

Feature-Relevant Data Reduction for In Situ Workflows

Will Fox¹, Matthew Wolf², Jeremy Logan³, Jong Youl Choi², Scott Klasky^{2,3}, Tahsin Kurc^{4,2}

¹Massachusetts Institute of Technology, Cambridge, MA, USA

²Oak Ridge National Laboratory, Oak Ridge, TN, USA

³University of Tennessee, Knoxville, TN, USA

⁴Stony Brook University, New York, NY, USA

Abstract—As the amount of data produced by HPC simulations continues to grow and I/O throughput fails to keep up, in situ data reduction is becoming an increasingly necessary component of HPC workflows. Application scientists, however, prefer to avoid reduction in order to preserve data fidelity for post-hoc analysis. In an attempt to compromise between data quality and data quantity, this work introduces the concept of feature-relevant compression. We explore two scientific datasets in an attempt to quantify the impacts of compression on features of interest by identifying such features and analyzing changes in their properties after compression. We find that it is indeed possible to compress simulation data in a lossy manner while preserving desired properties within a predetermined error rate. Additionally, we suggest that this error quantification could be applied as part of an in situ workflow to dynamically tune compression parameters during simulation, compressing aggressively when features are simple but preserving structure where data complexity increases. Future work should focus on implementation, extension to additional compression algorithms, and analysis of these techniques on quantities which are derived from original simulation data.

Index Terms—Data reduction, in situ, lossy compression, high performance computing, feature detection

I. INTRODUCTION

The amount of data produced by HPC simulations is growing incredibly quickly, in some cases exceeding petabytes in a single simulation run. Additionally, we are witnessing an increasing disparity between computation power and I/O throughput as simulations produce data more quickly than it can be moved to storage. There are two primary ways to combat this bottleneck—in situ analysis and data reduction. In situ analysis workflows interpret the simulation data before it is moved to storage, meaning less post-hoc analysis is needed. On the other hand, one can reduce the data output itself and thus alleviate the pressure on I/O services when moving to and from storage. While lossless compression algorithms do exist, in order to achieve the requisite compression levels ($>100\times$), this reduction increasingly demands the use of lossy algorithms.

One can achieve lossy reduction by temporal or spatial sampling – in other words, simply not writing out some of the data that is being generated. This is generally the first approach, but simple sampling (every 300th time step) does not offer any insight as to what features or scientifically-

relevant information might have been lost. In this work, we focus instead on a class of lossy compression algorithms that give us some leverage to enable more frequent temporal samples, but at a degraded accuracy of the saved data.

Previously, notable work has been done on development of lossy compression algorithms, both in general and for specific use cases. Algorithms such as SZ [1], ZFP [2], and MGARD [3] are finding new ways to bound, reduce, and assess [4] overall compression error, generally based on standard techniques like root-mean-squared (RMS) error or autocorrelation. On a separate front, significant effort has been applied to the area of automated feature detection using computer vision software. Researchers have developed very accurate algorithms for identifying nuclei in tissue samples [5]–[7], geological features for unmanned exploration vehicles [8], and facial recognition [9], among other applications. For scientific data, however, an under-explored area tends to be the intersection of the traditional data error metrics and errors in the emergent features of the data.

In this work, we seek to quantify the effects of compression with respect to scientific features of interest for some guiding examples. Such a capability would allow one to determine what compression level is acceptable for a data sample given certain restrictions on feature preservation. Specifically, we identify two scientific data sets, and then we define and extract features from the data after applying varied degrees of lossy compression in order to examine changes in feature properties. This is complicated by the fact that features and their properties tend to evolve in complex ways as one advances through a data set. From these metrics, we look towards assessing quality both of individual features and on an aggregate level. Here, the goal is to assess whether a regularly-applied in situ workflow element could serve as part of dynamic management of compression parameters in order to maximize data reduction/compression while also accounting for evolving feature complexity. This sort of dynamic compression would provide application scientists with a high-fidelity reconstruction when necessary but achieve higher compression ratios when possible.

II. METHODS

To explore the impact of lossy compression on data features, we selected two different sets of scientific images, one derived from a Tokamak fusion reactor simulation and the other containing microscopy images of tissue samples. We also selected a set of compression techniques that allow for variable degrees of lossy storage. To each of the images in these datasets, we applied compression at a range of different lossiness levels to generate a set from 100% accuracy to high relative error. After decompressing, we applied feature detection to the resulting image sets, and calculated a set of feature properties for each image in the set. Finally, we examined the impact of compression on the calculated feature properties. The following sections provide more detail about each of these steps.

Data Collection

Fusion data was generated by the simulation XGC [10], [11] which models behavior of magnetically confined plasma in a tokamak reactor. For each time step, we used visualizations of the derivative of potential variable from a cross-section in the reactor in order to identify features – regions of unusually high or low potential gradients – that the physicists identified as relevant. Using the set of rendered visualizations as the data set allowed us to have a common set of image-based feature recognition techniques between the fusion and cancer data. Note that the (lossy) compression is applied to the underlying binary data before the image is rendered.

The results shown here are also based in part upon data generated by the The Cancer Genome Atlas (TCGA) Research Network. TCGA provides an online repository for high-resolution images of cancerous tissue samples. For these images, identifying the darkly-colored nuclei were the features of greatest interest to our science collaborators.

Compression

Data from the above sources was then reduced using a lossy compression algorithm and a variety of parameters ranging from no compression to compression with very high per-element relative error. This data reduction was done using the algorithm SZ via the Python bindings for the Adaptable I/O System (ADIOS – v1.13.1). We varied the SZ relative error parameter from 0 (no compression) to 0.5. At the level of 0.5, features are obviously highly distorted or erased altogether. Original mesh data from the fusion simulation and pixel values from the cancer data were compressed before reconstructing images based on the new values.

Feature Detection

Image segmentation and feature detection was done with the Python computer vision library OpenCV [12]. The input was a data visualization in the form of an image (three-dimensional array of RGB pixel values). This image was converted to grayscale with a simple average across RGB channels. As evidenced by figures 1 and 2, features are generally closed surfaces with smooth edges, often in an oblong or elliptical

shape. For this reason, a gaussian blur was applied with a 5x5 pixel kernel in order to avoid overfitting before the image was thresholded so that only those pixels above a given value remained. Holes in features (defined as areas completely surrounded by feature pixels) were filled in to form a complete feature. A connected component analysis was then performed in order to filter out features below a predetermined minimum size (100 pixels). This procedure removes noise, leaving only the more significant features (high pixel-value regions of large size). Finally, the OpenCV function findContours() was applied in order to identify the perimeters of these remaining features, and ellipses were fit to these features using a bounding-box method. During this stage, properties of the features were calculated and saved for post-hoc analysis.

Calculating Properties

From the features which remained after image segmentation, we calculated a number of descriptive statistics or properties on individual features as well as for the image as a whole.

On an individual level, we considered both the original feature contours, or edges, as well as ellipses which were fitted to each contour. For the contours, we calculated total pixel area, contour perimeter, and contour centroid. For the ellipses, we calculated the major and minor axes lengths, center position, and angle of rotation.

On an aggregate level, we computed properties for each timestep of a fusion simulation or each 800x800 pixel tile of the cancer images as a whole. In these images, we calculated the total number of features, total area and perimeter of feature contours, total area and perimeter of fitted ellipses, difference in area between ellipses and contours, difference in perimeter between ellipses and contours, and the moment of inertia of the features, weighted by their areas.

III. RESULTS & DISCUSSION

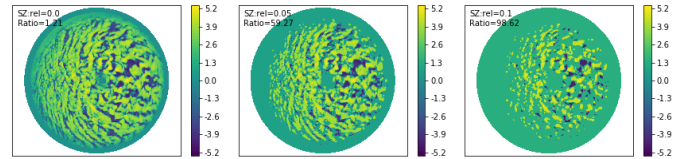


Fig. 1. An example timestep from XGC fusion simulation data under various compression parameters. Left image is uncompressed. Middle image is moderately compressed (about 60x). Right image is heavily compressed (about 100x). We can see the lossy compression affecting information content in the data.

The data set composed of all of the feature characteristics over different lossy compression values and over differing complexities of features is difficult to analyze succinctly. For the fusion data set, time is a useful proxy for the degree of complexity of the features, as the emergence of turbulence cases more and more fine-grained features to be tracked. The cancer data has no such easy external measure of complexity, but in a similar way the complexity of features can be directly measured from the uncompressed data and compared to the

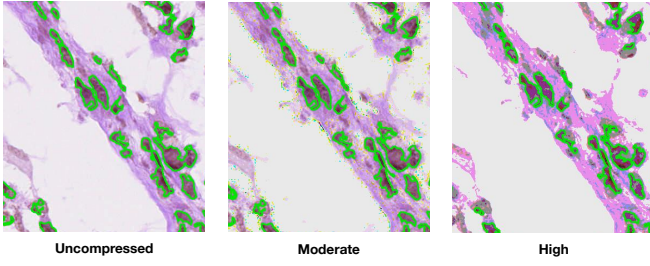


Fig. 2. Example TCGA tissue sample data under various compression parameters. Left image is the identified features in uncompressed data. Middle image is after moderate compression (about 50x). Right image is after heavy compression (about 200x). We can see the lossy compression affecting features, though the cancer images seem to be more resilient to compression than the fusion data. The green boundaries indicate features (contours) identified by the segmentation algorithm.

compressed values. We present in the following sections some views of the evolution of the trade-offs between compression error, compression value, and fidelity of feature capture as one or more of these are held fixed.

Properties vs. Compression

We began our search by examining how increasing compression affects features in both the fusion and cancer datasets for a given data complexity (though these results held generally across XGC timesteps and tissue sample tiles). We found that some amount of compression is almost always possible with very low feature distortion. Thus, for simulations with these types of small, smooth features, it may be advantageous to add a small but constant amount of compression for all timesteps and tiles. Take the compression 'SZ:rel=0.005' for example. We can see clearly in figures 3 and 4 that this compression has a very low error rate, minimally distorting features, while shrinking the data by a factor of approximately 30-40x in the fusion simulation. A slightly more aggressive approach such as 'SZ:rel=0.01' still has only a small impact on features present but allows for a compression ratio of about 60x for fusion data. This finding suggests that perhaps datasets with similar features should consider adding a small constant compression during analysis to improve efficiency without losing significant information content. It is unclear whether this recommendation extends to other types of features. In fact, we can even see that some datasets are more resilient and less predictable under compression by comparing the tissue sample images to fusion data in figure 4, suggesting that each dataset is unique and requires some amount of tuning and possibly domain expertise in order to identify compression quality.

As far as in situ workflows are concerned, our next key result was that, as expected, properties of features are altered by compression and that higher compression levels distort these properties more than lower compression levels. These effects of feature deterioration are evident in figures 3 and 4. Additionally, in figure 5, we see that this relationship generally persists over time in the fusion simulation. These trends held across a number of aggregate and individual properties including feature area, perimeter, and moment of inertia.

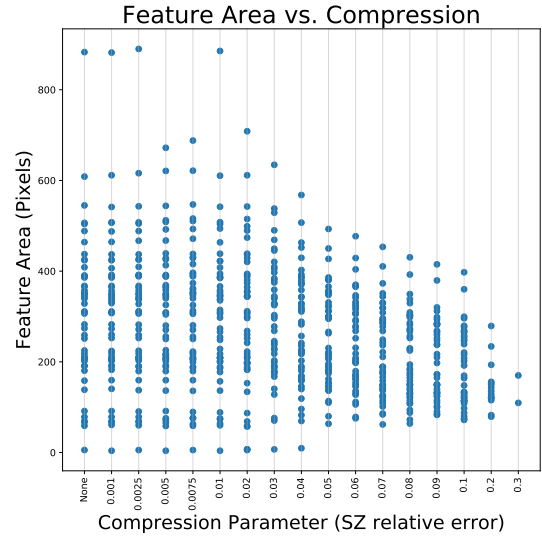


Fig. 3. Individual feature areas upon applying SZ compression with varying parameters (compression ratio increases to the right). Each data point represents the pixel area of a single identified feature. We see feature deterioration as compression increases. Shown for timestep 70 of an example XGC simulation.

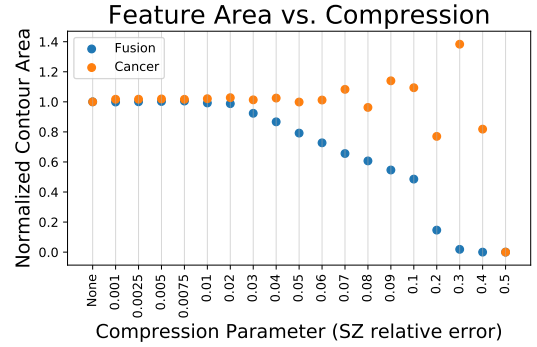


Fig. 4. Normalized feature area after applying SZ compression with varying parameters (compression ratio increases to the right). Features generally begin to disappear as the SZ parameter increases. Shown for timestep 70 of an example XGC simulation as well as a single 800x800 pixel tile of TCGA tissue sample data.

Though perhaps the least surprising result, predictable effects from compression are critical for the possibility of implementing feature-relevant reductions. The nearly monotonic deterioration of feature quality allows us to choose a strict cutoff parameter, above which image quality is unacceptable. In order to apply such a cutoff in a dynamic in situ algorithm, it is key that compression affects features in a generally predictable manner. We are thus able to efficiently search the space of possibilities in order to determine the optimal compression which preserves features within the desired error bounds.

Error vs. Time

Next, we looked to hold compression constant and explore the evolution of features in the fusion data, exploring feature property error against increasing complexity. As demonstrated in figures 6 and 7, the timestep value for fusion data is a good

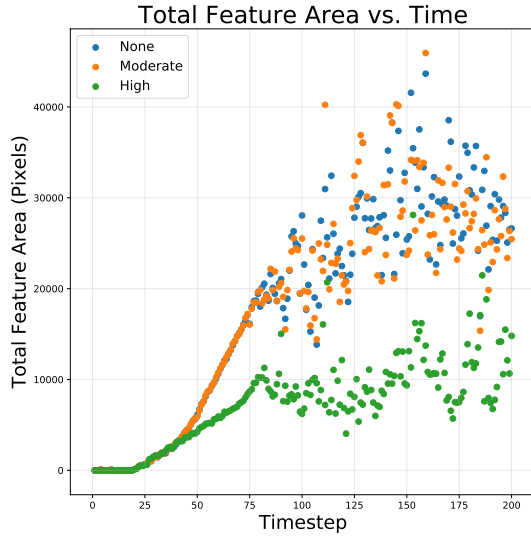


Fig. 5. Plot of the total feature area over time for various compression parameters, including no compression, moderate compression ($\text{rel}=0.01$), and high compression ($\text{rel}=0.1$) for XGC simulation run. Area increases smoothly in the beginning of the simulation, but becomes more variable as the simulation enters the more turbulent plasma phases.

proxy for the development of more complex feature structures. This dynamism means that compression levels which are acceptable at one point in a run (or at one level of complexity) may not be acceptable at a later one. In figure 6, we see the point where the p-value becomes very small and test statistic becomes large, indicating that compressed features are no longer representative of the original features as far as individual areas are concerned. We can use this statistical test as a proxy for compression acceptability and note the dramatic change near timestep 55. In figure 7, we see that the feature error of a given compression increases as the simulation progresses, and this pattern was found to hold generally across compression levels.

In the case of the fusion simulations, this data complexity is increasing with time as more features develop, allowing the possibility to start with a higher compression parameter and decrease to preserve features as they begin to appear. In the tissue sample dataset, a relatively small number of tiles contain most of the cells from the sample and there is no such smooth pattern as exists with the fusion dataset. However, these images also afford the possibility for very high compression on the regions which contain little to no interesting information (the blank tiles) while still preserving those regions which contain features.

In the context of an in situ application of feature-relevant reduction, this data variability is critical. If feature complexity was unchanging over the course of a simulation, the solution would simply be to apply constant compression. As this is not the case, we see that more consideration must be given to the state of the simulation data at a given timestep in order to compress without losing key information. Thus, a dynamic algorithm to characterize data and decide on an acceptable compression parameter could prove fruitful in managing the

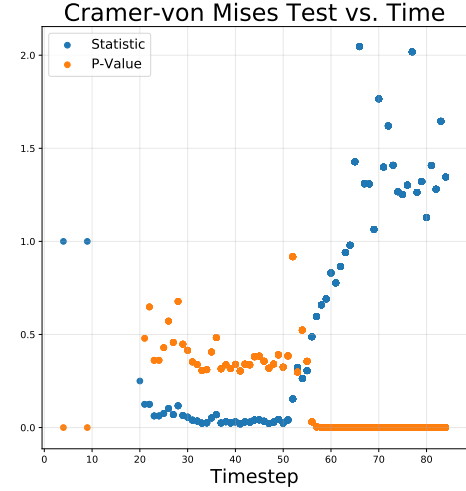


Fig. 6. Results of the Cramer-von Mises test to determine whether two sets of points are drawn from different distributions—in this case, the original individual feature areas versus those from the compressed image ($\text{rel}=0.1$). A small p-value and large statistic indicate that the compressed data does not match the original features for a given property. Data from XGC fusion simulation.



Fig. 7. Relative error (as compared to original, uncompressed data) for the specified properties over the course of an XGC simulation run compressed with SZ ($\text{rel}=0.1$).

growing amounts of simulation data while maintaining data fidelity.

Compression vs. Time

Finally, we explored how feature error is related to compression, examined again considering variation in data complexity. We can see from figure 7 that error does not necessarily evolve in the same way for all feature properties. In the fusion data with the given compression level, for example, error on the number of features begins to plateau around timestep 50 while

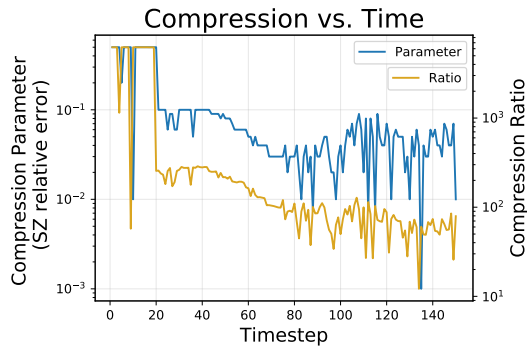


Fig. 8. XGC Simulation highest acceptable compression for 10% error on total feature area. We can compress aggressively early in the simulation but less so when the simulation becomes turbulent and variable.

error on area and perimeter continue to climb. This finding, along with the lack of a strong definition for a feature, pose a dilemma for application scientists: picking which feature property or linear combination of properties they most want to preserve in the dataset.

While it may be possible under low compression to retain all or nearly all properties with high fidelity, more aggressive compression requires domain specific knowledge to discriminate between those properties which are important to future analysis and those which are less so. Further, it hints that feature-relevant compression cannot be implemented in an in situ workflow simply by tracking number of features or total area. The procedure, instead, must be informed by the scientific question at hand with an eye on possible post-hoc analysis and input from application scientists.

Figure 8 represents a hypothetical case where an oracle could have predicted exactly what compression parameter would have been sufficient to maintain a 10% error in the total area feature characteristic. Using the fusion simulation timestep again as a proxy for the evolution of more complex feature structures, one can see that the acceptable compression parameter value varies over two orders of magnitude as complexity increases, with a decidedly non-monotonic functional form. In corresponding fashion, this means that the total data size could have been reduced to less than .1% of its original size at low complexity of features, with compression size with the most complexity of no worse than 10% of original.

More generally, this procedure could be used to find appropriate parameter values for other error bounds, for other feature characteristics, or even for some linear combination of characteristic values. Such curves would then stand as references for particular online or in situ systems that would need to make value choices without access to a perfect oracle.

Future Work

This work sought to demonstrate the validity and potential benefit of feature-relevant data reduction for a field where compression is becoming increasingly necessary. However, future work is required to expand on the ideas presented here. Specifically, we would like to expand the current analysis to include more compression algorithms, including ZFP and

MGARD, as well as new datasets including those based on derived quantities. Each algorithm is unique in its implementation, so exploring the types of features which are preserved could prove fruitful for achieving the optimal balance between compression quality and quantity. These additions would enable decisions on both compression ratio and algorithm, potentially enhancing performance.

Because it unclear whether the techniques outlined in this work will directly apply to other algorithms or datasets without modification, additional research should examine not only new derived quantities but new fields altogether. Generalization of the techniques introduced here are required in order to enable deployment across a broader spectrum of application codes. Emphasis should be placed on reducing the need for domain- and analysis-specific knowledge when defining features and error-bounds.

While further research on new compression techniques and unique data is important, continuing effort is needed to ultimately implement a dynamic reduction algorithm. Key analysis should focus on computational cost, exploring how to implement the quality calculation most efficiently and how frequently this data characterization needs to be done throughout a simulation run. Work on implementation promises significant storage and I/O savings as HPC enters the exascale era.

IV. CONCLUSION

In this work, we sought to explore the possibility of feature-relevant data reduction, specifically looking toward application as part of an in situ workflow. We examined whether it was possible to quantify the quality of compressed data with respect to features of scientific interest, resulting in several key findings. First, we discovered that it is indeed possible to use lossy compression methods and retain very high feature fidelity for the fusion and cancer datasets. As expected, we also found that increasing the compression parameter increases the feature distortion. Additionally, in these datasets, feature complexity varied across both space and time, demonstrating a need for dynamic adjustment of compression parameters tuned to the data at hand. Finally, we found that as the complexity of the feature structures evolves, the expressed values of the different feature characteristics evolved in distinct ways. This highlights the complex nature of the problem at hand and the need for domain- and analysis-specific knowledge when choosing compression parameters. In all, the results clearly indicate that the quantification of feature complexity is both possible and worthwhile to implement the required data reductions for the future of HPC simulations. Implementing such an algorithm to characterize data and choose optimal compression parameters accordingly would be a fruitful path for continuing work.

ACKNOWLEDGMENT

This research was supported in part by an appointment to the Oak Ridge National Laboratory HERE Program, sponsored by the U.S. Department of Energy and administered by the Oak Ridge Institute for Science and Education.

REFERENCES

- [1] S. Di and F. Cappello, "Fast error-bounded lossy hpc data compression with sz," in *2016 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, May 2016, pp. 730–739.
- [2] P. Lindstrom, "Fixed-rate compressed floating-point arrays," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 2674–2683, Dec 2014.
- [3] M. Ainsworth, S. Klasky, and B. Whitney, "Compression using lossless decimation: Analysis and application," *SIAM J. Scientific Computing*, vol. 39, 2017.
- [4] D. Tao, S. Di, H. Guo, Z. Chen, and F. Cappello, "Z-checker: A framework for assessing lossy compression of scientific data," *CoRR*, vol. abs/1707.09320, 2017. [Online]. Available: <http://arxiv.org/abs/1707.09320>
- [5] J. De Vylder, J. Aelterman, T. Lepez, M. Vandewoestyne, K. Douterloigne, D. Deforce, and W. Philips, "A novel dictionary based computer vision method for the detection of cell nuclei," *PLOS ONE*, vol. 8, no. 1, pp. 1–9, 01 2013. [Online]. Available: <https://doi.org/10.1371/journal.pone.0054068>
- [6] J. J. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, and H. J. Aerts, "Computational radiomics system to decode the radiographic phenotype," *Cancer Research*, vol. 77, no. 21, pp. e104–e107, 2017. [Online]. Available: <http://cancerres.aacrjournals.org/content/77/21/e104>
- [7] S. Sornapudi, R. Stanley, W. Stoecker, H. Almubarak, R. Long, S. Antani, G. Thoma, R. Zuna, and S. Frazier, "Deep Learning Nuclei Detection in Digitized Histology Images by Superpixels," *Journal of Pathology Informatics*, vol. 9, no. 1, p. 5, 2018. [Online]. Available: <http://www.jpathinformatics.org/article.asp?issn=2153-3539;year=2018;volume=9;issue=1;spage=5;epage=5;aulast=Sornapudi;t=6>
- [8] D. Thompson, S. Niekum, T. Smith, and D. Wettergreen, "Automatic detection and classification of features of geologic interest," in *2005 IEEE Aerospace Conference*, March 2005, pp. 366–377.
- [9] B. S. Manjunath, C. Shekhar, and R. Chellappa, "A new approach to image feature detection with applications," *Pattern Recognition*, vol. 29, no. 4, pp. 627–640, Apr 1996. [Online]. Available: <https://vision.ece.ucsb.edu/sites/vision.ece.ucsb.edu/files/publications/96PRManjunath.pdf>
- [10] C. Chang and S. Ku, "Spontaneous rotation sources in a quiescent tokamak edge plasma," *Physics of Plasmas (1994-present)*, vol. 15, no. 6, p. 062510, 2008.
- [11] S. Ku, C. Chang, and P. Diamond, "Full-f gyrokinetic particle simulation of centrally heated global itg turbulence from magnetic axis to edge pedestal top in a realistic tokamak geometry," *Nuclear Fusion*, vol. 49, no. 11, p. 115021, 2009.
- [12] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.