MNRAS **000**, 1–13 (2019) Preprint 6 September 2019 Compiled using MNRAS LATEX style file v3.0

Deep Transfer Learning for Star Cluster Classification: I. Application to the PHANGS-HST Survey

Wei Wei,[1,2⋆] E. A. Huerta,[1,3] Bradley C. Whitmore,[4] Janice C. Lee,[5] Stephen Hannon,[6] Rupali Chandar,[7] Daniel A. Dale,[8] Kirsten L. Larson,[5] David A. Thilker,[9] Leonardo Ubeda,[4] Médéric Boquien,[10] Mélanie Chevance,[11] J. M. Diederik Kruijssen,[11] Andreas Schruba[12], Guillermo Blanc[13,14,15], Enrico Congiu[16,13]

[1] *NCSA, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA*
[2] *Department of Physics, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA*
[3] *Department of Astronomy, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA*
[4] *Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, MD, USA*
[5] *Caltech/IPAC, California Institute of Technology, Pasadena, CA, USA*
[6] *Department of Physics and Astronomy, University of California, Riverside, CA, USA*
[7] *Department of Physics and Astronomy, University of Toledo, Toledo, OH USA*
[8] *Department of Physics and Astronomy, University of Wyoming, Laramie, WY, USA*
[9] *Department of Physics and Astronomy, The Johns Hopkins University, Baltimore, MD, USA*
[10] *Unidad de Astronomía, Universidad de Antofagasta, Antofagasta, Chile*
[11] *Astronomisches Rechen-Institut, Zentrum für Astronomie der Universität Heidelberg, Heidelberg, Germany*
[12] *Max-Planck-Institut für extraterrestrische Physik, Garching, Germany*
[13] *Observatories of the Carnegie Institution for Science, Pasadena, CA, USA*
[14] *Departamento de Astronomía, Universidad de Chile, Las Condes, Santiago, Chile*
[15] *Centro de Astrofísica y Tecnologías Afines (CATA), Las Condes, Santiago, Chile*
[16] *Las Campanas Observatory, La Serena, Chile*

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT

Deep learning is rapidly becoming a ubiquitous signal-processing tool in big-data experiments. Here, we present the results of a proof-of-concept experiment which demonstrates that deep learning can successfully be used for production-scale classification of compact star clusters detected in HST UV-optical imaging of nearby spiral galaxies ($D \lesssim 20\,\mathrm{Mpc}$) in the PHANGS-HST survey. Given the relatively small and unbalanced nature of existing, human-labelled star cluster datasets, we transfer the knowledge of state-of-the-art neural network models for real-object recognition to classify star clusters candidates into four morphological classes. We show that human classification is at the 66% : 37% : 40% : 61% agreement level for the four classes considered. On the other hand, our findings indicate that deep learning algorithms achieve 76% : 63% : 59% : 70% for a star cluster sample within 4Mpc $\leq D \leq$ 10Mpc. We further tested the robustness of our deep learning algorithms to generalize to different cluster images. For this experiment we used the first data obtained by PHANGS-HST of NGC1559, which is more distant at D = 19Mpc, and found that deep learning produces classification accuracies 73% : 42% : 52% : 67%. We furnish evidence for the robustness of these analyses by using two different state-of-the-art neural network models for image classification, which were trained multiple times from the ground up to assess the variance and stability of our results. Through ablation studies, we quantified the importance of the NUV, U, B, V and I images for morphological classification with our deep learning models, and find that, as expected, the V-band is the key contributor as human classifications are based on images taken in that filter. The methods introduced in this article lay the foundations to automate classification for these objects at scale, and motivate the creation of a standardized star cluster classification dataset, developed and agreed upon by a range of experts in the field.

Key words: keyword1 – keyword2 – keyword3

1 INTRODUCTION

Human visual classification of electromagnetic signals from astronomical sources is a core task in observational research with a long established history (Cannon & Pickering 1912, 1918; Hubble 1926, 1936; de Vaucouleurs 1963). It has been an essential means by which progress has been made in understanding the formation and evolution of structures from stars to galaxies. However, in the modern era of "Big Data" in Astronomy, with unprecedented growth in electromagnetic survey area, field of view, sensitivity, resolution, wavelength coverage, cadence, and transient alert production, it has become apparent that human classification is no longer scalable (Abbott et al. 2016; LSST Science Collaboration et al. 2009). This realization has motivated the use of machine learning techniques to automate image classification (Ball et al. 2008; Banerji et al. 2010; Carrasco Kind & Brunner 2013; Ishak 2017; Kamdar et al. 2016; Kim & Brunner 2017). Some of these machine learning algorithms have been integrated into widely-used methods for image processing, such as the neural networks trained for star/galaxy separation in the automated source detection and photometry software `SEXTRACTOR` (Bertin & Arnouts 1996). Other applications of machine learning for image classification include the use of so-called decision trees (Weir et al. 1995; Suchkov et al. 2005; Ball et al. 2006; Vasconcellos et al. 2011; Sevilla-Noarbe & Etayo-Sotos 2015) and support vector machines (Fadely et al. 2012; Solarz et al. 2017; Małek & et al 2013).

Visual object recognition has also been a core research activity in the computer science community. For instance, the PASCAL VOC challenge was initiated to develop software to accurately classify about 20,000 images divided into twenty object classes (Everingham et al. 2015). Over the last decade deep learning algorithms have rapidly evolved to become the state-of-the-art signal-processing tools for computer vision, to the point of surpassing human performance. The success of deep learning algorithms for image classification can be broadly attributed to the use of Graphics Processing Units (GPUs) to train, validate and test neural network models, and to the curation of high-quality, human-labeled datasets, such as the `ImageNet` dataset (Deng et al. 2009), which has over 14 million images divided into more than 1000 object categories.

The `ImageNet` Large Scale Visual Recognition Challenge (Russakovsky et al. 2015) has driven the development of deep learning models that have achieved remarkable breakthroughs for image classification. In 2012, the network architecture `AlexNet` (Krizhevsky et al. 2012) achieved a $\sim 50\%$ reduction in error rate in the `ImageNet` challenge—a remarkable feat at that time that relied on the use of GPUs for the training of the model, data augmentation (image translations, horizontal reflections and mean subtraction), as well as other novel algorithm improvements that are at the core of state-of-the-art neural network models today, e.g., using successive convolution and pooling layers followed by fully-connected layers at the end of the neural network architecture.

Within the next two years, the architectures `VGGNet` (Simonyan & Zisserman 2014b) and `GoogLeNet` (Szegedy et al. 2014) continued to improve the discriminative power of deep learning algorithms for image classification using deeper and wider neural network models, and innovating data augmentation techniques such as scale jittering. Furthermore, `GoogLeNet` provided the means to use wider and deeper neural network models to further improve image classification analysis by introducing multi-scale processing, i.e., allowing the neural network model to recover local features through smaller convolutions, and abstract features with larger convolutions. In 2015, the `ResNet` (He et al. 2015) model was the first architecture to surpass human performance on the `ImageNet` challenge. In addition to this milestone in computer vision, `ResNet` was also used to demonstrate that a naive stacking of layers does not guarantee enhanced performance in ultra deep neural network models, and may actually lead to sub-optimal performance for image classification.

In view of the aforementioned accomplishments, research in deep learning for image classification has become a booming enterprise in science and technology. This vigorous program has led to innovative ways to leverage state-of-the-art neural network models to classify disparate datasets. This approach is required because most applications of deep learning for image classification rely on supervised learning, i.e., neural network models are trained using large datasets of labelled data, such as the `ImageNet` dataset. Given that datasets of that nature are tedious and hard to obtain, deep transfer learning has provided the means to classify entirely new datasets by simply fine-tuning a pre-trained neural network model with the `ImageNet` dataset.

While deep transfer learning was initially explored to classify datasets that were of similar nature to those used to train state-of-the-art neural network models, the first application of deep transfer learning of a pre-trained `ImageNet` neural network model to classify small and unbalanced datasets of entirely different nature was presented in George et al. (2018, 2017), where a variety of neural network models were used to report state-of-the-art image classification accuracy of noise anomalies in gravitational wave data. That study triggered a variety of applications of pre-trained `ImageNet` deep learning algorithms to classify images of galactic mergers (Ackermann et al. 2018), and galaxies (Khan et al. 2019; Barchi et al. 2019; Domínguez Sánchez et al. 2018), to mention a few examples.

Building upon these recent successful applications of deep learning for image classification in physics and astronomy, in this paper we demonstrate that deep learning provides the means to classify images of compact star clusters in nearby galaxies obtained with the Hubble Space Telescope (HST), and that this approach can outperform human and traditional machine learning performance. A major motivation of this work is to determine whether deep learning techniques can be used to automate production-scale classification of candidate star clusters in data from the Cycle 26 `HST-PHANGS` (Physics at High Angular Resolution in Nearby GalaxieS[1]) Survey (PI: J.C. Lee, program 15654) for which observations commenced in April 2019. `HST-PHANGS` is anticipated to yield several tens of thousands of star cluster candidates for classification, only about a half of which will be true clusters.

This paper is organized as follows. In Section 2, we summarize the objectives of star cluster classification, and de-

[1] `www.phangs.org`

scribe the current classification system, which we employ in this paper. A review of the consistency between human classifications across prior studies is provided to establish the accuracy level to be achieved or surpassed by deep learning this initial proof-of-concept experiment. In Section 3, we describe the imaging data and classifications used to train our neural network (NN) models, and then provide an overview of the NN models employed in this work. We report our results in Section 4. We conclude in Section 5 with a summary of the results and next steps for future work.

2 CLASSIFICATION OF COMPACT STAR CLUSTERS IN NEARBY GALAXIES

The objects of interest in this study are compact star clusters and stellar associations in galaxies at distances between 4 Mpc to 20 Mpc. The physical sizes of these objects are characterized by effective radii between 0.5pc to about 5pc (Portegies Zwart et al. 2010; Ryon et al. 2017). Hence, only with the resolution of HST (1.2 pc at D=4 Mpc) can clusters be distinguished from individual stars and separated from other star clusters in galaxies beyond the Local Group.

Early attempts at classifying clusters in external galaxies with HST imaging focused mainly on old globular clusters, for example, the swarm of thousands of globular clusters around the central elliptical galaxy in the Virgo Cluster, M87 (Whitmore et al. 1995). This was a fairly straightforward process since the background was smooth and the clusters were well separated. With the discovery of super star clusters in merging galaxies (e.g, Holtzman et al. 1992), the enterprise of the identification and study of clusters in star-forming galaxies began, despite the fact that crowding and variable backgrounds in such galaxies make the process far more challenging. Studies of normal spiral galaxies pushed the limits to fainter and more common clusters (e.g, Larsen 2002; Chandar et al. 2010). In all these early studies, the primary objective was to distinguish true clusters from individual stars and image artifacts, and there were essentially no attempts to further segregate the clusters into different classes.

An exception, and one of the first attempts at a more detailed classification, was performed by Schweizer et al. (1996), who defined 9 object types and then grouped them into two classes: candidate globular clusters and extended stellar associations. More recently, Bastian et al. (2012), who studied clusters using HST imaging of the M83 galaxy, classified star clusters as either symmetric or asymmetric. Their analysis retained only symmetric clusters, which they posited were more likely to be gravitationally bound. Following this work, many studies in the field, most notably the Legacy ExtraGalactic UV Survey (`LEGUS`) (Calzetti et al. 2015) began differentiating clusters into two or three different categories, so that they could be studied separately or together depending on the goals of the project (see also the review by Krumholz et al. 2018, and their discussion of "exclusive" versus "inclusive" cluster catalogs). In `LEGUS`, cluster candidates are sorted into four classes as follows (Adamo et al. 2017; Cook et al. 2019):

- Class 1: compact, symmetric, single peak, radial profile more extended relative to point source
- Class 2: asymmetric, single peak, radial profile more extended relative to class 1 cluster
- Class 3: asymmetric, multiple peaks, sometimes superimposed on diffuse extended source
- Class 4: not a star cluster (image artifacts, background galaxies, pairs and multiple stars in crowded regions, stars)

We adopt the same classification system for this paper. In general, we refer to class 1, 2, and 3 as "compact symmetric cluster," "compact asymmetric cluster," and "compact association" respectively. Examples of objects in each of these classes are shown in Figure 1.

2.1 Consistency among Human Classifications

The stated goal of the current work is to provide automated cluster classifications that achieve accuracy levels at least comparable to human visual cluster classifications. In this section we establish this "accuracy" level, which we define as the consistency between different classifications for the same cluster populations as reported in the literature, as well as relative to classifications homogeneously performed by one of us (Bradley C. Whitmore, hereafter BCW).

A first look at the overall consistency between the clusters cataloged by different studies, but based on the same data and same limiting magnitude, is provided by the work on M83 by Bastian et al. (2012); Whitmore et al. (2014); Chandar et al. (2014). Comparisons reported in those papers show that about ∼70% of the clusters are in common between the studies. Later, Adamo et al. (2017) performed a similar comparison for the spiral galaxy NGC 628 for the catalogs from `LEGUS` and Whitmore et al. (2014), and finds a total-match fraction of ∼75%. Finally, the LEGUS study of M51 by Messa et al. (2018) find a total-match fraction of 73% in common with a study by Chandar et al. (2016).

However, these results are not based upon detailed analysis of human-vs-human cluster classifications for individual objects; they are statistical measures of overlap between samples where a mix of human classification/identification, and automated star/cluster separation based on the concentration index (i.e., the difference in magnitude in a 1 pixel vs. 3 pixel radius) were used across the studies.

To directly evaluate human-vs-human cluster classifications, we compare classifications assigned by BCW for NGC 4656 to those provided in the `LEGUS` public cluster catalog, which provides the average classification made by three other `LEGUS` team members (trained by BCW and Angela Adamo). Results are shown in Figure 2. For the combination of class 1 + 2 + 3, the total match fraction is 76%, which is similar or higher than the agreement in the prior studies M83, NGC 628 and M51 discussed above. If we combine only the class 1 + 2 clusters (to exclude compact associations which has a higher rate of confusion with class 4 non-clusters), the total match fraction is 67%. For the individual classes, the consistency of the assignments vary from 66%, 37%, 40%, 61% for class 1, 2, 3, and 4, respectively. We adopt these as the "accuracy" levels to be achieved or surpassed by the deep learning studies here.

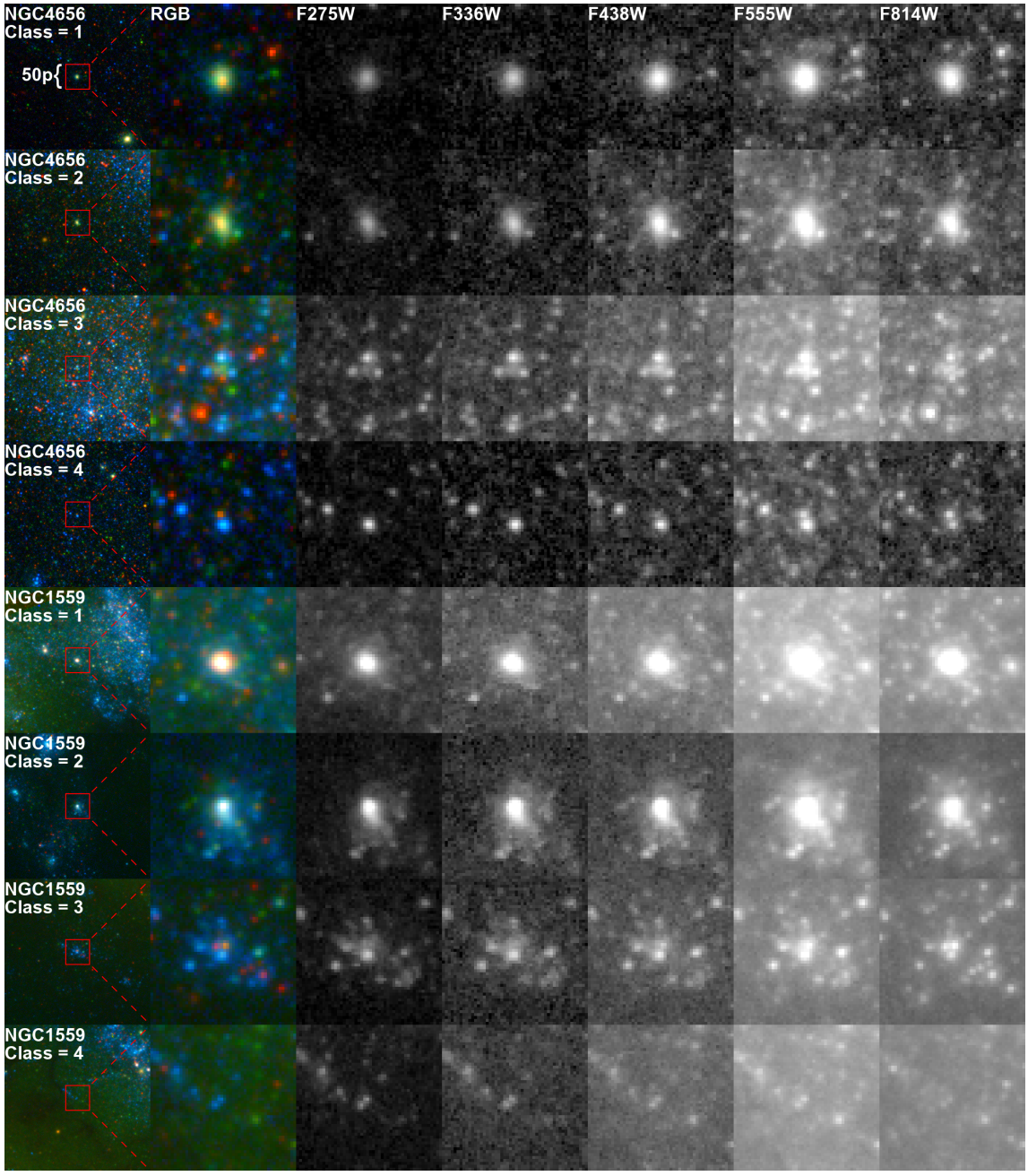

Figure 1. Examples of each of the four cluster classifications. The top four rows show clusters from NGC 4656, which are part of the training set, while the bottom four rows show clusters from PHANGS-HST observations of the spiral galaxy NGC 1559, which form our proof-of-concept test sample, and are not used for training. The first two columns show false-color RGB images for context: the first column displays a 299p x 299p RGB image (R = F814W, G = F438W + F555W, B = F275W + F336W) and the second column shows only the center 50p x 50p of the RGB image (184pc x 184pc for NGC1559, for example). Only the center 50p x 50p of individual NUV-U-B-V-I HST imaging is used for training and evaluation, and these are shown in grayscale in the last 5 columns (from left to right, 50p x 50p images taken with filters F275W, F336W, F438W, F555W, and F814W).

3 METHODS

In this section we describe the data sets used to train, validate and test our deep learning algorithms, and give an overview of the neural network models used. We approach this initial work as a proof of concept demonstration, with the intention of performing further optimization and more detailed tests in future work.

3.1 Data curation

For training, we use classifications which were performed by BCW for HST pointings in 10 galaxies, which are in both `LEGUS` and the Hα-`LEGUS` follow-up survey. In total, this provides samples of 1300, 1000, 700, and 2100 class 1, 2, 3, 4 objects for training, as described in Table 1. We use 80% of this sample for training, and reserve the remainder for validation.

Field	D (Mpc)	Class 1	Class 2	Class 3	Class 4
NGC3351	10.0	118	80	95	325
NGC3627	10.1	403	175	164	837
NGC4242	5.8	117	60	14	42
NGC4395N	4.3	8	19	21	20
NGC4449	4.31	120	261	213	0
NGC45	6.61	45	52	20	43
NGC4656	5.5	83	125	47	173
NGC5457C	6.7	287	108	81	436
NGC5474	6.8	48	95	34	144
NGC6744N	7.1	164	143	58	210
Total		1393	1118	747	2230
N $\geq$ 4		1271	1013	738	2125

Table 1. Number of sources in each of the ten HST LEGUS fields which have been classified by BCW and are used for training in this study. The number in each of morphological classes described in Section 2 is given. The total number of clusters with detection in at least four filters (a requirement for inclusion in the training and testing) are also given in the last row of the table. 80% of the latter are used for training, and the remaining 20% are reserved for testing. Distances compiled by (Calzetti et al. 2015) are listed.

Field	D (Mpc)	Class 1	Class 2	Class 3	Class 4
NGC1559	19.0	302	252	162	710

Table 2. Number of sources in the PHANGS-HST observation of NGC 1559 which have been classified by BCW. This cluster sample is used to test the neural networks trained with the data described in Table 1 as a proof-of-concept for production scale classification of PHANGS-HST compact clusters and associations.

To investigate whether networks trained in this manner can be used to automate classification of star clusters in the PHANGS-HST dataset in the future, we test the networks on the first observations obtained by PHANGS-HST of the spiral galaxy NGC 1559. NGC 1559 is about twice as distant as the most distant galaxy in the training sample. The PHANGS-HST NGC1559 observations provide 302, 252, 162, and 710 class 1, 2, 3, 4 objects, as determined by BCW (Table 1).

The available star cluster data sets are small and unbalanced (different number of images for each class), compared to the datasets used to successfully train state-of-the-art neural network models for image classification. Thus, we use two neural network models, `VGG19` (Simonyan & Zisserman 2014a) with batch normalization (`VGG19-BN`) and `ResNet18` (He et al. 2016), pre-trained with the `ImageNet` dataset (see Section 1), and then use deep transfer learning to leverage the knowledge of these models to classify real-object images to our task at hand, namely, the morphological classification of star clusters.

As input for the neural network training, we use cutouts from HST imaging obtained by `LEGUS` together with the classifications provided by BCW. LEGUS obtained WFC3 observations in 2013-2014 (GO-13364; PI Calzetti), and combined those data with ACS data taken in previous cycles by other programs to provide `NUV-U-B-V-I` coverage for a sample of 50 galaxies. The Hα-`LEGUS` follow-up program (GO-13773; PI Chandar) obtained additional imaging in a narrow-band filter covering the Hα emission-line (F657N) and a medium-band filter sampling line-free continuum (F547M) for the 25 galaxies with the highest star formation rates, but we note that the Hα data are not used in this work. Training and testing are based on HST broadband imaging in the NUV, U, B, V, and I filters. Sample images are presented in the last five columns of Figure 1. Including the nebular gas emission captured by the Hα imaging in the training can be explored in future work.

As of July 2019, BCW classifications for 5 of the 10 HST fields used here (Table 1) are available from the `LEGUS` public archive hosted by MAST[2]. Additional classifications performed by other `LEGUS` team members are publicly available for an additional 29 fields. In this work, we opt to only use the BCW dataset for two reasons. First, it represents a more homogeneous set of classifications, and internal consistency of the classifications is crucial for successful training of the networks. Moreover, the galaxies examined by BCW are more similar in star formation properties to those in the `HST-PHANGS` sample, to which we will apply deep learning techniques at production-scale. The `HST-PHANGS` sample primarily contains star-forming spiral galaxies with stellar masses greater than $\sim 10^{10} M_{\odot}$, whereas most of the additional galaxies which have `LEGUS` classifications are dwarf galaxies (Cook et al. 2019). 17/50 of the LEGUS galaxies can be considered to be dwarf galaxies. Consequently, another difference between the LEGUS and PHANGS-HST samples is that the PHANGS-HST galaxies are roughly twice as distant.

PHANGS-HST began observations on April 6, 2019 and is also obtaining observations in the NUV-U-B-V-I filters. The first galaxy to be observed is NGC 1559, at a distance

[2] https://archive.stsci.edu/prepds/legus/dataproducts-public.html

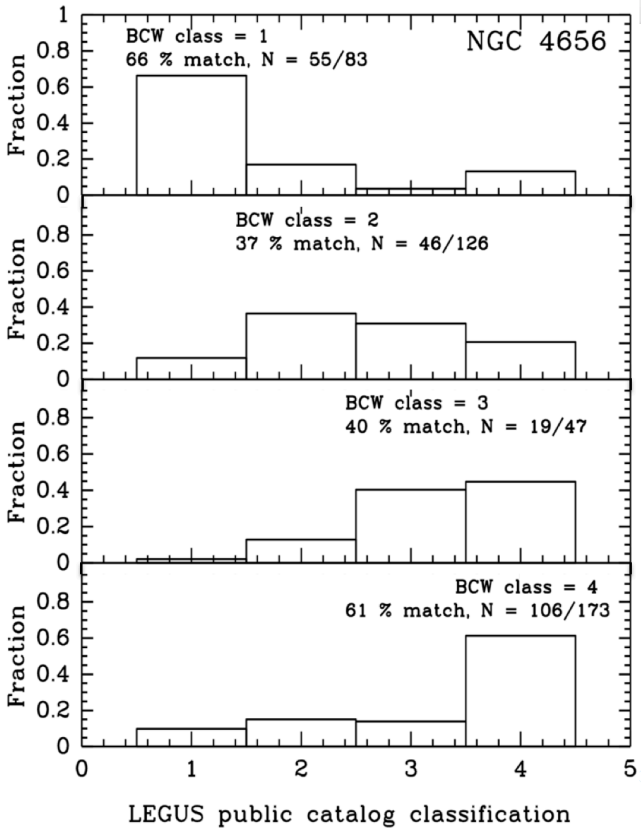

Figure 2. Comparisons between star cluster candidate classifications made by BCW and the mode of classifications made by three other LEGUS team members (trained by BCW and Angela Adamo) provided in the LEGUS public star cluster catalog for NGC 4656. Each panel shows the distribution of classifications given in the LEGUS catalog for BCW labelled class 1 (top, symmetric compact clusters), class 2 (upper middle, asymmetric compact clusters), class 3 (lower middle, compact associations) and class 4 (bottom, non-clusters) objects.

of 19 Mpc (A. Reiss, private communication), and we use BCW classifications for clusters in that galaxy for testing.

Bearing in mind that the VGG19-BN and ResNet18 models used in this study were pre-trained with the ImageNet dataset, in which images are resized to $299 \times 299 \times 3$, we follow best coding practices of neural network training, and curate our datasets so that star cluster images have size 299×299 pixels.

Given that the clusters subtend only a several to a dozen HST WFC3 pixels, we focus the training on a small area (see Figure 1). The central 50 x 50 pixels of the multi-extension fits (MEFs) are resized to fit in an 299 x 299 pixel area for the training. With WFC3's pixel size of .04 arcseconds, each postage stamp corresponds to a physical width between ∼40-100pc for our sample of galaxies. Testing whether the size of the cropped HST image influences the accuracy can be explored in future work.

Procedurally, from the HST mosaics, a 50 x 50 pixel .fits image (i.e., a "postage stamp") centered on each target cluster is cropped from each of the NUV-U-B-V-I bands. The five resultant stamps for each cluster are then stored in individual header data units (HDUs) within a single MEF file. We note that if there was no observation of the cluster in one of the filters, all pixel values for that particular filter's postage stamp were set to zero. If there was no observation in more than one filter, the cluster was removed from our sample.

We also introduce other modifications to our neural network models, described in more detail below, such that we can input these $299 \times 299 \times 5$ MEFs, and output probability distributions for four classes of images.

3.2 Neural network models

As mentioned before, we use two neural network architectures, VGG19-BN and ResNet18. These models have 3 input channels. However, since our images have 5 input channels available, we concatenate two copies of the same neural network architecture. The merged neural networks would have 6 input channels in total, so we set the input to the last channel to be constant zeros. We also apply one more matrix multiplication and an element-wise softmax function (see Appendix A) (Goodfellow et al. 2016) to make sure that for each image the output is a vector of size 4, representing the probability distribution over the 4 classes under consideration. We choose this particular combination given its simplicity and its expected performance for image classification.

Furthermore, following deep learning best practices, we quantify the variance in classification performance of our models by training them ten times independently and then presenting the mean accuracies and the corresponding standard deviations. We also compute the Shannon entropy Shannon (1948) of the output distribution over the four star cluster classes to quantify the uncertainty in each individual neural network model's prediction.

3.3 Training Experiments

To attain state-of-the-art classification accuracies with a star cluster data set that has just a few thousand images, we transfer the knowledge of VGG19-BN and ResNet18 for real-object recognition—which were trained with millions of high-quality, human-labelled, real-object images in the ImageNet dataset—for star cluster classification[3]. We use the pre-trained weights, except those for the last layers, of VGG19-BN and ResNet18 provided by PyTorch (Paszke et al. 2017) as the initial values for the weights in our models. The weights for the last layers in VGG19-BN and ResNet18 and the last fully connected layers are randomly initialized. We use cross-entropy as the loss function[4] and Adam (Kingma & Ba 2014) for optimization. The learning rate is set to 10^{-4}. The batch size for ResNet18 is 32, and for VGG19-BN is 16.

To fine-tune our neural network models through transfer learning, we use 80% of the BCW dataset in Table 1,

[3] A brief overview of transfer learning is presented in Appendix B.

[4] A loss function is used to evaluate and diagnose model optimization during training. The penalty for errors in the cross-entropy loss function is logarithmic, i.e., large errors are more strongly penalized.

and reserve the rest for testing. Absolute values of pixels are rescaled to be in the range $[0, 1]$, to avoid the brightness of the sources from becoming a parameter in the classification. During training we use several standard data augmentation strategies, such as random flips, and random rotations in the range $[0, 2\pi]$ to make sure that the trained neural networks are robust against those transformations. Taking into account the batch sized mentioned above for `ResNet18` and `VGG19-BN`, and bearing in mind that we trained the models using about 10,000 batches, this means that the nets were exposed to 320,000 and 160,000 images, respectively. Note, however, that the data augmentation techniques used during the training stage may produce very similar images to the actual star cluster images curated for this analysis.

4 RESULTS

We present four sets of results in this section. In Section 4.1, we present the classification accuracy for the four categories of star clusters candidates relative to the BCW determinations, and quantify the robustness of our neural network models to generalize to star cluster images in different galaxies, choosing the galaxy NGC 1559 as the driver of this exercise as discussed above. In Section 4.2, we determine the relative importance of different filters for image classification in our deep learning model, and in Section 4.3, we present the uncertainty quantification analysis of those models.

4.1 Classification Accuracy

First, we quantify the performance of our models for classification accuracy when we fine-tune the models from scratch ten times to determine whether the transfer learning was effective at learning the morphological features that tell apart the four classes of star clusters, and to assess the robustness of the optimization procedure for image classification. These results are summarized in Figure 3, which presents the mean classification accuracy averaged over the ten models.

The variance in the ten independent classification measurements provide another measure of the robustness of the models. We present results for the classification accuracy obtained by our neural network models, and the corresponding variance, in Tables 3 and 4. It is worth emphasizing that these results outperform human-classification results presented in Figure 2.

To further assess the robustness and resilience of our neural network models to classify images from galaxies not included in the original training dataset, we have used another batch of images from PHANGS-HST for testing purposes that correspond to the galaxy NGC 1559. This galaxy is two to four times further away than any of the images we used for training or for testing purposes in Tables 3 and 4. Notwithstanding this significant difference, we notice in Tables 5 and 6 that our neural network models still outperform human classification results, furnishing evidence for the suitability of neural network models for automated classification of PHANGS-HST sources at production-scale.

In sum, all of our classification algorithms outperform agreement in human comparisons as determined in Section 2.1 and illustrated in Figure 2. Tables 3 (5) and 4 (6) are very similar, showing that ResNet18 and VGG19-BN are roughly equal in their performance. The numbers are slightly lower for categories 2 and 3 for the PHANGS galaxy NGC 1559.

4.2 Classification accuracy as a function of input data

We have also quantified what filter has the leading contribution for classification accuracy. To do so, we perform the following experiment: using NGC 1559 images as testing dataset, we produced five different testing datasets in which one filter was set to zero. We then fed these 5 different testing datasets, one at a time, in our neural network models and quantified which missing filter leads to the most significant drop in classification accuracy. As shown in Figure 4, the key filter is F555W.

This finding is expected, since the BCW human classifications primarily rely on the F555W image (e.g., using DS9 and imexamine), with color images (F814, F555, F336W) generated by the Hubble Legacy Archive providing supporting morphological information. Therefore, our neural network models seem to use insights similar to human vision to classify star cluster images.

4.3 Uncertainty quantification

Having demonstrated that using different initial conditions for the training of our models produces consistent classification results, which is accomplished by training our models multiple times, in this section we address a complementary issue, that of uncertainty in the models' predictions.

A common method to address this is through the computation of entropy, which is done by using the output of our models, that consists of probability distributions for each of the cluster classes we are trying to classify. Intuitively, the more pronounced the peak is in the distribution, the more confident the neural network is about its prediction, and in this case, the entropy calculated from the prediction probability distribution will be lower. For example, if the probability distribution is only concentrated on one class, the network network in this case is 100% certain about its prediction and the entropy would be zero, i.e., there is no uncertainty in the prediction. On the other hand, if the prediction assigned the same probability for all the 4 classes under consideration equally, we would have maximum uncertainty in this case, since for the given input image, all the 4 classes are equally possible to be the predicted class, and in this case, the maximum entropy is $\ln 4 \approx 1.39$. Figure 5 shows the distribution of the entropies for the predictions of `VGG19-BN` when tested on NGC 1559 images.
Figure 5 is a reflection of the classification accuracies reported before, i.e., while our neural network models exceed the baseline for human classification established in Section 2.1, there is still a lot of work ahead to construct standardized datasets that may be used to further increase the classification accuracy of the models we have introduced herein. In that scenario, we would expect the entropy distributions presented in Figure 5 to be skewed even more towards entropy values around zero.

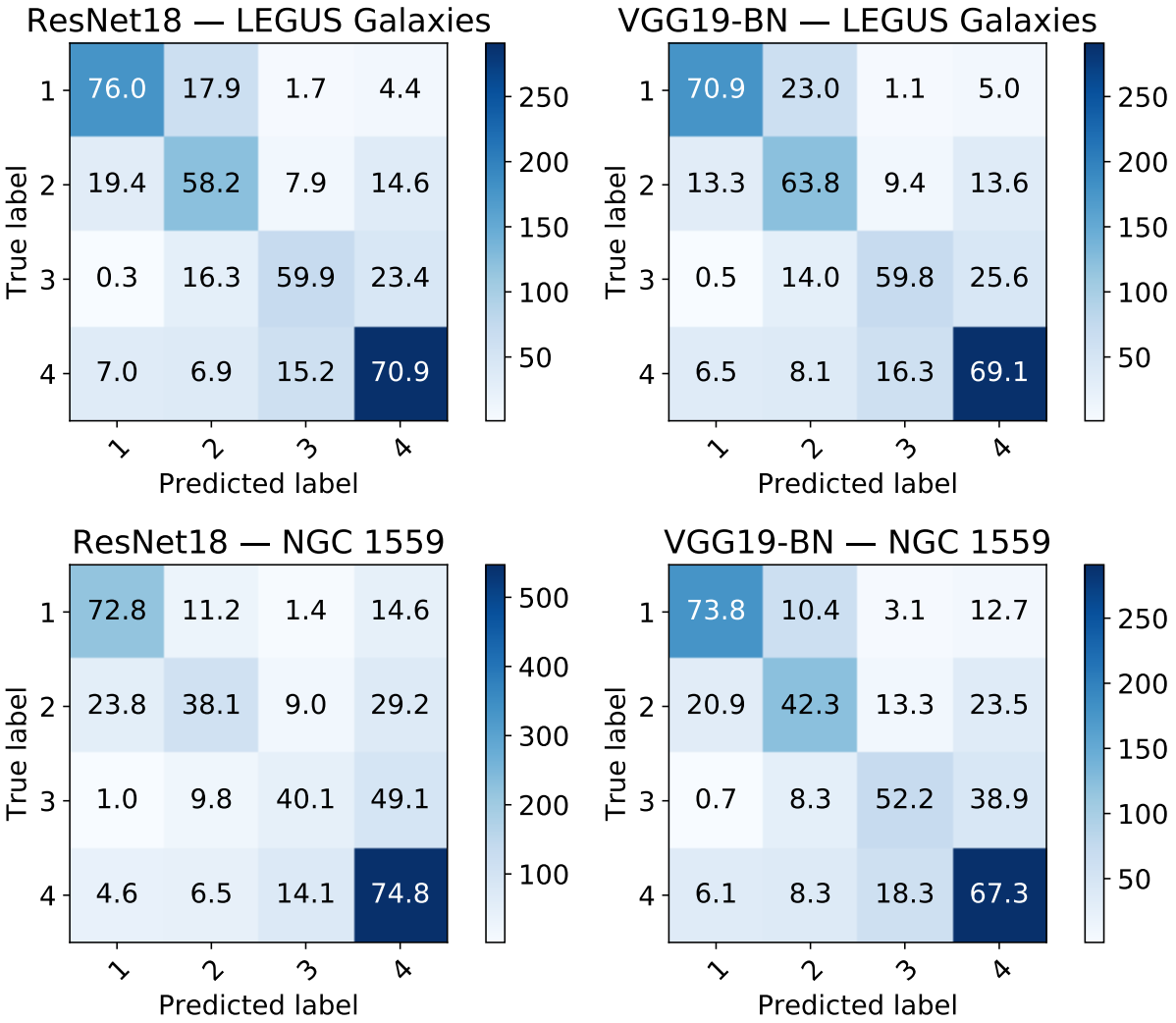

Figure 3. Top panels: prediction of `ResNet18` (left panel) and `VGG19-BN` (right panel) on data in Table 1, averaged over 10 models. Each row shows the averaged predictions for a particular class, and the number of ground truth images for that same class. Bottom lefts: as above panels but now using observations of spiral galaxy NGC 1559 obtained by the PHANGS-HST program. As before, results were obtained after averaging over 10 models. Note that in these confusion matrices each row corresponds to a predicted class, whereas each column corresponds to an actual class. Correct classification results are organized in a diagonal line from top left to bottom-right of the matrices.

	Class 1 [%]	Class 2 [%]	Class 3 [%]	Class 4 [%]	**Total**
Class 1	**76.0**± 4.2	17.9± 4.4	1.7±0.7	4.4±1.4	254
Class 2	19.4 ±3.5	**58.2**±5.3	7.9±3.5	14.6±3.0	202
Class 3	0.3 ±0.5	16.3±5.4	**59.9**±6.8	23.4±5.6	147
Class 4	7.0±2.1	6.9±2.9	15.2±3.1	**70.9**±4.8	425

Table 3. Prediction of `ResNet18` on data in Table 1, averaged over 10 models. Each row shows the averaged predictions and standard deviations for a particular class and the number of ground truth images for that same class. For reference, compare bold numbers across the diagonal with human classification results: 66% : 37% : 40% : 61%. The agreement reported here is higher than the agreement in human performance as determined in Section 2.1.

5 DISCUSSION & CONCLUSIONS

Using a homogeneous dataset of human-labeled star cluster images, we have leveraged a new generation of deep learning models for morphological classification of compact star clusters in nearby galaxies to distances of $\sim$ 20 Mpc. These results are very promising. Despite training with small and unbalanced datasets, we have demonstrated that deep learning can be successfully be used to automate classification of star cluster candidates identified in HST UV-optical imaging being obtained by PHANGS-HST. This is a milestone in the use of deep learning for this area of research, and progress

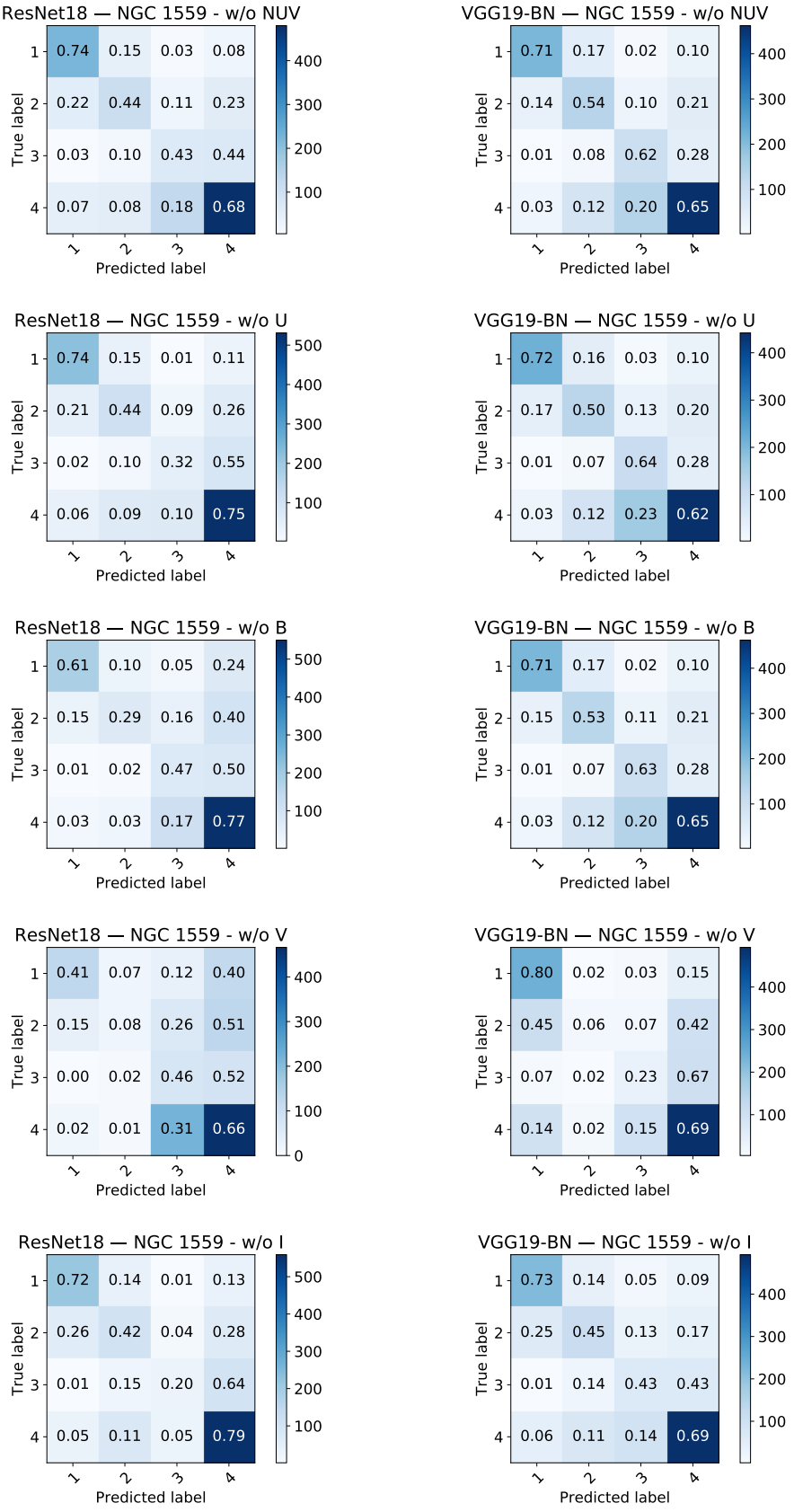

Figure 4. Left column: `ResNet` model classification results when the indicated filter is removed from the composite image. Right column: as before, but now for `VGG19-BN`.

	Class 1 [%]	Class 2 [%]	Class 3 [%]	Class 4 [%]	**Total**
Class 1	**70.9**±6.2	23.0±4.8	1.1±0.7	5.0±1.9	254
Class 2	13.3±4.3	**63.8**±4.8	9.4±2.9	13.6±3.6	202
Class 3	0.5±0.7	14.0±6.3	**59.8**±7.5	25.6±7.4	147
Class 4	6.5±2.4	8.1±2.6	16.3±3.8	**69.1**±6.8	425

Table 4. As Table 3, but now using `VGG19-BN` with batch normalization.

	Class 1 [%]	Class 2 [%]	Class 3 [%]	Class 4 [%]	**Total**
Class 1	**72.8**± 7.6	11.2± 3.8	1.4±0.6	14.6±5.1	302
Class 2	23.8±4.3	**38.1**±5.9	9.0±4.0	29.2±4.7	252
Class 3	1.0±0.5	9.8±4.2	**40.1**±7.1	49.1±6.1	162
Class 4	4.6±1.4	6.5±1.8	14.1±3.1	**74.8**±3.5	710

Table 5. Prediction of `ResNet18` on observations of spiral galaxy NGC 1559 obtained by the PHANGS-HST program, averaged over 10 models. Each row shows the averaged predictions and standard deviations for a particular class and the number of ground truth images for that same class. This experiment was performed to quantify the ability of this neural network model to generalize to new types of images, using a new, more distant PHANGS galaxy which is roughly twice as far away as any of the LEGUS galaxies in Table 1. We notice that even for this test case, deep learning outperforms human classification results as determined in Section 2.1: 66% : 37% : 40% : 61%.

	Class 1 [%]	Class 2 [%]	Class 3 [%]	Class 4 [%]	**Total**
Class 1	**73.8**±4.8	10.4±3.5	3.1±1.3	12.7±4.4	302
Class 2	20.9±6.4	**42.3**±7.9	13.3±2.6	23.5±8.0	252
Class 3	0.7±0.6	8.3±3.3	**52.2**±5.9	38.9±7.5	162
Class 4	6.1±2.4	8.3±3.3	18.3±3.0	**67.3**±6.8	710

Table 6. As Table 5, but now using `VGG19-BN` with batch normalization. As before, this neural network model outperforms human classification.

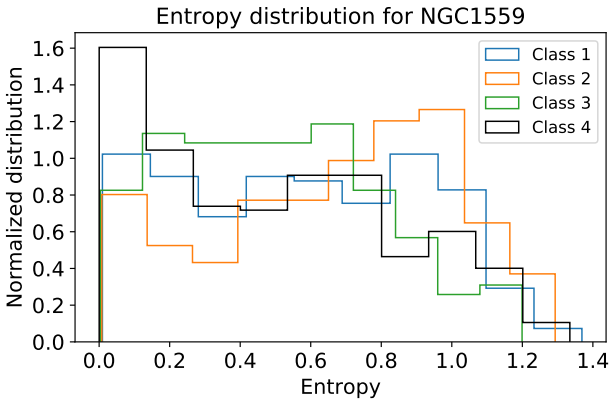

Figure 5. The uncertainty in our neural network's prediction is quantified by the entropy of the predicted probability mass distribution over the 4 classes. For a random guess over the 4 classes, the entropy is $\ln 4 \approx 1.39$. The lower the entropy, the higher the confidence the neural network has about its prediction. The figure shows the distribution of the entropies for the predictions made by one trained VGG19-BN model on NGC1559.

from early machine learning experiments reported in Messa et al. (2018). Messa et al. (2018) experimented with the use of an ML algorithm for classifying the approximately ten thousand clusters in the spiral galaxy M51, based on a human classified training set with approximately 2500 clusters from the LEGUS sample. While the recovery of class 1 and 2 clusters appears to be good (i.e., between 60 - 70% by inference from their Table 3), recovery of class 3 clusters is poor, with an apparently significant anti-correlation. Their ML algorithm only identifies 47 objects in the sample as class 3 when 1240 would be predicted based on the 14.7% fraction of class 3 objects in the human-labeled training set. Because of this, the Messa et al. (2018) Messa paper focused on class 1 & 2 objects for their analysis of cluster properties. In comparison, the recovery rates of the DL models presented here are at about 70%, 60%, 60%, and 70% for the class 1, 2, 3, 4 objects respectively.

We have conducted a detailed analysis of our results, using different neural network architecture, and multiple training sweeps for each model, to furnish evidence for the robustness of our results. We have also demonstrated that our deep learning algorithms can successfully classify star cluster candidate images from more distant galaxies not included in the training set.

In particular, our work motivates the development of a standardized dataset of human-labelled star cluster classifications, with classifications agreed upon by a range of experts in the field, which would be used as the basis for future network training. For this proof of concept experiment, we have opted to use classifications determined by a single expert (BCW). Our primary concern here is the internal consistency of the classifications, which is crucial for proper training, validation, and testing of the network, rather than whether the classifications are broadly agreed upon by experts in the community. The latter is of course a critical issue, but to make progress we have decoupled the issues for this study. A review of differences in star clus-

ter definitions between research groups, and their impact on conclusions about star cluster formation and evolution, can be found in Krumholz et al. (2018). To leverage upon deep learning techniques to not only rapidly produce reliable classifications and speed the time to science, but to significantly advance the field of star cluster evolution requires that deep learning networks be trained on such standardized datasets and broadly adopted by workers in the field. In the near future, this effort would benefit from a classification challenge, where experts can come to detailed agreement on the morphological features that constitute the criteria for classification, and explicitly describe where they disagree and why.

With this study we open a new chapter to explore in earnest the use of deep learning for the classification of very large datasets of star cluster galaxies in ongoing and future electromagnetic surveys, and application to the new PHANGS-HST data being obtained now.

ACKNOWLEDGEMENTS

Based on observations made with the NASA/ESA Hubble Space Telescope, obtained from the data archive at the Space Telescope Science Institute. STScI is operated by the Association of Universities for Research in Astronomy, Inc. under NASA contract NAS 5-26555. Support for Program number 15654 was provided through a grant from the STScI under NASA contract NAS5- 26555.

This research has made use of the NASA/IPAC Extragalactic Database (NED) which is operated by the Jet Propulsion Laboratory, California Institute of Technology, under contract with NASA.

This research is part of the Blue Waters sustained-petascale computing project, which is supported by the National Science Foundation (awards OCI-0725070 and ACI-1238993) and the State of Illinois. Blue Waters is a joint effort of the University of Illinois at Urbana-Champaign and its National Center for Supercomputing Applications.

This work utilized resources supported by the National Science Foundation's Major Research Instrumentation program, grant #1725729, as well as the University of Illinois at Urbana-Champaign.

We are grateful to NVIDIA for donating several Tesla P100 and V100 GPUs that we used for our analysis, and the NSF grants NSF-1550514, NSF-1659702 and TG-PHY160053.

This research used resources of the Argonne Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC02-06CH11357. We thank the NCSA Gravity Group for useful feedback.

MC and JMDK gratefully acknowledge funding from the Deutsche Forschungsgemeinschaft (DFG) through an Emmy Noether Research Group (grant number KR4801/1-1) and the DFG Sachbeihilfe (grant number KR4801/2-1). JMDK gratefully acknowledges funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme via the ERC Starting Grant MUSTANG (grant agreement number 714907)

REFERENCES

Abbott T., et al., 2016, Mon. Not. Roy. Astron. Soc., 460, 1270
Ackermann S., Schawinski K., Zhang C., Weigel A. K., Turp M. D., 2018, MNRAS, 479, 415
Adamo A., et al., 2017, ApJ, 841, 131
Ball N. M., Brunner R. J., Myers A. D., Tcheng D., 2006, ApJ, 650, 497
Ball N. M., Brunner R. J., Myers A. D., Strand N. E., Alberts S. L., Tcheng D., 2008, ApJ, 683, 12
Banerji M., et al., 2010, MNRAS, 406, 342
Barchi P. H., de Carvalho R. R., Rosa R. R., Sautter R., Soares-Santos M., Marques B. A. D., Clua E., 2019, arXiv e-prints, p. arXiv:1901.07047
Bastian N., et al., 2012, MNRAS, 419, 2606
Bengio Y., 2011, in Proceedings of the 2011 International Conference on Unsupervised and Transfer Learning workshop-Volume 27. pp 17–37
Bertin E., Arnouts S., 1996, A&AS, 117, 393
Calzetti D., et al., 2015, AJ, 149, 51
Cannon A. J., Pickering E. C., 1912, Annals of Harvard College Observatory, 56, 65
Cannon A. J., Pickering E. C., 1918, Annals of Harvard College Observatory, 91, 1
Carrasco Kind M., Brunner R. J., 2013, MNRAS, 432, 1483
Chandar R., et al., 2010, ApJ, 719, 966
Chandar R., Whitmore B. C., Calzetti D., O'Connell R., 2014, ApJ, 787, 17
Chandar R., Whitmore B. C., Dinino D., Kennicutt R. C., Chien L. H., Schinnerer E., Meidt S., 2016, ApJ, 824, 71
Cook D. O., et al., 2019, MNRAS, 484, 4897
Deng J., Dong W., Socher R., Li L.-J., Li K., Fei-Fei L., 2009, in CVPR09. http://www.image-net.org/
Domínguez Sánchez H., Huertas-Company et al., 2018, preprint, p. arXiv:1807.00807 (arXiv:1807.00807)
Everingham M., Eslami S. M. A., Van Gool L., Williams C. K. I., Winn J., Zisserman A., 2015, International Journal of Computer Vision, 111, 98
Fadely R., Hogg D. W., Willman B., 2012, ApJ, 760, 15
George D., Shen H., Huerta E. A., 2017, arXiv e-prints, p. arXiv:1711.07468
George D., Shen H., Huerta E. A., 2018, Phys. Rev. D, 97, 101501
Goodfellow I., Bengio Y., Courville A., 2016, Deep Learning. MIT Press
He K., Zhang X., Ren S., Sun J., 2015, arXiv e-prints, p. arXiv:1512.03385
He K., Zhang X., Ren S., Sun J., 2016, in Proceedings of the IEEE conference on computer vision and pattern recognition. pp 770–778, https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html
Holtzman J. A., et al., 1992, AJ, 103, 691
Hubble E. P., 1926, ApJ, 64, 321
Hubble E. P., 1936, Realm of the Nebulae
Ishak B., 2017, Contemporary Physics, 58, 99
Kamdar H., Turk M., Brunner R., 2016, Monthly Notices of the Royal Astronomical Society, 455, 642
Khan A., Huerta E. A., Wang S., Gruendl R., Jennings E., Zheng H., 2019, Phys. Lett., B795, 248
Kim E. J., Brunner R. J., 2017, MNRAS, 464, 4463
Kingma D. P., Ba J., 2014, Adam: A Method for Stochastic Optimization (arXiv:1412.6980)
Krizhevsky A., Sutskever I., Hinton G. E., 2012, in Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1. NIPS'12. Curran Associates Inc., USA, pp 1097–1105, http://dl.acm.org/citation.cfm?id=2999134.2999257
Krumholz M. R., McKee C. F., Bland -Hawthorn J., 2018, arXiv

e-prints, p. arXiv:1812.01615
LSST Science Collaboration et al., 2009, preprint (arXiv:0912.0201)
Larsen S. S., 2002, AJ, 124, 1393
LeCun Y., Bengio Y., Hinton G., 2015, nature, 521, 436
Małek K., et al 2013, A&A, 557, A16
Messa M., et al., 2018, MNRAS, 473, 996
Paszke A., et al., 2017, in NIPS-W.
Portegies Zwart S. F., McMillan S. L. W., Gieles M., 2010, ARA&A, 48, 431
Russakovsky O., et al., 2015, International Journal of Computer Vision (IJCV), 115, 211
Ryon J. E., et al., 2017, The Astrophysical Journal, 841, 92
Schweizer F., Miller B. W., Whitmore B. C., Fall S. M., 1996, AJ, 112, 1839
Sevilla-Noarbe I., Etayo-Sotos P., 2015, Astronomy and Computing, 11, 64
Shannon C. E., 1948, Bell system technical journal, 27, 379
Simonyan K., Zisserman A., 2014a, arXiv preprint arXiv:1409.1556
Simonyan K., Zisserman A., 2014b, arXiv e-prints, p. arXiv:1409.1556
Solarz A., Bilicki M., Gromadzki M., Pollo A., Durkalec A., Wypych M., 2017, A&A, 606, A39
Suchkov A. A., Hanisch R. J., Margon B., 2005, AJ, 130, 2439
Szegedy C., et al., 2014, arXiv e-prints, p. arXiv:1409.4842
Vasconcellos E. C., de Carvalho R. R., Gal R. R., LaBarbera F. L., Capelato H. V., Frago Campos Velho H., Trevisan M., Ruiz R. S. R., 2011, AJ, 141, 189
Weir N., Fayyad U. M., Djorgovski S., 1995, AJ, 109, 2401
Whitmore B. C., Sparks W. B., Lucas R. A., Macchetto F. D., Biretta J. A., 1995, ApJ, 454, L73
Whitmore B. C., et al., 2014, ApJ, 795, 156
de Vaucouleurs G., 1963, ApJS, 8, 31

APPENDIX A: STATISTICAL FOUNDATIONS OF DEEP LEARNING CLASSIFIERS

Within the framework of statistical learning, an image X can be modeled as a random matrix that takes value in set $\mathcal{X}$, and the corresponding class can be treated as a random variable Y that takes value in set $\mathcal{Y}$. Since we use 299×299 images with 5 channels, we treat a cluster image as random matrix of size $299 \times 299 \times 5$. Similarly, as we are trying to classify the images into 4 classes, Y is a discrete random variable that takes values in $\mathcal{Y}$ with cardinality $|\mathcal{Y}| = 4$.

We assume that the star images and the corresponding class labels follow some unknown but fixed joint probability distribution, with the probability density function (pdf) $f_{XY}x, y$. We also use $\Delta_{\mathcal{Y}}$ to denote set of all possible distribution over $\mathcal{Y}$. Since in our case, $|\mathcal{Y}| = 4$, we have $\Delta_{\mathcal{Y}} = \{\pi = \pi_1, \pi_2, \pi_3, \pi_4 : \sum_{i=1}^{4} \pi_i = 1, \pi_i \geq 0, \forall i \in 4\}$

Under these conventions, the goal of classification is to find a classifier or function $h : X \to \Delta_{\mathcal{Y}}$ that minimizes the expectation of the cross entropy between the predicted and the ground truth probability mass distribution (pmf) over the classes given the input image X, namely,

$$Lh = \mathbf{E}HhX, f_{Y|X}\cdot|X \tag{A1}$$

$$= HhX, f_{Y|X}\cdot|xf_X xdx\,, \tag{A2}$$

where $f_X x$ is the marginal distribution of X over $\mathcal{X}$, and H is the cross entropy between the predicted and the ground truth pmf over classes,

$$Hhx, f_{Y|Y}\cdot|x = -\sum_{i=1}^{4} f_{Y|X}Y = i|x \log hx_i, \tag{A3}$$

and the $f_{Y|X}y|x$ is the conditional distribution of Y given X.

In most cases, we only know the empirical distribution $\hat{f}_{XY}x, y$ of X, Y and $\hat{f}_{Y|X}y|x$ of Y, which are determined by the empirical data. So the quantity we can directly minimize is

$$\hat{L}h = \hat{\mathbf{E}}Hh_X\cdot, \hat{f}_{Y|X}\cdot|X \tag{A4}$$

$$= Hh_x\cdot, \hat{f}_{Y|X}\cdot|x\hat{f}_X xdx\,, \tag{A5}$$

In practice, if the choice of $h\cdot$ is arbitrary, then finding an optimal solution is computationally unfeasible. Therefore, we often restrict the searching space to a class of parameterized functions, $h_{\mathbf{w}}\cdot$, where $\mathbf{w}$ is a vector of parameters. In this case, the optimization problem can be posed as

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \hat{L}h_{\mathbf{w}}\,. \tag{A6}$$

The choice of the parameterized function class is critical to the success of any statistical learning algorithm. In recent years, a deep-layered structure of functions has received much attention (LeCun et al. 2015; Goodfellow et al. 2016),

$$h_{\mathbf{w}}\mathbf{x} = h_{\mathbf{w}_n}h_{\mathbf{w}_{n-1}}\cdots h_{\mathbf{w}_1}\mathbf{x}, \tag{A7}$$

where n is the number of layers or the depth. Usually, we choose, $h_{\mathbf{w}_i}\mathbf{x} = g\mathbf{w}_i\mathbf{x}$, where $\mathbf{w}_i$ is a matrix, $\mathbf{x}$ is an input vector, and $g\cdot$ is a fixed non-linear function, e.g., $\max\{\cdot, 0\}$ (also known as `ReLU`), $\tanh\cdot$, etc, that is applied element-wise. For the classification problems, we usually apply the so-called softmax function after the last linear transformation. The softmax function on a vector $\mathbf{x}$ is a normalization after an element-wise exponentiation,

$$\text{softmax}\mathbf{x}_i = \frac{\exp x_i}{\sum_{i=1}^{n} \exp x_i}, \qquad \forall i = 1, ..., n, \tag{A8}$$

where n is the length of $\mathbf{x}$.

This function class and its extensions, also dubbed neural networks, combined with simple first-order optimization algorithms such as stochastic gradient descent (SGD), and improved computing hardware, has lead to disruptive applications of deep learning (LeCun et al. 2015; Goodfellow et al. 2016).

APPENDIX B: DEEP TRANSFER LEARNING

In practice, Eq. A6 is usually iteratively solved by using variants of SGD. Thus, the choice of initial value for weights $\mathbf{w}$ is critical to the success of the training algorithm. If we have some prior knowledge about what initial wights $\mathbf{w}_0$ works better, then it is highly possible that the numerical iteration can converge faster and return better weights $\mathbf{w}$. This is the idea behind deep transfer learning (Bengio 2011; Goodfellow et al. 2016).

For a deep learning neural network, such as the one defined by Eq. A7, the layered structure can be intuitively interpreted as different levels of abstraction for the learned

features. In other words, layers that are close to the input learn lower-level features, such as different shapes and curves in the image, and layers that are close to the final output layer learn higher-level features, such as the type of the input image. Suppose we have a trained model that works well in one setting, with probability distribution f^1_{XY}, and now we would like to train another model in a different setting, with with probability distribution f^2_{XY}. If the images drawn from the distributions f^1_{XY} and f^2_{XY} share some features, then it is possible to transfer weights from the model trained on images sampled from f^1_{XY}, to the model that we would like to train, using images sampled from f^2_{XY}, with the assumption that the weights from the model trained on images sampled from f^1_{XY}, can also be useful in extracting features from images drawn from the distribution f^2_{XY}. So, instead of training the second model from scratch, we can initialize the weights of the second model to those of the first model that we trained in a different setting (e.g., distribution f^1_{XY}), and utilize the common features we have already learned in the previous setting.

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