Using SCONCE SCMS on COSMOS2020 to find Cosmic Filaments

Seri Nakamura[a,b], Kiana Ejerito[a,b], Sina Taamoli[c], Alexander de la Vega[c]

[a] *University of Hawaii at Manoa, 2500 Campus Rd, Honolulu, HI 96822, USA*

[b] *Institute for Astronomy, University of Hawai'i, 2680 Woodlawn Drive, Honolulu, HI 96822, USA*

[c] *University of California Riverside, 900 University Ave, Riverside, CA 92521*

Abstract

Identifying filaments is computationally expensive and arbitrary. Machine learning techniques improve computation time and when trained successfully, produce promising results. In order to produce a machine learning algorithm that identifies filaments, the algorithm must be trained using a dataset of the base image and its mask identifying where the filament exists in the base image. This paper explores an algorithm SCONCE SCMS and determines whether the algorithm is capable of producing a mask of the filament for the machine learning algorithm's training dataset.

Keywords: , Machine learning, filaments, mask, SCONCE SCMS

1. Introduction

As understood through simulations and observations Libeskind et al. (2017), matter is distributed in the universe in a web-like structure, the cosmic web. Studies investigate the role of the effect of the environment a galaxy resides in on its evolution Taamoli et al. (2024). The observations described in Taamoli et al. (2024), has investigated the role of the environment in galaxy evolution and uses COSMOS2020 data to estimate environment densities and relate the findings to star formation rates (SFR). In studying the effect of a galaxy's environment on its star formation rate thing to consider is the correlation between filaments and its properties. In order to properly study this, filaments must be identified first. There are multiple methods of identification of cosmic filaments, including graphing techniques, statistical evaluations of geometric concepts, and geometric methods with Hessian matrices of density, tidal, or velocity fields Libeskind et al. (2017). Identification of filaments proves to be a difficult task with the arbitrary shape of filaments and the computationally inefficient methods that attempt to define filaments. Machine learning can be a useful aid to both filament identification issues. The data we will be using is surveys of galaxy points and density fields are unlabeled–pixels are not yet defined by whether it belongs to a filament or not–thus, an unsupervised machine learning algorithm is necessary. MaLeFiSenta: Machine Learning for FilamentS Identification and orientation in the ISM, Alina et al. (2022),

uses Mask R-CNN and a U-Net architecture for filament identification. The Mask R-CNN algorithm is trained on images that have been passed through the rolling hough transform (RHT), which identifies the shapes (filaments). This technique was also used on solar filaments; a Mask R-CNN was used to identify solar filaments in H alpha images Salasa and Arymurthy (2019). Both Alina et al. (2022) and Salasa and Arymurthy (2019) discuss the training process to use a mask of the filament within the data to train the Mask R-CNN. order to develop these filament masks, a non-machine learning technique will used: SCONCE SCMS Zhang et al. (2022). Spherical and Conic Cosmic Web Finders with the Extended SCMS Algorithm, SCONCE-SCMS, is a Python library for cosmic web filament detection. This algorithm uses the subspace-constrained meanshift (SCMS) to estimate density ridges, the structures of density peaks. For our purposes, we will be producing a Python script in order to use this algorithm on our fields.

In this project, we expand on the work of Taamoli et al. (2024), testing multiple methods of filament detection in order to produce masked datasets, correlating to the density maps produced by Taamoli et al. (2024). We classify the filaments in the galaxies to determine the relationship between the environment and SFRs at different redshifts.

2. Methodology

2.1. COSMOS2020 Catalog

Taamoli et al. (2024) produced density maps of galaxies in 135 redshift slices spanning $0.4 < z < 6$ using COSMOS2020 survey. COSMOS2020 survey includes approximately 1.7 million sources on a 2deg^2 COSMOS field, constructed by a profile-fitting photometric tool The Farmer Weaver et al. (2023) based on The Tractor Lang et al. (2016). The Tractor fits galaxy profiles to derive photometry, utilizing a decision tree to find the best-fit model. The final sample included 211,431 galaxies in a photo-z range of $0.4 < z < 6$ and have a magnitude cut of value $K_s < 24.5$.

From the COSMOS2020 catalog, the data is split into 135 redshift slices of constant comoving width $\Delta\chi = 35h^{-1}\text{Mpc}$ and calculate the weights, with a Gaussian probability distribution, for every galaxy in the redshift slice, which is noted as the probability of the galaxy being in the redshift slice. A density estimation around the masked regions starts with a mesh of uniformly distributed artificial sources. The optimal bandwidth in each redshift slice is then used to smooth the mesh according to the density field.

2.2. SCONCE SCMS

Spherical and Conic Cosmic Web Finder with the extended SCMS algorithms (SCONCE SCMS) Chen et al. (2015) Zhang et al. (2022) is a Python library for identifying cosmic filament and associated cosmic node structures from a collection of discrete observations. The subspace constrained mean shift (SCMS) algorithm is a modified mean shift, non-parametric

iterative algorithm that identifies density ridges in a density field. The algorithm uses Kernel Density Estimator (KDE) in correlation with SCMS algorithm to estimate filaments at density ridges. The KDE uses a weight that is calculated by multiplying the calculated likelihood the galaxy is within the redshift slice and the mass of the point. A smoothing bandwidth, provided with the data from Taamoli et al. (2024) is passed as parameters with the weight into the KDE in order for high performance on the estimator and density ridge filament model. For 2D coordinates and weighted data, a Directional-SCMS (DirSCMS) algorithm is used. In order to utilize the DirSCMS algorithm, the density field is calculated for the data. A threshold calculation is set for the accuracy of the ridge estimator and is used while producing the thresholded mesh of galaxy points for the DirSCMS algorithm. The initial mesh of uniformly distributed points is then sent through the KDE function with the threshold to get the mesh of the KDE. This mesh is then passed through the DirSCMS algorithm with the data to identify the convergence between the mesh on the data to identify the filaments. In order to more accurately identify density regions, the SCMS algorithm was developed to take into account data in the Celestial Sphere and the inherent spherical geometry correlating to the data. SCONCE SCMS converts the (RA, DEC) coordinates of the observation to the Cartesian representations on a unit sphere.

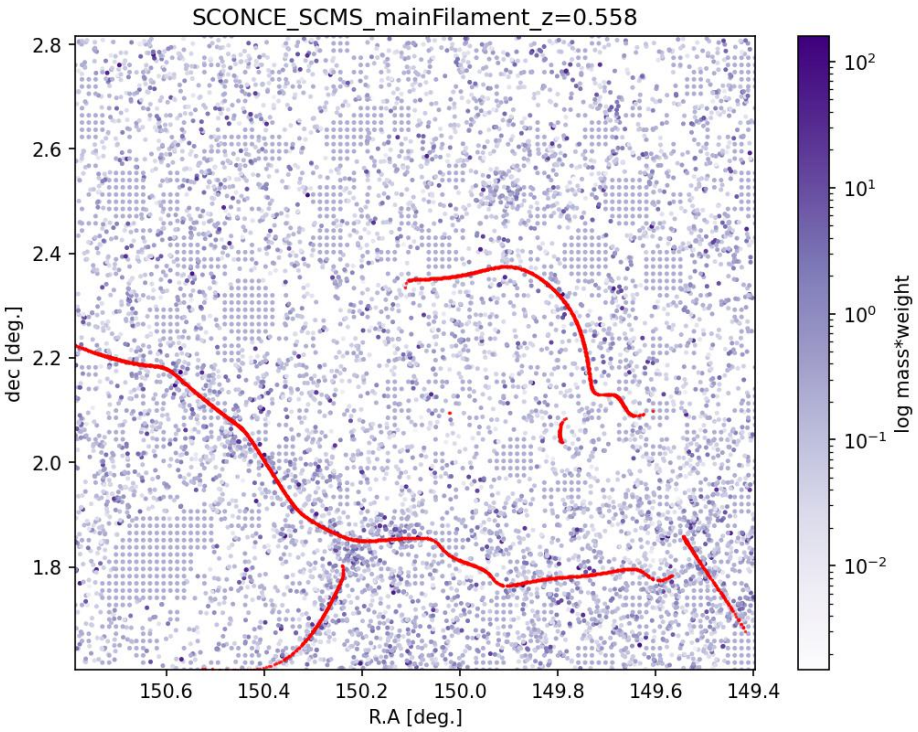

Figure 1: SCONCE SCMS produces the array of points that form the tracing of the filament as shown in red. This plot has the initial input of the galaxy points plotted in purple with a calculated log of the weight shown with a gradient.

2.3. Neural Network Training

In order to train the neural network for filament identification with the dataset, Taamoli et al. (2024), a mask must also exist of the filament in the image. This tracing, mask, of the filament is used to identify the parts of the original image that correspond to the filament in order to determine where the filament will be in new datasets. In order to train the neural network, a decent-sized training dataset must be established. The dataset from Taamoli et al. (2024) consists of 135 redshift slices, thus we will need to slightly augment these images in order to obtain a viable dataset. Augmenting includes rotations and reflections of the original dataset. Augmentation produces 7 new images, for a total of 8 images for each redshift slice.

3. Analysis

3.1. Comparisons

We compare SCONCE SCMS to the topological algorithm, Discrete Persistent Structures Extractor, DisPerSE `DisPerSE`, that uses Discrete Morse Theory (Forman, 2002), to identify filaments (Sousbie, 2011; Sousbie et al., 2011). Morse Theory relates the gradient of a field found from discrete data points, and the topology of the manifold and gradient to one another for smooth function.

Using a threshold of about a standard deviation above the median, the filaments found by the two algorithms are compared in 2. We can consider the identified filaments at this threshold to be the main filament. SCONCE SCMS, however, does not recognize the obvious filaments that DisPerSE maps.

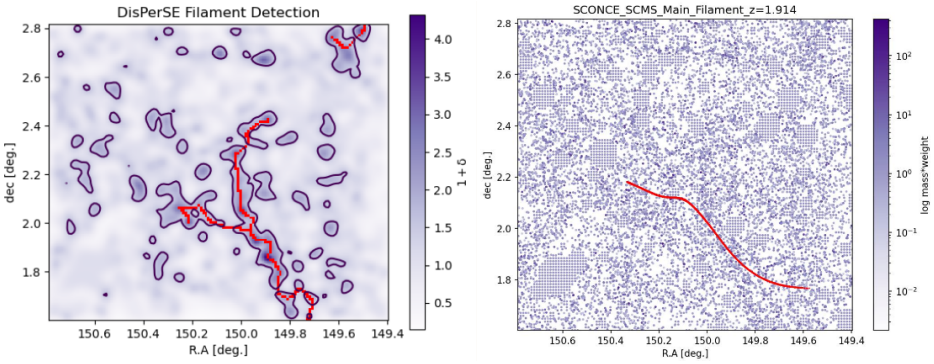

Figure 2: When the threshold is held to about a standard deviation over the median, a main filament can be identified. On the left, DisPerSE identifies a main filament with a branch stretching to higher declination and a small filament in the upper right corner whereas SCONCE SCMS, on the right, shows a single filament.

In producing a catalog that contains the main filament of the redshift slice in order to analyze the correlation between the environment's shape and the galaxy's SFR, DisPerSE proves to be a useful method of identifying the main filament to study this.

3.2. Establishing Training datasets

SCONCE SCMS was tested on all 135 redshift slices. The threshold is a tool that we use to manipulate in order to prepare a training dataset for the neural network. In lower

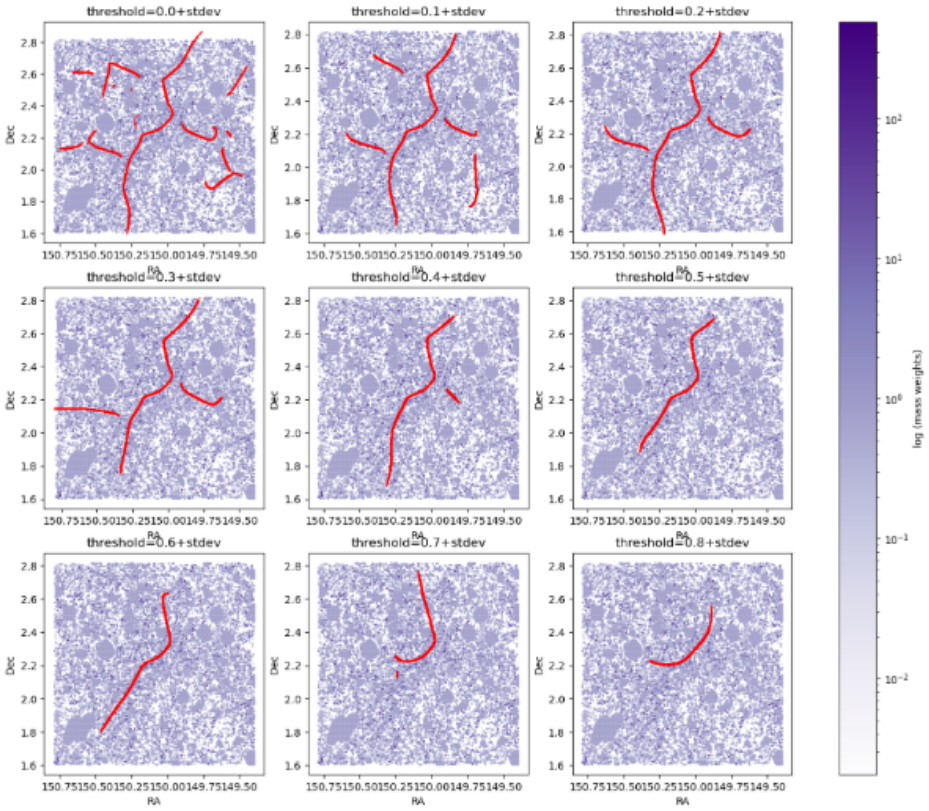

Figure 3: The above grid shows that more filaments are shaped with lower thresholds.

thresholds, SCONCE SCMS solves for more filaments within the redshift slice, however, it is unclear how we can determine a threshold in which SCONCE SCMS will find the general filaments tracing in the redshift slice for a training dataset without over estimating filament identifications. We can arbitrarily determine that the threshold of data points above a standard deviation of the data and this can be used to produce a dataset of general filament tracings.

4. Results and Discussion

SCONCE SCMS has proved capable of establishing a main filament, however, the missing filaments may be caused by the steps taken to decrease computation time: placing additional thresholds. Filament tracings have been produced for all 135 redshift slices, 25 of which the redshift slice doesn't have enough galaxy points to successfully identify a valid filament. With the application of augmentation, we have 1080 images, yet we will still be lacking in

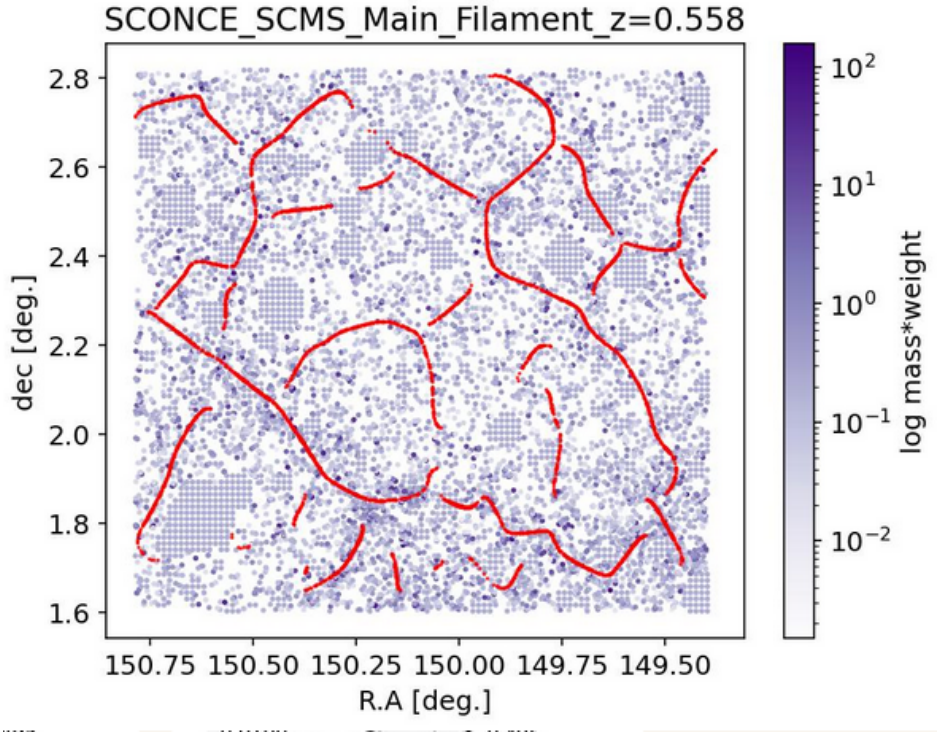

Figure 4: Identifying all the potential filaments in the field, despite the length,

training datasets. SCONCE SCMS does not perform as well as DisPerSE, 2 shows that SCONCE SCMS fails to find filaments that are multi-branched. SCONCE SCMS identifies filaments that follow a singular path, but is not capable of recognizing a complex structure, which will not be of use when creating tracings (masks) for the machine learning training dataset. As a result, SCONCE SCMS will not be used to create a training dataset for a neural network to identify filaments.

5. Conclusions

SCONCE SCMS shows a high unlikelihood of successful application in producing a training dataset to train the Mask R-CNN network. SCONCE SCMS also does not perform as well as the algorithm DisPerSE when finding the main filament(s). To further this study, a new plan can be utilized in place of SCONCE SCMS. In order to improve, we would most likely need to implement a two-level image segmentation process, where the machine learning algorithm first find the general outline of dense areas on survey data of galaxy points and the second part of the algorithm will identify the filament within the filament-segmented areas of the image.

Acknowledgements

We would like to acknowledge the NSF Grant to fund the California Hawaii Astronomy Mentorship Program. We would like to thank everyone who organized the program, including director Christopher Storfer, Grey Murphree for their guidance, Bahram Mobasher, and David Sanders.

References

D. Alina, A. Shomanov, and S. Baimukhametova. MaLeFiSenta: Machine Learning for FilamentS Identification and Orientation in the ISM. *IEEE Access*, 10:74472–74482, Jan. 2022. doi: 10.1109/ACCESS.2022.3189646.

Y.-C. Chen, S. Ho, P. E. Freeman, C. R. Genovese, and L. Wasserman. Cosmic web reconstruction through density ridges: method and algorithm. MNRAS, 454(1):1140–1156, Nov. 2015. doi: 10.1093/mnras/stv1996.

R. Forman. A user's guide to discrete morse theory. *Séminaire Lotharingien de Combinatoire [electronic only]*, 48:B48c–35, 2002.

D. Lang, D. W. Hogg, and D. Mykytyn. The Tractor: Probabilistic astronomical source detection and measurement. Astrophysics Source Code Library, record ascl:1604.008, Apr. 2016.

N. I. Libeskind, R. van de Weygaert, M. Cautun, B. Falck, E. Tempel, T. Abel, M. Alpaslan, M. A. Aragón-Calvo, J. E. Forero-Romero, R. Gonzalez, S. Gottlöber, O. Hahn, W. A. Hellwing, Y. Hoffman, B. J. T. Jones, F. Kitaura, A. Knebe, S. Manti, M. Neyrinck, S. E. Nuza, N. Padilla, E. Platen, N. Ramachandra, A. Robotham, E. Saar, S. Shandarin, M. Steinmetz, R. S. Stoica, T. Sousbie, and G. Yepes. Tracing the cosmic web. *Monthly Notices of the Royal Astronomical Society*, 473(1):1195–1217, Aug. 2017. ISSN 1365-2966. doi: 10.1093/mnras/stx1976. URL `http://dx.doi.org/10.1093/mnras/stx1976`.

R. P. Salasa and A. M. Arymurthy. Solar filament detection using mask r-cnn. In *2019 International Workshop on Big Data and Information Security (IWBIS)*, pages 67–71, 2019. doi: 10.1109/IWBIS.2019.8935810.

T. Sousbie. The persistent cosmic web and its filamentary structure - I. Theory and implementation. MNRAS, 414(1):350–383, June 2011. doi: 10.1111/j.1365-2966.2011.18394.x.

T. Sousbie, C. Pichon, and H. Kawahara. The persistent cosmic web and its filamentary structure - II. Illustrations. MNRAS, 414(1):384–403, June 2011. doi: 10.1111/j.1365-2966.2011.18395.x.

S. Taamoli, B. Mobasher, N. Chartab, B. Darvish, J. R. Weaver, S. Hemmati, C. M. Casey, Z. Sattari, G. Brammer, P. L. Capak, O. Ilbert, J. S. Kartaltepe, H. J. McCracken, A. Moneti, D. B. Sanders, N. Scoville, C. L. Steinhardt, and S. Toft. Large-scale Structures in COSMOS2020: Evolution of Star Formation Activity in Different Environments at 0.4 ¡ z ¡ 4. ApJ, 966(1):18, May 2024. doi: 10.3847/1538-4357/ad32c5.

J. R. Weaver, L. Zalesky, V. Kokorev, C. J. R. McPartland, N. Chartab, K. M. L. Gould, M. Shuntov, I. Davidzon, A. Faisst, N. Stickley, P. L. Capak, S. Toft, D. Masters, B. Mobasher, D. B. Sanders, O. B. Kauffmann, H. J. McCracken, O. Ilbert, G. Brammer, and A. Moneti. The Farmer: A Reproducible Profile-fitting Photometry Package for Deep Galaxy Surveys. ApJS, 269(1):20, Nov. 2023. doi: 10.3847/1538-4365/acf850.

Y. Zhang, R. S. de Souza, and Y.-C. Chen. ¡scp¿sconce¡/scp¿: a cosmic web finder for spherical and conic geometries. *Monthly Notices of the Royal Astronomical Society*, 517(1):1197–1217, Oct. 2022. ISSN 1365-2966. doi: 10.1093/mnras/stac2504. URL `http://dx.doi.org/10.1093/mnras/stac2504`.