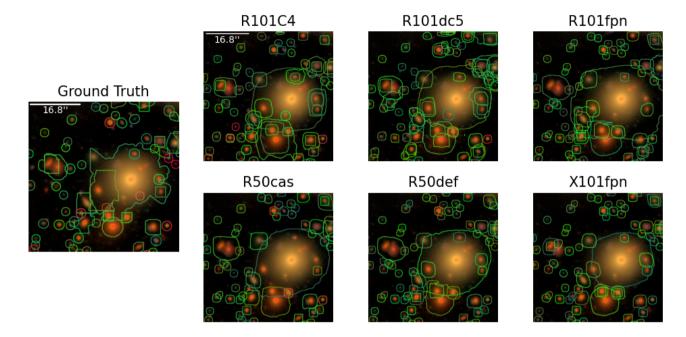
Figure 8. Inference on a close blend. The ground truth is shown on the left. RGB images are created with a Lupton contrast scaling. Other panels show model inference of segmentation maps and classes. Top row, left to right: R101C4, R101dc5, R101fpn. Bottom row, left to right: R50cas, R50def, X101fpn. The colors indicate classes, green for galaxy and red for star. Differences in detections are solely due to the different backbones. While the networks do not pick up every ground truth object, they are also able to detect real objects that were missed by our ground truth labelling.

Figure 9. Inference on the same close blend as Figure 8, but only with a Swin architecture. The ground truth is shown on the left most panel, and the effect of lowering the detection confidence threshold to 0.5, 0.4, 0.3 is shown in left to right, respectively. As the threshold is lowered, objects within a larger footprint are detected.

Many models, especially transformers, are trained with state-of-the-art computing resources at FAIR or other organizations, and subsequently retraining them demands significant resources. Tests could be done on other sets of real data, with other downstream tasks in mind. For example, González et al. (2018) investigate the application of instance segmentation models on SDSS data to classify galaxy morphologies. It would be straightforward to add additional classes, or implement a redshift estimation network using the modular nature of DETECTRON2. In future work we plan to add a photo-z estimator branch to the Mask R-CNN/transformer networks and interface with the LSST software RAIL (Redshift Assessment Infrastructure Layers)[2]. The availability of realistic LSST-like simulations (LSST

Dark Energy Science Collaboration (LSST DESC) et al. 2021) for training will allow us to avoid biases from label generation. The efficiency of neural networks and the ability to perform multiple tasks at once is now a necessity with the amount of survey data pouring into pipelines.

As surveys push deeper into the sky, they will produce unprecedented amounts of objects that will be necessary to process. LSST will provide the deepest ground-based observations ever, and survey terabytes of data every night, highlighting a need for accurate and precise object detection and classification, potentially in real-time. Correctly classifying and and deblending sources will be necessary for a wide range of studies, and deep instance segmentation models will be a valuable tool in handling these tasks.

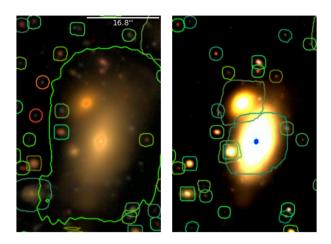[2] https://github.com/LSSTDESC/RAIL

Figure 10. The effect of using a different contrast scaling on a close blend. We show inference of a R50cas network when trained on Lupton scaled images (left) and z-scaled images (right). The objects are more easily distinguished with a z-scaling.

REFERENCES

Aihara H., et al., 2018a, Publications of the Astronomical Society of Japan, 70, S4

Aihara H., et al., 2018b, PASJ, 70, S8

Aihara H., et al., 2022, Publications of the Astronomical Society of Japan, 74, 247

Alam S., et al., 2015, ApJS, 219, 12

Amiaux J., et al., 2012, Euclid Mission: building of a reference survey, doi:10.1117/12.926513, https://doi.org/10.1117/12.926513

Andreon S., Gargiulo G., Longo G., Tagliaferri R., Capuano N., 2000, MNRAS, 319, 700

Arcelin B., Doux C., Aubourg E., Roucelle C., LSST Dark Energy Science Collaboration 2021, Monthly Notices of the Royal Astronomical Society, 500, 531

Astropy Collaboration et al., 2013, A&A, 558, A33

Bertin E., Arnouts S., 1996, A&AS, 117, 393

Bochkovskiy A., Wang C.-Y., Liao H.-Y. M., 2020, arXiv e-prints, p. arXiv:2004.10934

Bosch J., et al., 2018, Publications of the Astronomical Society of Japan, 70, S5

Boucaud A., et al., 2020, MNRAS, 491, 2481

Bretonnière H., Boucaud A., Huertas-Company M., 2021, Probabilistic segmentation of overlapping galaxies for large cosmological surveys, doi:10.48550/arXiv.2111.15455, http://arxiv.org/abs/2111.15455

Burke C. J., Aleo P. D., Chen Y.-C., Liu X., Peterson J. R., Sembroski G. H., Lin J. Y.-Y., 2019, MNRAS, 490, 3952

Cai Z., Vasconcelos N., 2018, in Proceedings of the IEEE conference on computer vision and pattern recognition. pp 6154–6162

Caron M., Touvron H., Misra I., Jégou H., Mairal J., Bojanowski P., Joulin A., 2021, in Proceedings of the IEEE/CVF international conference on computer vision. pp 9650–9660

Cheng J., 2017, PhD thesis, Purdue University

Cheng B., Parkhi O., Kirillov A., 2022, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp 2617–2626

Dai J., Qi H., Xiong Y., Li Y., Zhang G., Hu H., Wei Y., 2017, arXiv e-prints, p. arXiv:1703.06211

Dark Energy Survey Collaboration et al., 2016, MNRAS, 460, 1270

Dawson W. A., Schneider M. D., Tyson J. A., Jee M. J., 2016, ApJ, 816, 11

Deng J., Dong W., Socher R., Li L.-J., Li K., Fei-Fei L., 2009, in 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp 248–255, doi:10.1109/CVPR.2009.5206848

Dey A., et al., 2019, AJ, 157, 168

Dosovitskiy A., et al., 2020, arXiv e-prints, p. arXiv:2010.11929

Fan H., Xiong B., Mangalam K., Li Y., Yan Z., Malik J., Feichtenhofer C., 2021, in Proceedings of the IEEE/CVF international conference on computer vision. pp 6824–6835

Flaugher B., et al., 2015, AJ, 150, 150

Girshick R., 2015, in 2015 IEEE International Conference on Computer Vision (ICCV). pp 1440–1448, doi:10.1109/ICCV.2015.169

González R. E., Muñoz R. P., Hernández C. A., 2018, Astronomy and Computing, 25, 103

Grogin N. A., et al., 2011, ApJS, 197, 35

Hausen R., Robertson B., 2020, ApJS, 248, 20

He K., Zhang X., Ren S., Sun J., 2016, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 770–778

He K., Gkioxari G., Dollár P., Girshick R., 2017, in Proceedings of the IEEE international conference on computer vision. pp 2961–2969

He Z., Qiu B., Luo A.-L., Shi J., Kong X., Jiang X., 2021, Monthly Notices of the Royal Astronomical Society, 508, 2039

Hemmati S., et al., 2022, ApJ, 941, 141

Huertas-Company M., Lanusse F., 2023, Publ. Astron. Soc. Australia, 40, e001

Hunter J. D., 2007, Computing in Science & Engineering, 9, 90

Ibrahim M. R., Haworth J., Cheng T., 2020, Cities, 96, 102481

Ivezić Ž., et al., 2019, ApJ, 873, 111

Jarvis J. F., Tyson J. A., 1981, AJ, 86, 476

Kawanomoto S., et al., 2018, PASJ, 70, 66

Kindratenko V., et al., 2020, in Practice and Experience in Advanced Research Computing. PEARC '20. Association for Computing Machinery, New York, NY, USA, p. 41–48, doi:10.1145/3311790.3396649

Koekemoer A. M., et al., 2011, ApJS, 197, 36

Kroupa P., 2001, MNRAS, 322, 231

LSST Dark Energy Science Collaboration (LSST DESC) et al., 2021, ApJS, 253, 31

Leauthaud A., et al., 2007, ApJS, 172, 219

Li Y., Wu C.-Y., Fan H., Mangalam K., Xiong B., Malik J., Feichtenhofer C., 2022, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp 4804–4814

Lin T.-Y., Maire M., Belongie S., Hays J., Perona P., Ramanan D., Dollár P., Zitnick C. L., 2014, in European Conference on Computer Vision (ECCV). Zürich

Lin T.-Y., Dollár P., Girshick R. B., He K., Hariharan B., Belongie S. J., 2017, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 936–944

Lintott C., et al., 2011, MNRAS, 410, 166

Liu Z., Lin Y., Cao Y., Hu H., Wei Y., Zhang Z., Lin S., Guo B., 2021, in Proceedings of the IEEE/CVF international conference on computer vision. pp 10012–10022

Lupton R., Blanton M. R., Fekete G., Hogg D. W., O'Mullane W., Szalay A., Wherry N., 2004, PASP, 116, 133

Madau P., Dickinson M., 2014, ARA&A, 52, 415

Mahabal A., et al., 2019, PASP, 131, 038002

Malanchev K. L., et al., 2021, MNRAS, 502, 5147

Melchior P., Moolekamp F., Jerdee M., Armstrong R., Sun A.-L., Bosch J., Lupton R., 2018, Astronomy and Computing, 24, 129

Melchior P., Joseph R., Sanchez J., MacCrann N., Gruen D., 2021, Nat Rev Phys, 3, 712

Miller A. A., Hall X. J., 2021, PASP, 133, 054502

Miyazaki S., et al., 2017, Publications of the Astronomical Society of Japan, 70

Morganson E., et al., 2018, PASP, 130, 074501

Muyskens A. L., Goumiri I. R., Priest B. W., Schneider M. D., Armstrong R. E., Bernstein J., Dana R., 2022, AJ, 163, 148

Oquab M., et al., 2023, arXiv e-prints, p. arXiv:2304.07193

Pavel M. I., Tan S. Y., Abdullah A., 2022, Applied Sciences, 12

Peterson J. R., et al., 2015, ApJS, 218, 14

Price-Whelan A. M., et al., 2018, AJ, 156, 123

Reiman D. M., Göhre B. E., 2019, Monthly Notices of the Royal Astronomical Society, 485, 2617

Ross A. J., et al., 2011, MNRAS, 417, 1350

Russeil E., Ishida E. E. O., Le Montagner R., Peloton J., Moller A., 2022, arXiv e-prints, p. arXiv:2211.10987

Scoville N., et al., 2007, ApJS, 172, 1

Spergel D., et al., 2013, arXiv e-prints, p. arXiv:1305.5422

Tachibana Y., Miller A. A., 2018, PASP, 130, 128001

Tan C., Sun F., Kong T., Zhang W., Yang C., Liu C., 2018, A Survey on Deep Transfer Learning: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III. pp 270–279, doi:10.1007/978-3-030-01424-7_27

Wu Y., Kirillov A., Massa F., Lo W.-Y., Girshick R., 2019, Detectron2, https://github.com/facebookresearch/detectron2

Xie S., Girshick R., Dollár P., Tu Z., He K., 2017, in Proceedings of the IEEE conference on computer vision and pattern recognition. pp 1492–1500

Zhou S. K., et al., 2021, Proceedings of the IEEE, 109, 820

APPENDIX A: DECAM RESULTS

For a baseline comparison of network performances, we utilize the PhoSim dataset created and used by Burke et al. (2019). We refer to the earlier work for a full description, but provide a brief summary here. Crowded fields as taken with DECam are produced using the Photon Simulator code (Peterson et al. 2015). Simulations account for equipment optics (Cheng 2017), telescope options (Flaugher et al. 2015) and atmospheric conditions. Spiral, elliptical and irregular galaxies are produced by sampling three-dimensional sersic profiles with additional parameters for extra morphological features. Stars are modeled as point sources and created following an initial mass function from (Kroupa 2001). For both stars and galaxies, SEDs and metallicities are assigned based on physical models. Cosmic star formation history (Madau & Dickinson 2014) is used to assign galaxy number density and population, while the distribution of stars is based on galactic latitude. To simulated crowded fields, the galactic overdensity is boosted by a factor of 4. A 512x512 pixel^2 image is produced with g,r, and z DECam bands. Integration time and magnitude ranges are assigned to roughly correspond to DECaLS DR7 coadds (Dey et al. 2019). In order to assign object masks, a g-band image without background is produced for every object in the field. The PSF is configured to ~1 arcsec. In total, 1000 images are produced for our training set, while an additional 250 are used for validation and another 50 as our test set for evaluation. Each image contains roughly 150 objects.

Here we present the results of two runs on the simulated DECam data, using the R101fpn and MViTv2 backbones. These backbones are chosen to compare the performance of convolutional versus transformer-based architectures. We use the same contrast scalings that were applied to the HSC data, but change the stretch parameter to 100 and Q to 10 for the Lupton and Lupton high-contrast scalings. The dynamic range of the simulated data is different from the HSC data, so the adjustment is done to make galaxy features more distinguishable. AP scores for each configuration are shown in Table A1. We adapt the ranges for Small, Medium, and Large bounding box sizes to match those used in Burke et al. (2019). Overall, we find that a Lupton scaling with a ResNet backbone works best for this dataset, giving the highest AP scores for almost all categories. This is in contrast to the results on HSC data, however we note that a transformer backbone is again more robust to contrast scalings. Although Burke et al. (2019) use a z-scale with a R101fpn backbone, our results are different as we use a slightly altered z-scale formula in that we rescale each band by a constant σ_I rather than a per-band scale factor. This alteration makes galaxy classification performance worse (AP=29.80 compared to AP=49.6) but star classification per-

		R101fpn	MViTv2
Galaxies	Lupton	65.8	62.5
	LuptonHC	58.3	62.2
	zscale	29.8	60.4
Stars	Lupton	70.1	68.0
	LuptonHC	64.3	68.6
	zscale	54.3	66.4
Small	Lupton	68.3	65.7
	LuptonHC	61.8	65.7
	zscale	42.3	63.8
Medium	Lupton	36.1	31.6
	LuptonHC	29.0	46.7
	zscale	16.3	31.7
Large	Lupton	72.6	54.9
	LuptonHC	49.1	65.2
	zscale	38.0	68.0

Table A1. AP scores for DECam runs. Galaxy and star AP scores improve over the results of Burke et al. (2019) when different contrast scalings and backbones are applied. Transformer-based models are more robust to contrast scalings, consistent with results on real HSC data.

formance better (AP=54.32 compared to AP=48.6). The large drop in galaxy AP suggests that the R101fpn backbone is very sensitive to the contrast scaling. All other configurations result in better galaxy and star AP scores than the Burke et al. (2019) results. Our AP scores for Small objects are lower, but Medium and Large are much higher. For size categories, we use the same size definitions as in Burke et al. (2019), but compute an average AP for all IOU thresholds, rather than the AP at only the lowest threshold IOU=0.5. Thus, our results can be thought of as a kind of lower bound, as AP score tends to increase with a lower IOU threshold.

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