Multi-Branch Attention Networks for Classifying Galaxy Clusters

Yu Zhang[1]*, Gongbo Liang[1], Yuanyuan Su[2], Nathan Jacobs[1]

1 Department of Computer Science, University of Kentucky, USA
2 Department of Physics & Astronomy, University of Kentucky, USA
Email: y.zhang@uky.edu*

***Abstract*—This paper addresses the task of classifying galaxy clusters, which are the largest known objects in the Universe. Galaxy clusters can be categorized as cool-core (CC), weak-cool-core (WCC), and non-cool-core (NCC), depending on their central cooling times. Traditional classification approaches used in astrophysics are inaccurate and rely on measuring surface brightness concentrations or central gas densities. In this work, we propose a multi-branch attention network that uses spatial attention to classify a given cluster. To evaluate our network, we use a database of simulated X-ray emissivity images, which contains 954 projections of 318 clusters. Experimental results show that our network outperforms several strong baseline methods and achieves a macro-averaged F1 score of 0.83. We highlight the value of our proposed spatial attention module through an ablation study.**

I. Introduction

Clusters of galaxies are the most massive collapsed objects in the cosmos. A large quantity of valuable information related to dark energy and dark matter is carried by galaxy clusters. With the rapid development of satellites, space telescopes can capture long-distance astronomical images that the unaided human eye could never capture. However, classifying these large scale images is still a challenging task, mainly because most of the useful information is concentrated in small central regions.

Galaxy clusters can be categorized into cool-core (CC), weak-cool-core (WCC) and non-cool-core (NCC) clusters based on their central cooling times [1]. The formation process of hierarchical structures shock-heats the intracluster medium (ICM) to to $10^7 - 10^8 K$, resulting radiating the emission of X-rays [2]. X-ray emissivity images of the ICM show that the central cooling times vary significantly between different clusters. Clusters with shorter cooling times ($t_{cool} \leq 1.0 h_{71}^{-1/2} Gyr$) are known as CC clusters. CCs have a systematic central temperature while temperature profiles of WCCs ($1.0 h_{71}^{-1/2} Gyr < t_{cool} < 7.7 h_{71}^{-1/2} Gyr$) are decreasing slightly towards the center or flat. NCCs are characterized as having highest core temperature in the center ($t_{cool} \geq 7.7 h_{71}^{-1/2} Gyr$) [3]. See Figure 1 for examples of simulated X-ray images of CC, WCC and NCC galaxy clusters in multiple dimensions.

Conventional methods in astrophysics for conducting this classification are through measuring other physical quantities. For example, central gas densities have been used in [2] [4] [5] to identify whether the galaxy cluster contains a cool core or not. The problem for this strategy is that for a modest exposure time, directly measuring gas densities in X-ray images is still a challenging task. Measuring X-ray surface brightness concentrations is another approach to identifying cool-core clusters [2] [6] [7], but this approach leads to inaccurate predictions of core types [8].

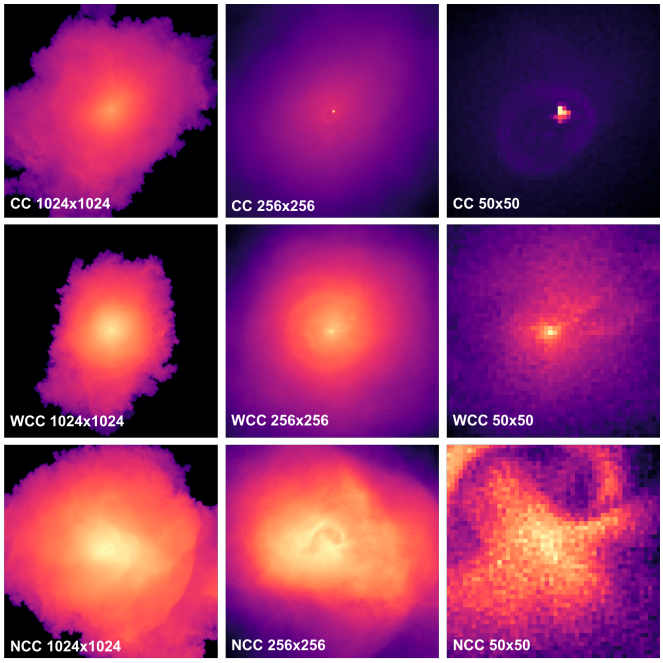

Fig. 1: Simulated X-ray emissivity images of cool-core (CC), weak-cool-core (WCC) and non-cool-core (NCC) galaxy clusters in 1024×1024 pixels (left column), 256×256 pixels (mid column), and 50×50 pixels (right column).

Deep learning has been applied in different vision tasks [9] [10] [11]. Naïve deep learning approaches for solving this problem would be taking entire images as inputs and predicting cluster types directly. However, this strategy is limited in that central cooling times are usually related to only small informative regions near the center, and feeding unrelated regions into the network may bring unnecessary noise and decrease the performance of the model.

In this work, we address the existing issues and propose

improved multi-branch attention networks that utilize attention and bivariate Gaussian distribution to identify the galaxy cluster type. We use ResNet-18 [12] as backbones for both primary branch and auxiliary branch in our architecture. The primary branch takes as input an original X-ray image and outputs a categorical distribution over a discrete label space. Considering the fact that unrelated regions in the image may be noisy for predicting the central cooling time, an attention module is attached at the end of the last residual block to guide the network to focus on the small region that is strongly relevant to the prediction. By taking advantage of the fact that the cooling time is more relevant to the central region in a galaxy cluster than outer regions, we utilize the bivariate Gaussian distribution to generate masks. Combining outputs from attention and Gaussian modules, we get a binary mask which can be used to crop the small region in the original image. The auxiliary branch takes as input a cropped region and outputs a distribution. We concatenate feature vectors from two branches together for the final prediction.

Our loss function is designed by encompassing our domain knowledge that the central cooling times of three different types of galaxy clusters (CC, WCC, NCC) vary continuously. We address this as an ordinal classification task. For this work, we propose to use cross entropy for classification and Cramer distance for regression. Our final loss incorporates both a classification and regression component.

The main contributions of this work are summarized as follows:

- introducing a new simulated X-ray Emission dataset.
- proposing a multi-branch attention network architecture for galaxy cluster classification.
- integrating attention and bivariate Gaussian distribution to generate core masks.
- designing loss functions by encompassing our domain knowledge.

II. Related Work

Our work builds upon previous works in several areas: machine learning in astronomy, visual attention mechanisms, X-ray emission from clusters of galaxies, and other similar works.

Machine Learning in Astronomy. Astronomy is undergoing a fast growth in data size and complexity. Machine learning approaches have become increasingly popular among astronomers and have been broadly applied in multiple tasks [8] [13]. Deep learning methods, especially the convolutional neural network (CNN) and its varieties, can achieve better performance than many conventional methods in image-related tasks and facilitate new discoveries. For example, in the task of predicting galaxy cluster X-ray masses, CNNs learn from a low resolution spatial distribution of photon counts and achieve higher accuracy compared to a more standard core-excised luminosity method [14]. For 2D photometric galaxy profile modelling, DeepLeGATo [15] is more accurate than GALFIT [16] and about 3000 times faster on GPU. In the task of improving galaxy morphologies, deep learning approaches show smaller offset and scatter than previous models trained with support vector machines [17].

Visual Attention Mechanisms. The galaxy cluster classification task needs to tell the relatively subtle differences between different cooling times. Typically, central cooling times are related to only small areas near the center. It is beneficial to induce the network to localize and focus on those small informative regions to minimize the negative effect of the noisy regions. Attention mechanisms are applied in the neural networks to help models learn significant features efficiently [18] [19] [20]. Unlike the convolution kernel that calculates over a local neighborhood region, the attention module can calculate the similarity of pixels in the whole feature map. Visual attention mechanisms are broadly applied in multiple vision tasks. Self-attention is used in improving the training stability and performance of Generative Adversarial Networks [21]. Non-local models utilize attention for video processing [22]. Residual attention is proposed for image classification [23]. Attention has also been applied in medical imaging tasks. For example, an attention guided model is proposed for the task of thorax disease classification on chest X-ray images [24]. In our work, we utilize the Class Activation Mapping [25] to allow the classification-trained CNN to both classify the image and localize the region that is most relevant to the central cooling time.

Spatial Transformer Networks The purpose of spatial transformer networks [26] is similar to our work. STN introduces a new learnable module, which can explicitly allow the spatial manipulation of data within the network, so neural networks can actively spatially transform feature maps to disentangle object pose and part deformation from texture and shape. However, STN is not designed for large images like galaxy cluster images, and it is difficult to train, especially when the target objects exist in such small central regions.

X-ray Emission from Clusters of Galaxies. Galaxy clusters are the largest known gravitationally bound structures in the Universe. Typically a regular cluster contains hundreds of galaxies spreading over a region with a diameter of roughly $10^{23}m$ [27]. A galaxy cluster can be hundreds to thousands times more massive than the Milky Way. The space between cluster member galaxies is suffused with hot gas, which is heated to temperatures of millions of Kelvin. This gas emits high-energy radiation so it can be studied with X-ray telescopes. X-ray emission was detected from the galaxy M87 in the center of the Virgo cluster in 1966 [28]. In 1971, X-ray sources were detected in the directions of the Perseus and Coma clusters by balloons or sounding rockets [29] [30]. These discoveries suggested that galaxy clusters might be bright X-ray sources, and the launch of the Uhuru X-ray astronomy satellite verified that this hypothesis was correct [31]. In 1999, the flagship X-ray telescope of NASA was launched with the goal of detecting X-ray emission from

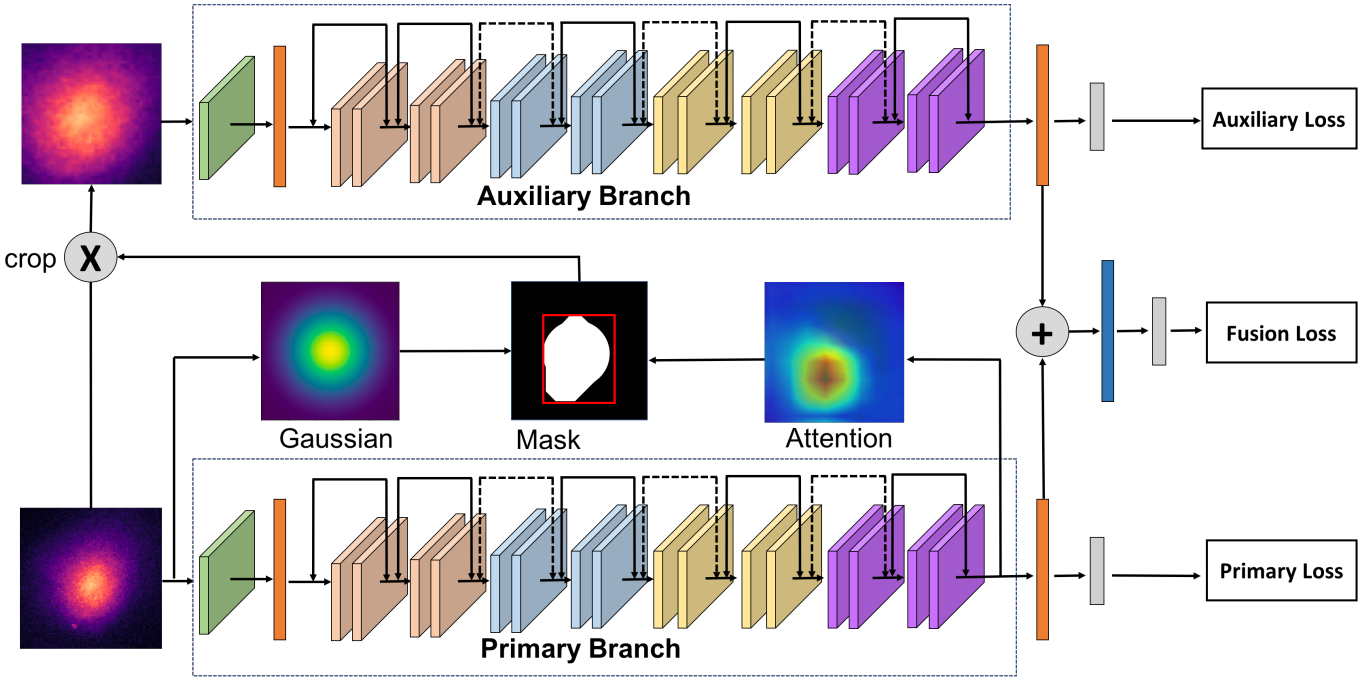

Fig. 2: Overview of our network architecture.

very hot regions of the Universe such as exploded stars, clusters of galaxies, and matter around black holes, and it provides massive high-quality X-ray images for astronomy researchers.

III. Approach

We propose multi-branch attention networks for galaxy cluster classification. We optimize the parameters of this model by minimizing a loss function that combines classification and regression component losses. We begin by outlining our base architecture, which is used in computing all component loss functions.

A. Architecture

Our proposed CNN architecture is shown in Figure 2. It takes as input a galaxy cluster X-ray image and outputs three categorical distributions over the same discrete label space. Both the primary and auxiliary branches are classification networks that predict the core type of the galaxy cluster, and consist of a portion of ResNet-18 architecture [12]. In the primary branch, ResNet architecture is used for high-level feature map extraction. We use the output of the last residual block, before global average pooling. We denote this output feature map as M, which is a 8×8 tensor with 512 channels. M is then used to generate an attention map. We also denote the feature vector extracted by the pooling layer as V_1, which is a 1×1 tensor with 512 channels. For the input image, we formulate a parametric function using bivariate Gaussian distribution. We create a binary mask by taking the union of the attention mask and the Gaussian mask. We use that binary mask to crop an informative region from the input image and pass it to the auxiliary branch for classification. In the auxiliary branch, we denote the output of the last global average pooling layer as V_2, which is also a 1×1 tensor with 512 channels. By concatenating V_1 and V_2, we get a new tensor with the size of $1 \times 1 \times 1024$, and this tensor is passed to a fully connected layer to predict a categorical distribution over 3 galaxy cluster classes.

B. Attention Map

We propose to construct attention maps to locate the most informative regions in the input images for galaxy cluster classification. By applying thresholds on attention maps, binary masks are constructed.

In this work, Class Activation Mapping [25] is performed to localize the discriminative regions used by the CNN to identify the cluster type. We compute a weighted sum of M to obtain a class activation map (CAM) for the input image. The activation of channel k in the feature map M at coordinate (x, y) is $m_k(x, y)$. The result of global average pooling for that unit is $M_k = \sum_{x,y} m_k(x, y)$. For a given class c, the input to the softmax is $S_c = \sum_k w_k^c M_k$, where w_k^c is the weight corresponding to class c for channel k. We obtain the class score

$$S_c = \sum_k w_k^c \sum_{x,y} f_k(x, y) = \sum_{x,y} \sum_k w_k^c f_k(x, y) \qquad (1)$$

and by summing the scores for location (x, y), we get the attention map

$$A_c(x, y) = \sum_k w_k^c f_k(x, y) \qquad (2)$$

where $A_c(x, y)$ is the value that represents the importance of the activation at (x, y) of the input image for class c. The final probability for class c is defined by

$$P_c = \frac{\exp(S_c)}{\sum_c \exp(S_c)} \tag{3}$$

The size of the attention map A we get is 8×8, and we resize that to 256×256 to fit the original size of the input image. We then normalize the attention map to $[0, 1]$ and get a heatmap H_1. To generate a binary mask B_1, we set the threshold to τ_1 ($0 \leq \tau_1 \leq 1$). Specifically,

$$B_1(x, y) = \begin{cases} 1, & H_1(x, y) \geq \tau_1 \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

C. Gaussian Mask

Considering the fact that the most informative regions for classification are close to centers, we formulate a parametric function by using bivariate Gaussian distribution to guide the model to focus on the central regions during training. For a given image with the size of $w \times h$, we use the following bivariate Gaussian distribution to sample the probability for the pixel at location (x, y) in the image:

$$p(x, y) = \frac{1}{2\pi\sigma_x\sigma_y} exp\left[-\left(\frac{(x-\mu_x)^2}{2\sigma_x^2} + \frac{(y-\mu_y)^2}{2\sigma_y^2}\right)\right]. \tag{5}$$

We use the coordinate of the brightest point in the input image as (μ_x, μ_y), which is the center of this distribution. The covariance matrix is $\sum = diag(\sigma_x^2, \sigma_y^2)$ since x and y are considered as independent variables. We represent σ_x and σ_y in form:

$$\sigma_x = \sqrt{\frac{w^2}{\lambda}}, \sigma_y = \sqrt{\frac{h^2}{\lambda}} \tag{6}$$

where λ is a hyper-parameter. To get the heatmap H_2, we then normalized the Gaussian mask to $[0, 1]$. Similar to the operation in attention map, we set the threshold to τ_2 ($0 \leq \tau_2 \leq 1$) to generate a binary mask B_2. Specifically,

$$B_2(x, y) = \begin{cases} 1, & H_2(x, y) \geq \tau_2 \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

The final binary mask is created by taking the union of B_1 and B_2. Specifically,

$$B(x, y) = \begin{cases} 0, & B_1(x, y) \cdot B_2(x, y) = 0 \\ 1, & \text{otherwise.} \end{cases} \tag{8}$$

We take the minimum and maximum coordinates in the binary mask B and get a bounding box $[x_{min}, y_{min}, x_{max}, y_{max}]$. The bounding box is used for cropping a small informative region in the original input image for the classification of the auxiliary branch.

D. Loss Function

Avoiding overfitting is always a key challenge in training large networks with relatively small datasets. In this work, we propose to use a combination of classification and regression component losses to train our model. The classification loss contains the primary loss, the auxiliary loss, and the fusion loss. The regression loss utilizes Cramer distance to take the order of CC, WCC, and NCC into consideration. The total loss function is:

$$L = \alpha_p L_p + \alpha_a L_a + \alpha_f L_f + \alpha_r L_r \tag{9}$$

In this section we are going to describe the different components of L in detail.

Classification Loss. The first three components L_p, L_a, and L_f, of our total loss, L, are classification loss and correspond to the main goal of our network: identifying the galaxy cluster core type of an input X-ray image. Each X-ray image is associated with a core type, $t \in \{1, 2, 3\}$. We apply our proposed network architecture with 3 outputs, representing a categorical distribution over CC, WCC and NCC. We denote the predicted distribution of the primary component, the auxiliary component and the fusion component as $\hat{y_p}$, $\hat{y_a}$ and $\hat{y_f}$, respectively. We use the weighted cross entropy loss between the predicted distribution and target distribution, y, so the classification loss functions of three components are represented as:

$$L_p = -\frac{1}{N}\sum_{i=1}^{N} w_{t_i} y_i(t_i) \log \hat{y_{p_i}}(t_i) \tag{10}$$

$$L_a = -\frac{1}{N}\sum_{i=1}^{N} w_{t_i} y_i(t_i) \log \hat{y_{a_i}}(t_i) \tag{11}$$

$$L_f = -\frac{1}{N}\sum_{i=1}^{N} w_{t_i} y_i(t_i) \log \hat{y_{f_i}}(t_i) \tag{12}$$

respectively, where N is the number of training examples, and w_t is the weight for class t, deployed to avoid poor fitting caused by the unbalanced distribution of labels. Specifically,

$$w_t = \frac{1}{\sqrt{count(t)}} \tag{13}$$

where $count(t)$ represents the number of training examples which are in class t.

Regression Loss. The last component, L_r, represents the regression loss. We propose to use the Cramer distance between $\hat{y_p}$ and y:

$$L_r = \frac{1}{N}\sum_{i=1}^{N} \|F(\hat{y_{p_i}}) - F(y_i)\|_2^2. \tag{14}$$

$F_X(x)$ is the cumulative distribution function (CDF) of x, representing the probability that the discrete random variable X takes on a value less than or equal to x. Specifically,

$$F_X(x_i) = P(X \leq x_i) = \sum_{x_i \leq x} p(x_i). \tag{15}$$

TABLE I: Evaluation results of our approaches trained on different settings vs. baseline.

Approach	Attention	Gaussian	Regression	macro-avg. f1	class	precision	recall	f1
Baseline	✗	✗	✗	0.803	CC	0.59	0.79	0.68
					WCC	0.92	0.85	0.88
					NCC	0.84	0.86	0.85
Ours(Att)	✓	✗	✗	0.823	CC	0.62	0.81	0.70
					WCC	0.93	0.86	0.89
					NCC	0.86	0.90	0.88
Ours(Gauss)	✗	✓	✗	0.813	CC	0.58	0.79	0.67
					WCC	0.93	0.85	0.89
					NCC	0.86	0.90	0.88
Ours(Att+Gauss)	✓	✓	✗	0.827	CC	0.67	0.79	0.73
					WCC	0.91	0.86	0.89
					NCC	0.84	0.88	0.86
Ours(all)	✓	✓	✓	**0.830**	CC	0.65	0.86	0.74
					WCC	0.94	0.85	0.89
					NCC	0.83	0.90	0.86

TABLE II: Hyper-parameter settings for loss function.

Method	α_p	α_a	α_f	α_r	macro-avg. f1
Baseline	0.8	0.1	0.1	–	0.803
Ours(all)	0.8	0.1	0.1	1	0.813
Ours(all)	0.8	0.1	0.1	100	0.817
Ours(all)	0.1	0.1	0.8	0.1	0.823
Ours(all)	0.1	0.8	0.1	0.1	0.820
Ours(all)	0.8	0.1	0.1	0.1	**0.830**

IV. Experiments

We evaluate our approaches on various metrics. Below we describe the datasets used for these experiments, and explore the performance of our model through an extensive analysis.

A. Datasets

The IllustrisTNG project [32] contains a large number of state-of-the-art cosmological magnetohydrodynamical simulations of the formation of galaxies. TNG300 is the simulation with the largest volume in IllustrisTNG [33] [34] [35] [36] [37]. We obtain a total number of 954 X-ray emissivity images from 318 maassive clusters in TNG300. These images can be categorized into three different classes based on their central cooling times. CC clusters are defined as $t_{cool} \leq 1.0Gyr$. WCC clusters are those with $1.0Gyr < t_{cool} < 7.7Gyr$. NCCs are characterized as having highest core temperatures in the center with $t_{cool} \geq 7.7Gyr$ [3].

B. Preprocessing

For each sample, we crop the central 256×256 pixels to get rid of a huge number of purely dark pixels in outer regions. Since the original values of pixels are extremely large, we scale pixel values smaller by taking the natural log. We then get the final ready-to-use images by normalizing all pixel values to $[0, 1]$.

To create cluster-level train/val/test splits for training and evaluating our network, we utilize a 10-fold cross validation strategy, and projections in the same cluster are always in the same fold. During the training process, we use 8 folds for training, 1 fold for validation and 1 fold for testing. We repeat this process until all folds are tested. We report the results tested on the entire 10 folds by using this cross validation strategy.

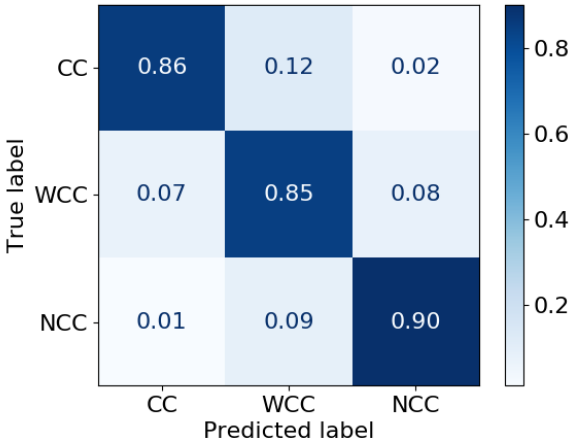

Fig. 3: The confusion matrix (row normalized) for our best method.

C. Implementation Details

Our model is developed by using PyTorch [38], and optimized using Adam [39] with a step learning rate decay strategy. We start to optimize the network with the learning rate 0.001, and decay the learning rate using the step size 10, by the factor 0.1. We initialize ResNet-18 [12] using ImageNet [40] pre-trained weights. We notice that using these pre-trained weights achieves significantly better performance than random initialization.

Based on our experiments, we notice that optimizing the total loss with $\alpha_p = 0.8$, $\alpha_a = 0.1$, $\alpha_f = 0.1$, and $\alpha_r = 0.1$ offers the best performance. In attention map generation, we set the hyper-parameter τ_1 to 0.7. For binary masks generation, we set the threshold to $\tau_2 = 0.7$, and set $\lambda = 25$.

D. Results

Our results are shown in Table I. In our experiments, ResNet-18 is used as the baseline method. We compare our

complete architecture with three variants as well as the baseline method. $Ours(Att)$ represents the method that only uses attention maps to generate the binary masks. $Ours(Gauss)$ shows the results obtained by using bivariate Gaussian distribution to generate binary masks. $Ours(Att + Gauss)$ represents that we generate binary masks by taking the union of Gaussian and attention. All methods mentioned above use cross entropy as their loss functions. From Table I we can notice that using the combination of attention and Gaussian methods achieves better performance than using only one of them. The last row in Table I, $Ours(all)$, shows that adding the regression component in the loss function can achieve the best performance, and get the highest macro-averaged F1 score (0.830) among all experiments. Table II shows results of various combinations of hyper-parameters for the total loss function.

Figure 3 shows the confusion matrix for our best method. The results are satisfactory, especially for CC clusters, when considering the fact that our training set is extremely imbalanced, and only about 10% of all samples are CC clusters.

V. Conclusion

Identifying various types of galaxy clusters using X-ray images is important in astronomy. We introduce a novel approach for identifying cool-core, weak-cool-core, and non-cool-core galaxy clusters. We demonstrate how a combination of an attention map and a bivariate Gaussian distribution helps crop an informative region from the input image for better classification performance. We design a loss function that encompasses the domain knowledge and utilizes both classification components and regression components. In several critical experiments, we demonstrate our proposed approach outperforms the baseline and many variant methods. We hope this work can be a baseline as well as a guideline for future classification research using large astronomical images.

References

[1] Y. Su, D. A. Buote, F. Gastaldello, and R. van Weeren, "Chandra observation of abell 1142: A cool-core cluster lacking a central brightest cluster galaxy?" *The Astrophysical Journal*, 2016.

[2] D. J. Barnes, M. Vogelsberger, R. Kannan, F. Marinacci, R. Weinberger, V. Springel, P. Torrey, A. Pillepich, D. Nelson, R. Pakmor *et al.*, "A census of cool-core galaxy clusters in illustristng," *Monthly Notices of the Royal Astronomical Society*, 2018.

[3] D. S. Hudson, R. Mittal, T. H. Reiprich, P. E. Nulsen, H. Andernach, and C. L. Sarazin, "What is a cool-core cluster? a detailed analysis of the cores of the x-ray flux-limited hiflugcs cluster sample," *Astronomy & Astrophysics*, 2010.

[4] Y. Su, R. Kraft, P. E. Nulsen, C. Jones, T. J. Maccarone, F. Mernier, L. Lovisari, A. Sheardown, S. Randall, E. Roediger *et al.*, "Extended x-ray study of m49: The frontier of the virgo cluster," *The Astronomical Journal*, 2019.

[5] L. Lovisari, T. Reiprich, and G. Schellenberger, "Scaling properties of a complete x-ray selected galaxy group sample," *Astronomy & Astrophysics*, 2015.

[6] J. S. Santos, P. Rosati, P. Tozzi, H. Böhringer, S. Ettori, and A. Bignamini, "Searching for cool core clusters at high redshift," *Astronomy & Astrophysics*, 2008.

[7] F. Andrade-Santos, C. Jones, W. R. Forman, L. Lovisari, A. Vikhlinin, R. J. Van Weeren, S. S. Murray, M. Arnaud, G. W. Pratt, J. Démoclès *et al.*, "The fraction of cool-core clusters in x-ray versus sz samples using chandra observations," *The Astrophysical Journal*, 2017.

[8] Y. Su, Y. Zhang, G. Liang, J. ZuHone, D. Barnes, N. Jacobs, M. Ntampaka, W. Forman, P. Nulsen, R. Kraft *et al.*, "A deep learning view of the census of galaxy clusters in illustristng," *Monthly Notices of the Royal Astronomical Society*, vol. 498, no. 4, pp. 5620–5628, 2020.

[9] T. Salem, S. Workman, and N. Jacobs, "Learning a dynamic map of visual appearance," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[10] Y. Zhang, G. Liang, T. Salem, and N. Jacobs, "Defense-pointnet: Protecting pointnet against adversarial attacks," in *2019 IEEE International Conference on Big Data (Big Data)*, 2019.

[11] G. Liang, X. Wang, Y. Zhang, X. Xing, H. Blanton, T. Salem, and N. Jacobs, "Joint 2d-3d breast cancer classification," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2019, pp. 692–696.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[13] D. Baron, "Machine learning in astronomy: A practical overview," *arXiv preprint arXiv:1904.07248*, 2019.

[14] M. Ntampaka, J. ZuHone, D. Eisenstein, D. Nagai, A. Vikhlinin, L. Hernquist, F. Marinacci, D. Nelson, R. Pakmor, A. Pillepich *et al.*, "A deep learning approach to galaxy cluster x-ray masses," *The Astrophysical Journal*, 2019.

[15] D. Tuccillo, M. Huertas-Company, E. Decencière, S. Velasco-Forero, H. Domínguez Sánchez, and P. Dimauro, "Deep learning for galaxy surface brightness profile fitting," *Monthly Notices of the Royal Astronomical Society*, 2018.

[16] C. Y. Peng, L. C. Ho, C. D. Impey, and H.-W. Rix, "Galfit: Detailed structural decomposition of galaxy images," *ascl*, 2011.

[17] H. Domínguez Sánchez, M. Huertas-Company, M. Bernardi, D. Tuccillo, and J. Fischer, "Improving galaxy morphologies for sdss with deep learning," *Monthly Notices of the Royal Astronomical Society*, 2018.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017.

[19] G. Lin, C. Shen, A. Van Den Hengel, and I. Reid, "Efficient piecewise training of deep structured models for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[20] T. Shen, T. Zhou, G. Long, J. Jiang, S. Pan, and C. Zhang, "Disan: Directional self-attention network for rnn/cnn-free language understanding," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[21] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *International Conference on Machine Learning*, 2019.

[22] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018.

[23] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

[24] Q. Guan, Y. Huang, Z. Zhong, Z. Zheng, L. Zheng, and Y. Yang, "Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification," *arXiv preprint arXiv:1801.09927*, 2018.

[25] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[26] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," in *Advances in neural information processing systems*, 2015.

[27] M. Wolf, "Über einen nebelfleck-haufen im perseus," *Astronomische Nachrichten*, 1906.

[28] E. Byram, T. Chubb, and H. Friedman, "Cosmic x-ray sources, galactic and extragalactic," *Science*, 1966.

[29] G. Fritz, A. Davidsen, J. F. Meekins, and H. Friedman, "Discovery of an x-ray source in perseus," *The Astrophysical Journal*, 1971.

[30] W. Forman, E. Kellogg, H. Gursky, H. Tananbaum, and R. Giacconi, "Observations of the extended x-ray sources in the perseus and coma clusters from uhuru," *The Astrophysical Journal*, 1972.

[31] R. Giacconi, S. Murray, H. Gursky, E. Kellogg, E. Schreier, and H. Tananbaum, "The uhuru catalog of x-ray sources." *The Astrophysical Journal*, 1972.

[32] D. Nelson, V. Springel, A. Pillepich, V. Rodriguez-Gomez, P. Torrey, S. Genel, M. Vogelsberger, R. Pakmor, F. Marinacci, R. Weinberger *et al.*, "The illustristng simulations: public data release," *Computational Astrophysics and Cosmology*, 2019.

[33] A. Pillepich, D. Nelson, L. Hernquist, V. Springel, R. Pakmor, P. Torrey, R. Weinberger, S. Genel, J. P. Naiman, F. Marinacci *et al.*, "First results from the illustristng simulations: the stellar mass content of groups and clusters of galaxies," *Monthly Notices of the Royal Astronomical Society*, 2018.

[34] V. Springel, R. Pakmor, A. Pillepich, R. Weinberger, D. Nelson, L. Hernquist, M. Vogelsberger, S. Genel, P. Torrey, F. Marinacci *et al.*, "First results from the illustristng simulations: matter and galaxy clustering," *Monthly Notices of the Royal Astronomical Society*, 2018.

[35] D. Nelson, A. Pillepich, V. Springel, R. Weinberger, L. Hernquist, R. Pakmor, S. Genel, P. Torrey, M. Vogelsberger, G. Kauffmann *et al.*, "First results from the illustristng simulations: the galaxy colour bimodality," *Monthly Notices of the Royal Astronomical Society*, 2018.

[36] J. P. Naiman, A. Pillepich, V. Springel, E. Ramirez-Ruiz, P. Torrey, M. Vogelsberger, R. Pakmor, D. Nelson, F. Marinacci, L. Hernquist *et al.*, "First results from the illustristng simulations: A tale of two elements–chemical evolution of magnesium and europium," *Monthly Notices of the Royal Astronomical Society*, 2018.

[37] F. Marinacci, M. Vogelsberger, R. Pakmor, P. Torrey, V. Springel, L. Hernquist, D. Nelson, R. Weinberger, A. Pillepich, J. Naiman *et al.*, "First results from the illustristng simulations: radio haloes and magnetic fields," *Monthly Notices of the Royal Astronomical Society*, 2018.

[38] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in neural information processing systems*, 2019.

[39] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Machine Learning (ICML)*, 2014.

[40] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009.