2023

Towards Explainable AI Using Attribution Methods and Image Segmentation

Garrett J. Rocks
University of Central Florida

TOWARDS EXPLAINABLE AI USING ATTRIBUTION METHODS AND IMAGE SEGMENTATION

by

Garrett J. Rocks
University of Central Florida

A thesis submitted in partial fulfillment of the requirements
for the Honors in the Major Program in Electrical Engineering
in the College of Engineering and Computer Science
and in the Burnett Honors College
at the University of Central Florida
Orlando, Florida

Fall Term, 2023

Thesis Chair: Rickard Ewetz Ph.D.

ABSTRACT

With artificial intelligence (AI) becoming ubiquitous in a broad range of application domains, the opacity of deep learning models remains an obstacle to adaptation within safety-critical systems. Explainable AI (XAI) aims to build trust in AI systems by revealing important inner mechanisms of what has been treated as a black box by human users. This thesis specifically aims to improve the transparency and trustworthiness of deep learning algorithms by combining attribution methods with image segmentation methods. This thesis has the potential to improve the trust and acceptance of AI systems, leading to more responsible and ethical AI applications. An exploratory algorithm called ESAX is introduced and shows greater performance on PIC testing than other top attribution methods in some cases. These results lay a foundation for future work in segmentation attribution.

ACKNOWLEDGEMENTS

I would like to thank Dr. Rickard Ewetz for the teaching and guidance provided to me in the completion of this research, as well as Dr. Hao Zheng. I would also like to acknowledge the support provided by Lockheed Martin Corp. as well as the Florida High Tech Corridor.

CONTENTS

FIGURES

TABLES

INTRODUCTION

Among the many applications of machine learning, from medical diagnosis to autonomous cars and military systems, a problematic pattern has been emerging. The performance of neural networks and machine learning continues to skyrocket. Across the board industries seek to apply ML concepts to anything quantifiable, and researchers pour their efforts into propelling the field forward as quickly as possible. Neural networks get deeper and more accurate by the year and are able to outperform humans on a number of tasks. As they become more complex however, they are more and more opaque to even their designers.

There are many different aspects of neural network performance. Some are quantifiable, such as accuracy and speed, while others remain more qualitative, like interpretability. Among even the quantifiable characteristics like accuracy, there are varying levels of completeness of data available. For example, Holzinger et al. stated that avid proponents of AI in the medical field stress the that attaining an artificial intelligence which would replace a doctor entirely is exceedingly difficult. This is because the number of factors which contribute to a diagnosis of a certain kind of cancer for example is very wide, and it is impossible with most traditional AI techniques to know whether the right variables are being considered in the diagnosis. As such, a doctor or medical professional would still need to be on site and engaged in the process of diagnosis.

The incompleteness of the medical example above is crucial in understanding the need for explainable AI. For example, aircraft collision detection has been operating fine for decades without explainability. This because the problem is well understood and can be defined precisely

[Bukart and Huber 2011]. The limited number of variables associated with the problem of aircraft collision detection can be enumerated. Due to the extreme complexity of many real world systems and patterns which AI can be applied to, it is essential for human understandability of these AI to grow along with their capacities, and for trust of such AI to be not only built, but validated.

This paper introduces ESAX, a technique that builds on existing AI attribution methods, incorporating additional heuristics to enhance segmentation accuracy. By comparing among the results generated by ESAX and with other existing algorithms, insights into attribution by segmentation were gathered and a path forward for future work was paved.

BACKGROUND

The first neural network known as the single-layer perceptron dates to the 1950's. The single-layer perceptron can recognize very basic shapes. An article from The New York Times cited the navy in 1958 as saying that the perceptron will lead to development which will eventually end in a computer which is capable of self-consciousness, which current day GPT-4 appears closer to than ever. In the 1960's the single-layer perceptron was proved to be incapable of learning other relatively simple fundamental patterns, such as the XOR function. In the 1980's the multilayer perceptron was invented, but also quickly hit a wall in its capacity to learn more complex problems, The chief obstacle being that of the vanishing gradient problem. Rumelhart et al. (1986) showed how backpropagating errors can be used to learn patterns in neural networks. The vanishing gradient problem appears when using backpropagation to train a multi-level neural network. As the number of layers in a perceptron increases, the gradient rapidly approaches zero which makes training high level perceptrons infeasible. This problem caused a near 20-year gap in major improvements to neural networks between the 1980's to 2006, when Hiton et al. (2006) showed that Restricted Boltzmann Machine initialization could be used to offset the vanishing gradient problem. Since then, deep learning has been used to great effect, with recent RNN (Recurrent Neural Network) architectures achieving over 100 layers while avoiding the vanishing gradient.

A Convolutional Neural Network uses several hidden layers in order to ascertain deeper and more complex relationships between features in the input than that which a simple perceptron or a fully connected multi-layer perceptron is capable of. CNNs have a broad range of

use cases, but they are particularly common and powerful when used for object recognition/classification.

There are several insights which motivate the use of CNNs for classification and object recognition tasks. On the input image, there are patterns which can be used to identify the image. These patterns are smaller than the entire image, therefore some kind of smaller window view of the image can be used to detect patterns. These patterns can appear anywhere on the image, so the subsampling method will need to be applied to every area on the image to find the patterns. The main patterns of an image are also generally such that the image can be subsampled or degraded in resolution in a particular way while the patterns remain intact. These factors contribute to the architecture of CNNs.

CNNs use a kernel to convolve over the image in the convolution layer. The kernel can be thought of as the "window" which a CNN uses to view only a small part of an image at a time, in order to extract the patterns smaller than the whole image as discussed previously. The convolution of the kernel with the input is fed into an activation function, in the case of Figure 1 this is the Rectified Linear Unit (ReLU) function. The output of the activation function is then fed into a pooling function which is the down-sampling of resolution as discussed previously, and various functions are used in this step depending on the application, sometimes it is omitted entirely as in the notable case of AlphaGo.

The hidden layers of a CNN are called hidden for a reason, and after a relatively small amount of complexity has been reached by a network, its decisions and inner workings are completely opaque even to the designer of the network. This is where the field of Explainable Artificial Intelligence (XAI) enters the scene. XAI aims to shed light on what exactly the hidden

layers of a CNN are doing. Attribution is the act of determining the effect that something has on the output of a network. Attribution is commonly applied on at least 3 different levels of analysis: the feature level, the layer level, and the neuron level. Performing attribution at these different levels would correspond to explaining how a specific feature of the input, an individual layer of the network, or a single neuron of the network, respectively, affect the output or the "decision" of the network.

Attribution is the process of assigning a score to an element of the input according to the element's effect on the output relative to a baseline. The use of a baseline is crucial because an element's effect on the NN output cannot be determined relative to nothing. Any explanation presupposes a counterfactual situation which the true explanation is relative to.

Neural networks are often used to aid in the task of classification, or the separation of things into categories based on the similarities of their attributes. CNNs seem particularly well suited to this task, because of their capacity to differentiate objects on a feature-by-feature basis, due to their multi-level construction.

During the training process of CNNs, a technique called gradient descent is used with backpropagation to optimize the weight parameters of the network. The gradients of the weights with respect to the loss function are essentially indicative of the importance of those weights in the accuracy of the network's inference. Using methods to tap into these gradients is one of the primary ways of obtaining information about the inner workings of an AI. Model explainability is the extent to which the parameters of a network can justify the results given by the network. Model interpretability is the ability of a network to correctly draw a relationship between a cause and an effect [1].

PREVIOUS WORK

Using various gradient based techniques, a variety of different XAI methods have been developed over the years. Being a field still in its infancy, XAI has experienced multiple revolutions, including several paradigm shifts in the approaches used to generate saliency maps, from gradient based, to perturbation based and post processing techniques.

Saliency maps in general are tool used in computer vision consisting of an image with different elements highlighted to show where attention is drawn. These are used in applications of image compression and quality, and studies of human attention. These maps are crucial in XAI, not for understanding human attention but for understanding the "attention" of a classification model. For our purposes, a saliency map is a vector generally in the same shape as a given input, which contains information about the relevance of each feature in the input vector to the classifying model.

Among the first attempts to visualize the innerworkings of a deep network advanced enough to be considered a black box is from Erhan et al. in 2009. The authors show how deep architectures had advanced in the context of vision datasets, but that there was a dearth of qualitative analysis of the models used, which was necessary to motivate and inform further development in the field of XAI. To address this emerging need, the idea of maximizing the activation of a hidden unit within the model. The thought process behind this is that if, for example, an image was found that maximized the activation of a single neuron, then that image

would be a meaningful and human understandable representation of what said neuron is doing. An interesting insight gained while testing the activation maximization technique in this paper is

In Simoyan et al. a technique for class model visualization was developed. In the technique, an input image was found that maximized the neuron activation using gradient descent to reveal the most salient input. This technique was extended in the same paper to image-specific class saliency visualization, which is a clear precursor to more current methods of attribution, producing results which are similar in appearance to Integrated Gradients.

Integrated gradients (IG) is a state-of-the art method used by researchers. The fundamental process that IG uses is to multiply the difference between the input and a baseline by the gradient of the weights with respect to the loss function. The IG function is shown below.

$$\mathsf{IntegratedGrads}_i(x) ::= (x_i - x_i') \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} \, d\alpha$$

Figure 1: Integrated Gradients Formula

The original equation cannot be implemented due to the integral, so a summation is used instead in practice to receive an approximate solution. The importance of a baseline as discussed earlier is apparent in this equation, where the x_i' in the equation is the given baseline. In IG, multiple baselines are used, and their results are combined in order to see the best explanation of the network's decision. In the standard implementation of IG, a black baseline is used, and the input image is interpolated over the black baseline over a series of dozens of steps, with the IG

algorithm being run at each step, see Figure 2.

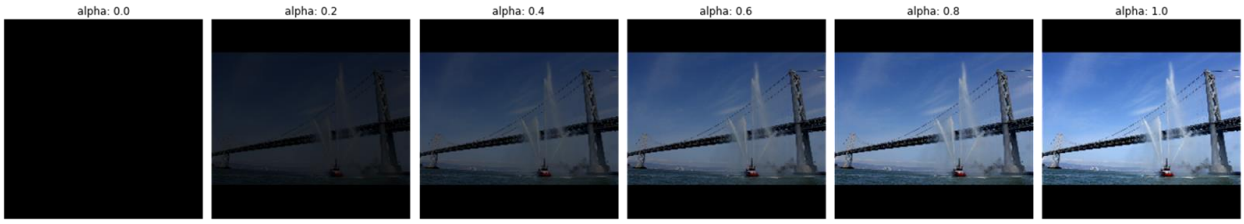


Figure 2: IG Black Baseline Interpolation

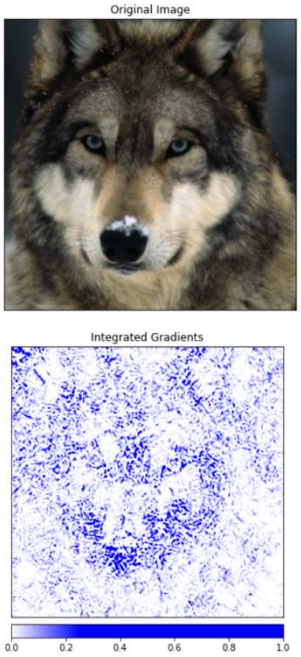


Figure 3 IG Attribution

As shown on the left timber wolf example of figure X, IG can produce a very noisy output. Certain parts of the image are being given good attribution, but exactly what the AI is looking at is difficult to ascertain from this map. Other researchers have tried to create new ways of using IG to get a more understandable explanation of the AI's "thought process"; enter XRAI.

XRAI's main addition to the scene of explainable AI is using a segmentation algorithm to divide up an image into many segments, and then calculate the weight of each segment instead of every individual pixel. This is thought to create a much more human understandable map of the attribution. When compared to the vanilla IG results, the XRAI heatmap is generally better understood and can be used to create a mask of the original image which reveals highly salient regions of the image, see Figure 4. For example, in the case of a timber wolf image, the XRAI heatmap can be used to identify the upper and middle parts of the wolf's face as highly salient regions that contribute to the network's prediction, and it is subjectively very easy to see what regions are relevant.

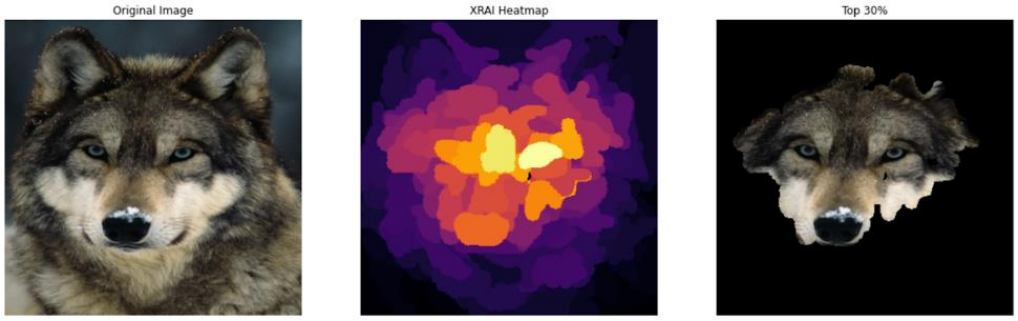


Figure 4: XRAI Heatmap

Evaluation of XAI Methods

There are qualitative and quantitative methods for evaluating XAI techniques. Qualitative methods generally require some human input, where opinions are gathered on outputs for different XAI methods. According to the PyTorch Captum website, explanation infidelity is a measure of the mean-squared error between the explanation multiplied by a perturbation function, and the difference between the prediction of the original input and the perturbed input. Captum describes Sensitivity-n, on the other hand, as a function which correlates the attribution to differences of the predictor between an input and a baseline.

Sensitivity is a direct measure of the change in the XAI output due to a perturbation added to the input. In this case, the perturbations are intentionally very small. Sensitivity-n is a property proposed in the paper “Towards better understanding of gradient-based attribution methods for deep neural networks” by Ancona et al.

Two methods for evaluation are proposed by the creators of XRAI in their original paper: Accuracy Information Curve (AIC) and SoftMax Information Curve, SIC. The authors state that

these approaches are similar to receiver operating characteristic (ROC). “A receiver operating characteristics (ROC) graph is a technique for visualizing, organizing and selecting classifiers based on their performance.” [1] The graphs are a plot of the true positive rate over the false positive rate. ROC is commonly used for evaluating the performance of classifiers, as opposed to a simple measure of classification accuracy, which can be misleading in the case of unbalances classes.

AIC and SIC, collectively known as Performance Information Curves (PIC) are inspired by the “bokeh” effect in photography, where the subject of the image is shown in focus, but the rest of the image is blurred. Likewise, PIC applies a blur to the input image, decreasing the information content of the image, and then a selective focusing occurs to the image according to the output of the saliency model. In other words, the highly salient sections of the original input are shown in focus, while the rest of the image remains out of focus. This focusing occurs in steps. After the information content of the image was brought down to a low level by the blurring, the step-wise re-focusing once again increases the information content. At each step of increased focus, the entropy of the image is approximated as the compressed image size, and recorded along with both the softmax output and the model accuracy. The accuracy (AIC) and softmax output (SIC) are then averaged at every step of re-focusing across many images and are plotted over the level of focus. The area under this curve is taken as the performance of the saliency method.

Methodology

The Exploratory Segmentation for Attribution based on XRAI (ESAX) methodology was to closely follow the algorithm of XRAI and change certain variables associated with segmentation. The XRAI authors stated that they aimed to avoid reliance on any specific hyperparameter opting to over sample the image to the point that any scale or sampling hyperparameter's effect on the attribution would be lost. ESAX aims to explore the effect of the segmentation size parameters on the attribution, as well as the method of assigning value to each segment. Whether or not dilation of the segments makes a significant difference is tested as well. Additionally, Felzenszwalb segmentation as well as multiple other segmentation techniques are explored to determine the effect that the shape and nature of the segments truly has on the attribution of the algorithm.

The main idea of ESAX is to investigate how combining contours and attributions can lead to better understandability of the attributions, and to improve objective measures of attribution quality. The algorithm follows XRAI in most of its high level steps but adds a variation of segmentation techniques instead of over-segmenting using the Felzenszwalb technique. Segmentation techniques have multiple parameters which determine the number, shape, and size of segments. The authors of XRAI state that they do not want the attribution results to depend on a particular set of such parameters, so to remove the dependance on these parameters they use Felzenszwalb's method which generally creates segments appealing to the human eye. They then over segment the image by a factor of 6, only later selecting the segments

with the highest attribution method to be added to the final segmentation to achieve a total area covering 100% of the original attribution shape.

This approach is called into question, and different methods of segmentation are explored with ESAX to determine the validity of XRAI's approach. To test this, four different segmentation methods are used with no over-segmentation to remove the potential effects that this may have. In future work over-segmentation should be explored as well to see how different segmentation methods interact with this effect. The four segmentation methods explored in this paper are Quickshift, Watershed, Slic, and the original Felzenszwalb. These are all implemented in the Skimage Python package.

ESAX begins by taking an input image and processing it to fit the size of the classification model. The processed image is classified and checked against the groundtruth to ensure proper classification. This step is taken to ensure that there is no biasing introduced by using incorrectly classified images, as the attribution given will not accurately reflect the relevance of features. Next, a segmentation algorithm is applied to the processed images that returns an array in the same shape as the image containing numbered segments. The image is fed into an attribution method, and the attributions are returned in the same shape as the original image. Then the attribution within each segment is averaged across the segment, and each element in the segment is set to the average attribution value of said segment. This is the final ESAX attribution.

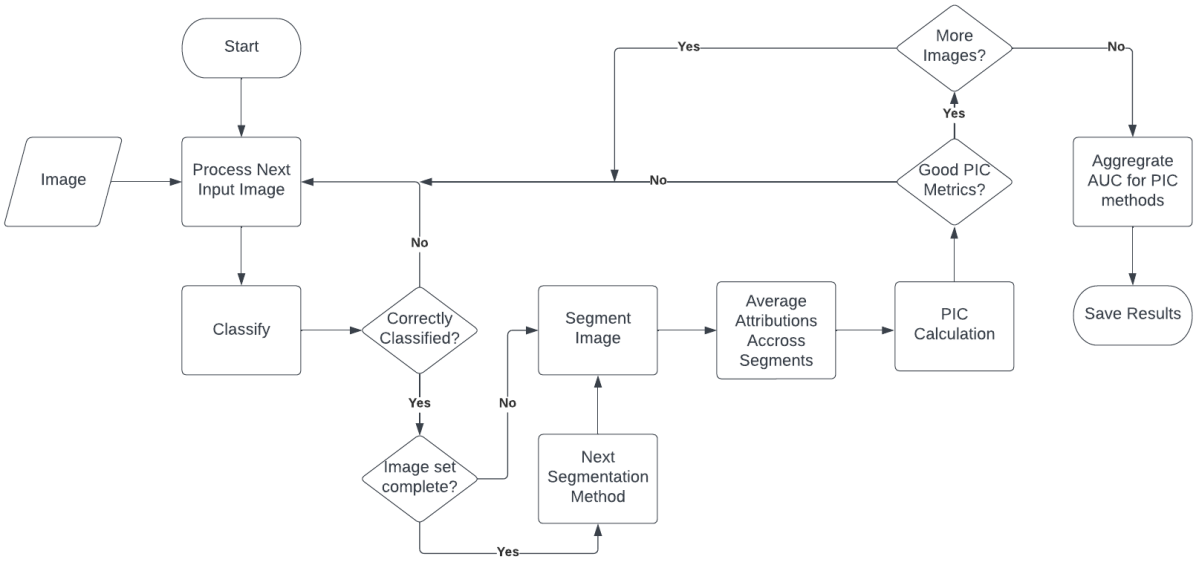


Figure 5: Flowchart for ESAX algorithm

Quickshift was created by Andrea Vedaldi and Stefano Soatto in 2008. It is a local mode-seeking algorithm which is based on mean-shift. They designed the algorithm to address apparent weaknesses in the medioid-shift segmentation method. It uses a Gaussian kernel which is controlled systematically by ESAX to give greater or fewer numbers of segments. Interestingly, there is a built-in scale factor which is assessed during the computation of the algorithm, inherently breaking free from some of the hyper-parameter constraints that XRAI aimed to avoid with over-segmentation. Quickshift tends to be useful for object tracking because it can keep segments consistent in the face of changes in object appearance.

The Watershed algorithm is inspired by natural watersheds, where all water in a region tends to run to a certain local geographical minimum, i.e., a basin. The algorithm takes a grayscale gradient version of the input image, which uses bright pixels to show boundaries between regions. This grayscale image is thought of as a landscape with the brightness of pixel

representing elevation. The landscape is then flooded from specified markers to create segments. This is often applied in medical imaging among other problems due to its ability to segment connected regions well.

The Slic segmentation algorithm is a K-means clustering based method, where K-means is performed on color and image location information. The algorithm groups pixels with similar color and texture characteristics into what are called superpixels which are the results of perceptual grouping of pixels that tend to retain a generally boxy shape. This algorithm is very fast and as such is useful in robotics and computer vision for real time applications.

The Felzenszwalb segmentation algorithm uses a graph to represent images, and then performs segmentation using texture and color to group regions together. The authors of this technique state that it can preserve detail in low-variability regions, while ignoring detail in high-variability regions. Because of its capacity to create highly irregular shapes in its graph-based approach, the Felzenszwalb algorithm is very popular in object detection and recognition.

EXPERIMENTAL EVALUATION

Setup

The attribution method used as the foundation of ESAX is the PyTorch Captum implementation of Integrated Gradients, which uses a black baseline. All data was collected on the PyTorch ResNet101 with ImageNet-1K_V2 weights. To accomplish the goal of exploring different segmentation method's effects of the attribution, a parameter was chosen from each of these algorithms to be modified systematically to alter the size of the segments provided by the algorithms. The parameters for different number of segments within a segmentation algorithm was chosen with the aim of producing three qualitatively different segmentations for every method, one with the *most* segments, one with *fewer* segments, and one with the *least* segments. See Table 1.

	Quickshift *kernel size*	Watershed *markers*	Slic *segments*	Felzenszwalb *scale*
most segments	1	500	500	50
fewer segments	5	200	200	250
least segments	10	20	20	500

Table 1: Segmentation Parameters

Skimage's dilation function was then applied to the segments, yielding a dilated and un-dilated segmentation map. This is a morphological dilation method which grows bright regions and shrinks dark regions, meaning that higher attribution segments will be grown and lower attribution segments will be shrunk. The footprint used in this case is a disk with a radius of 2,

yielding a relatively small dilation. In total, this yields 24 different segmentation maps, a dilated and un-dilate most, fewer, and least segments view of each method. See the appendix for an example of each.

The attributed segments produced by ESAX are fed into the Performance Information Curve algorithm, and the area under the curve for SIC and AIC is aggregated and averaged across the number of images used in the test set. In this experiment, 200 images were used from the ImageNet LSVRC 2012 Validation Images (all tasks) dataset. Approximately the same 200 images were used in every test, but in some cases the segmentation yields bad PIC results, and the image may be discarded. The area was limited to the top attributed 20% of the total area before calculating PIC curves. This is done to focus on the most informative regions. By giving an area threshold with the highest attribution segments being selected first, the PIC scores reflect the capacity of a segmentation method to segment the most relevant parts of an image together.

Results

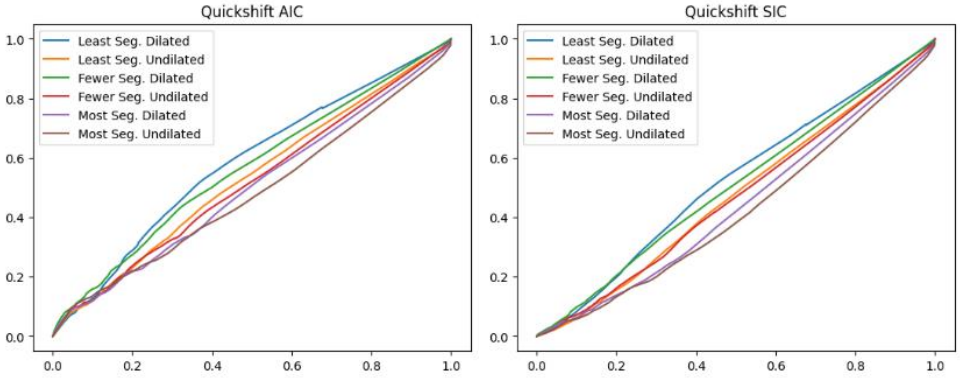


Figure 6: Quickshift PIC

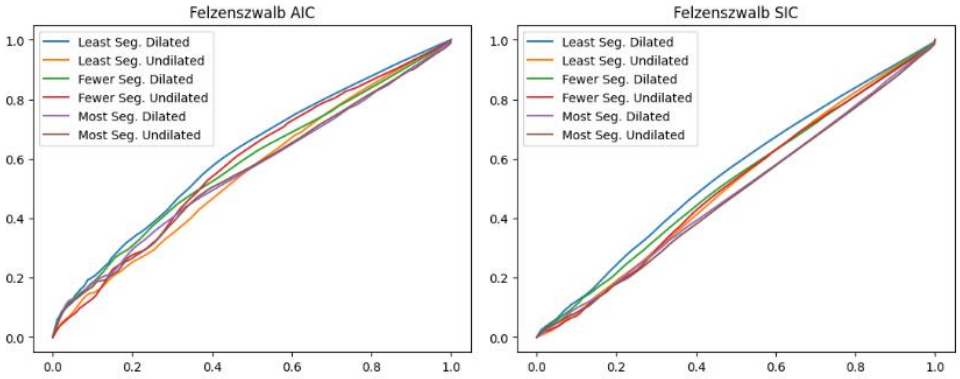


Figure 7: Felzenszwalb PIC

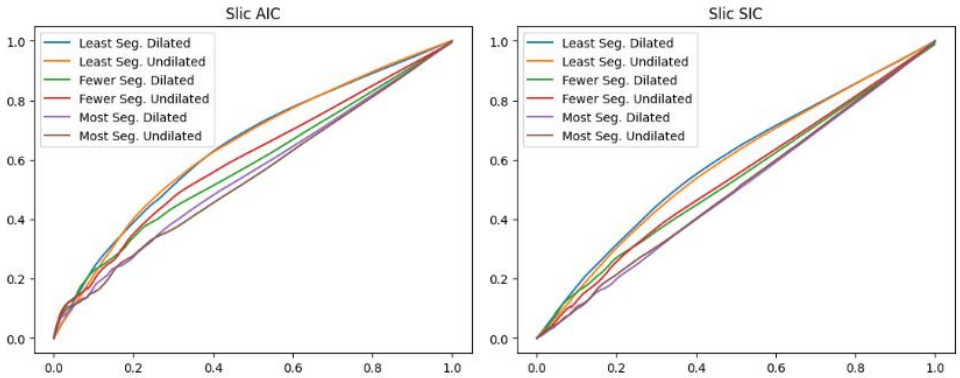


Figure 8: Slic PIC

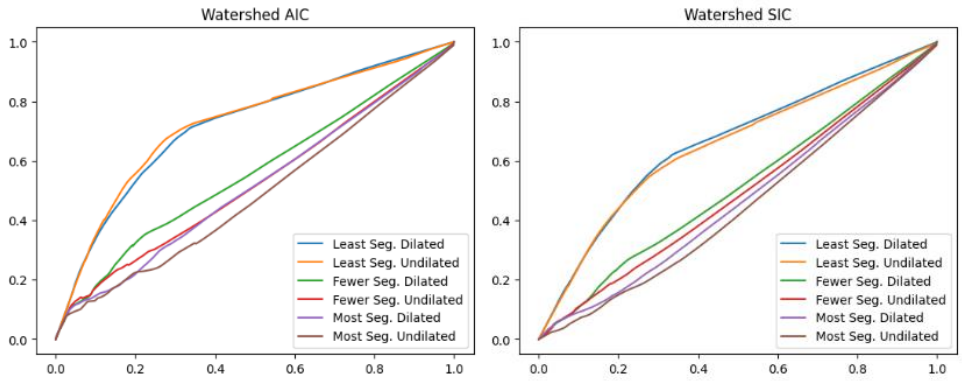


Figure 9: Watershed PIC

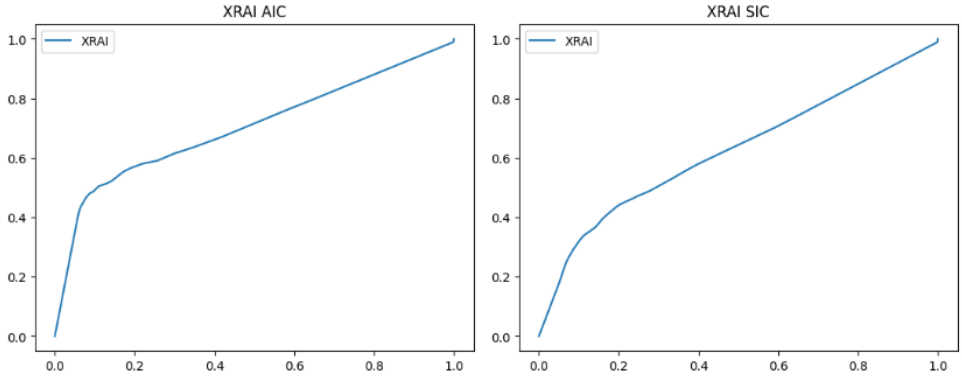


Figure 10: XRAI PIC

Method	AIC AUC
Dilated Watershed Least Segments	0.711
XRAI	0.702
Watershed Least Segments	0.692
Dilated Slic Least Segments	0.622
Dilated Felzenszwalb Least Segments	0.614
Slic Least Segments	0.597
Slic Fewer Segments	0.585
Dilated Slic Fewer Segments	0.583
Slic Most Segments	0.580
Dilated Quickshift Least Segments	0.572
Dilated Watershed Fewer Segments	0.571
Dilated Felzenszwalb Fewer Segments	0.568
Dilated Quickshift Fewer Segments	0.559
Felzenszwalb Fewer Segments	0.559
Watershed Fewer Segments	0.549
Dilated Slic Most Segments	0.548
Quickshift Fewer Segments	0.545
Dilated Felzenszwalb Most Segments	0.545
Felzenszwalb Most Segments	0.544
Quickshift Least Segments	0.537
Dilated Watershed Most Segments	0.536
Felzenszwalb Least Segments	0.534
Dilated Quickshift Most Segments	0.503
Watershed Most Segments	0.495
Quickshift Most Segments	0.488

Method	SIC AUC
Dilated Watershed Least Segments	0.632
Watershed Least Segments	0.625
XRAI	0.624
Dilated Slic Least Segments	0.562
Slic Least Segments	0.544
Dilated Felzenszwalb Least Segments	0.533
Slic Fewer Segments	0.527
Dilated Slic Fewer Segments	0.525
Slic Most Segments	0.514
Dilated Watershed Fewer Segments	0.509
Dilated Quickshift Least Segments	0.504
Dilated Felzenszwalb Fewer Segments	0.500
Dilated Quickshift Fewer Segments	0.493
Felzenszwalb Fewer Segments	0.489
Dilated Felzenszwalb Most Segments	0.488
Watershed Fewer Segments	0.484
Dilated Slic Most Segments	0.482
Felzenszwalb Most Segments	0.479
Quickshift Fewer Segments	0.479
Quickshift Least Segments	0.478
Felzenszwalb Least Segments	0.468
Dilated Watershed Most Segments	0.461
Dilated Quickshift Most Segments	0.445
Watershed Most Segments	0.442
Quickshift Most Segments	0.423

Table 2: AIC and SIC area under the curve for each method

The segmentation methods give a mean AIC AUC of 0.568 with a standard deviation of 0.053, and a mean SIC AUC of 0.504 with a standard deviation of 0.050. The highest scoring method in both metrics, Dilated Watershed Least Segments (DWLS) is 2.7 standard deviations above the AIC mean, and 2.6 standard deviations above the SIC mean. DWLS is followed closely XRAI, and then by Watershed Least Segments (WLS), which in turn is over a standard

deviation away from the next highest scoring. Some segmentations score significantly higher when dilated, with Dilated Felzenszwalb Least Segments (DFLS) scoring nearly 2 standard deviations higher than Felzenszwalb Least Segments (DLS). Notably, DWLS outperformed XRAI, showing that fixed hyper-parameters can perform better than oversampling on PIC testing.

Discussion

In most cases the dilated segmentations score higher than the non-dilated, and the highest scoring segmentation is dilated. This suggests that edge data is important for attribution. Because the segmentation methods generally groups regions with similar characteristics in a segment, the resulting segment boundaries are likely to be along the edges of objects. This is potentially undesirable if the goal is to create segments which fully capture the most salient regions, however, as edges may be crucial for the recognition of objects. It is generally easier to tell what something is from its outline or sketch, as opposed to a zoomed in feature of something which is often difficult to interpret. This is true for humans, and this data supports its validity for AI as well.

The segmentations with the least segments generally outperform their more subdivided counterparts. This may be because the image set used generally contains one main object, and this object generally takes up a large part of the area of the image. This would naturally lead to larger segments being more able to capture the important parts of the object and make the noise outside of the object less likely to impact the attribution after it is averaged across a very large segment. Conversely, smaller segments may be more vulnerable to noise because there is less

area for the high attribution noise to be averaged over. It is possible that for an image set where objects take up a smaller portion of the image would lead to higher scores for the segmentations with a greater number of smaller segments.

The analysis uncovered a small group of outliers, DWLS and WLS, with distinct features that significantly deviate from the general trends observed in the data. This could be interpreted as an exception, due to the fact that the large segments produced by these images tend to resemble superpixels more than segments, though watershed seems to have an ability to highlight very distinct features (see the elephant's ear and tusk in the appendix). Although DLSW performed better than Felzenszwalb (used in XRAI) in PIC measurement, Felzenszwalb tends to create segments which are subjectively more visually appealing, at least in the grayscale shown in the appendix.

FUTURE WORK

A direction for future work in the field of segmenting is laid out by ESAX. Experimentation with different segmentation techniques on image sets where the size and prominence of objects of interest is more variable would yield understanding for specific applications like surveillance or warfare. An exploration of the application of segmented attributions to video feeds where the capacity of segmentation methods to resolve objects over translations and rotations would expand the scope of XAI as well. Application of deep learning in the segmentation methods themselves, perhaps in the choosing of hyper-parameters, could yield significantly more interesting and understandable results, where the dependance on any specific hyper-parameter could be removed entirely by outsourcing to a DNN. Utilization of oversampling on different techniques than Felzenszwalb is a logical next step as well, with Watershed showing promise. It is likely that choosing the optimal method is a domain specific decision.

A combination of segmentation methods like what is known as an ensemble method would be appropriate for maximizing the intersection of performance and understandability of attribution methods. By representing segmentation as a tree problem, bootstrapping and bagging [9] could be applied to find more optimal segmentations. Training a deep network on subjective experimental data on explainability as well as PIC metrics could yield a model which selects features to group together based on their maximization of these scores. Some combination of these techniques and others will help XAI to adapt to ever increasing complexity of models, making total explainability a moving goalpost.

Because the objective measurements used in this study are inconclusive for actual utility in human understanding, subjective testing of different attributions methods is necessary for the goal of trust to be reached. Future work specifically identifying the effectiveness of current methods of attribution on the trustworthiness of AI is necessary to determine further direction for the field. With more complex language processing

CONCLUSION

In a world of rapidly evolving AI being incorporated into sensitive and safety-critical systems, the need for a basis of trust of AI is more evident than ever. XAI aims to address this by revealing the inner workings of black box algorithms. Attribution methods highlight features based on their relevancy to neural network decision making. Segmentation attribution methods aim to align attribution with our perceptions by showing relevant regions of interest. This thesis furthers the precedence for the use of segmentation in AI. It outlines a direction for future work and provides the basis for deeper research into the understanding of visual AI via segmentation.

The results find two high performing outliers among the 24 segmentation methods tests, which outperform XRAI on the image set. This study's results emphasize the importance of domain-specific approaches, and the need to reconcile the tradeoff between interpretability and objective measurements. By setting the stage for deeper research into the understanding of visual AI via segmentation, this thesis not only addresses immediate challenges in the field of XAI but also encourages future work that can help bridge the gap between human perceptions and AI decision-making. A roadmap is defined for future work in XAI, and the need is established for further research into the subjective effectiveness of segmentation for understanding AI decisions.

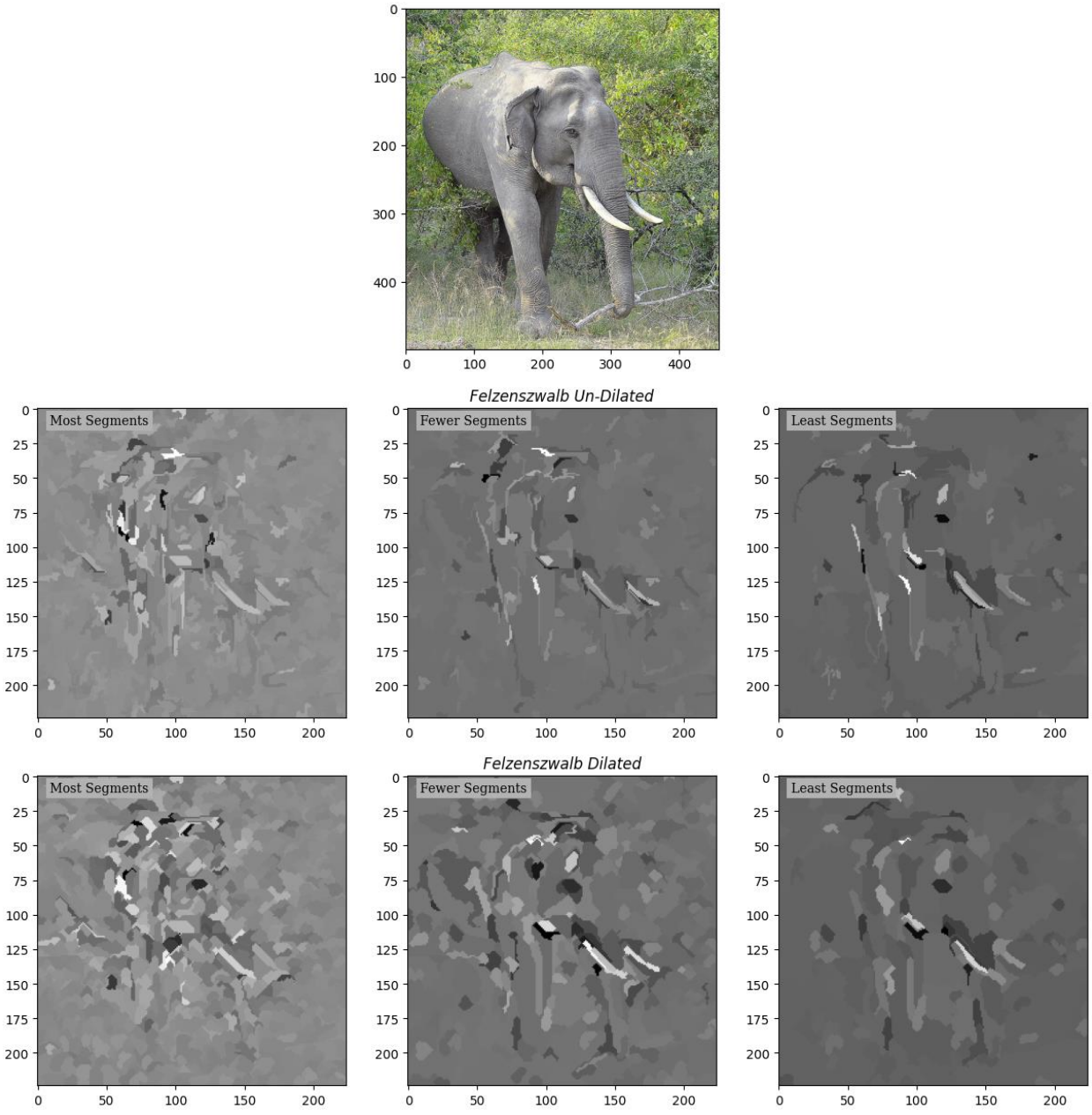
LIST OF REFERENCES

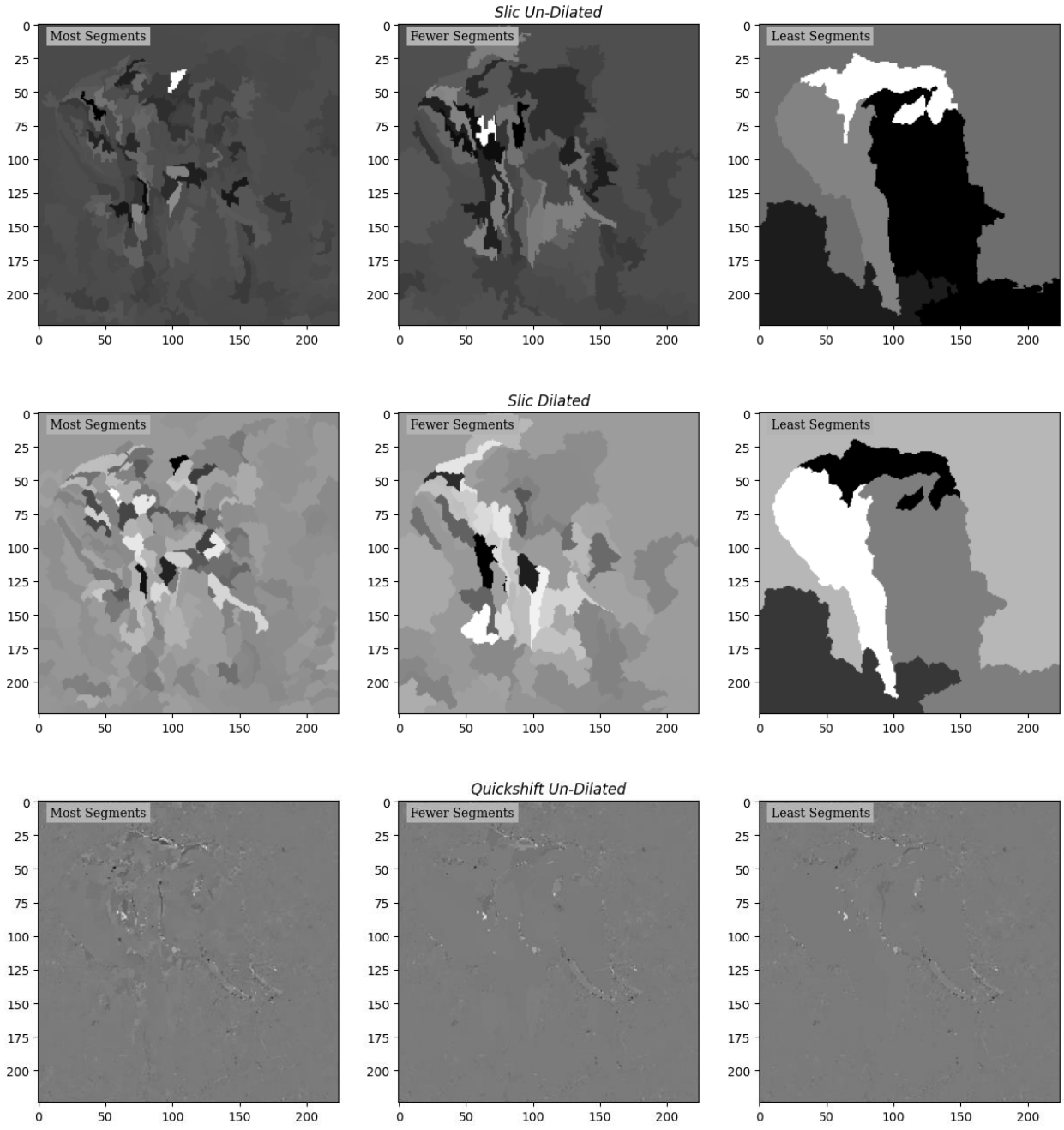
1. A. Kapishnikov, T. Bolukbasi, F. Viegas, and M. Terry, “XRAI: Better attributions through regions,” *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Aug. 2019.
2. D. Erhan, Y. Bengio, A. Courville, and P. Vincent, "Visualizing higher-layer features of a deep network," Technical Report 1341, University of Montreal, Montreal, QC, Canada, Jun. 2009.
3. K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps," arXiv preprint arXiv:1312.6034, Dec. 2013.
4. M. Olazaran, "A Sociological Study of the Official History of the Perceptrons Controversy," Social Studies of Science, vol. 26, no. 3, pp. 611-659, 1996.
5. G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," Neural Comput., vol. 18, no. 7, pp. 1527-1554, Jul. 2006.
6. Scikit-image Development Team, "Gallery of examples: Comparing different segmentation algorithms on the same image," Scikit-image, 2021. [Online]. Available: https://scikit-image.org/docs/stable/auto_examples/segmentation/plot_segmentations.html. [Accessed: April 16, 2023].

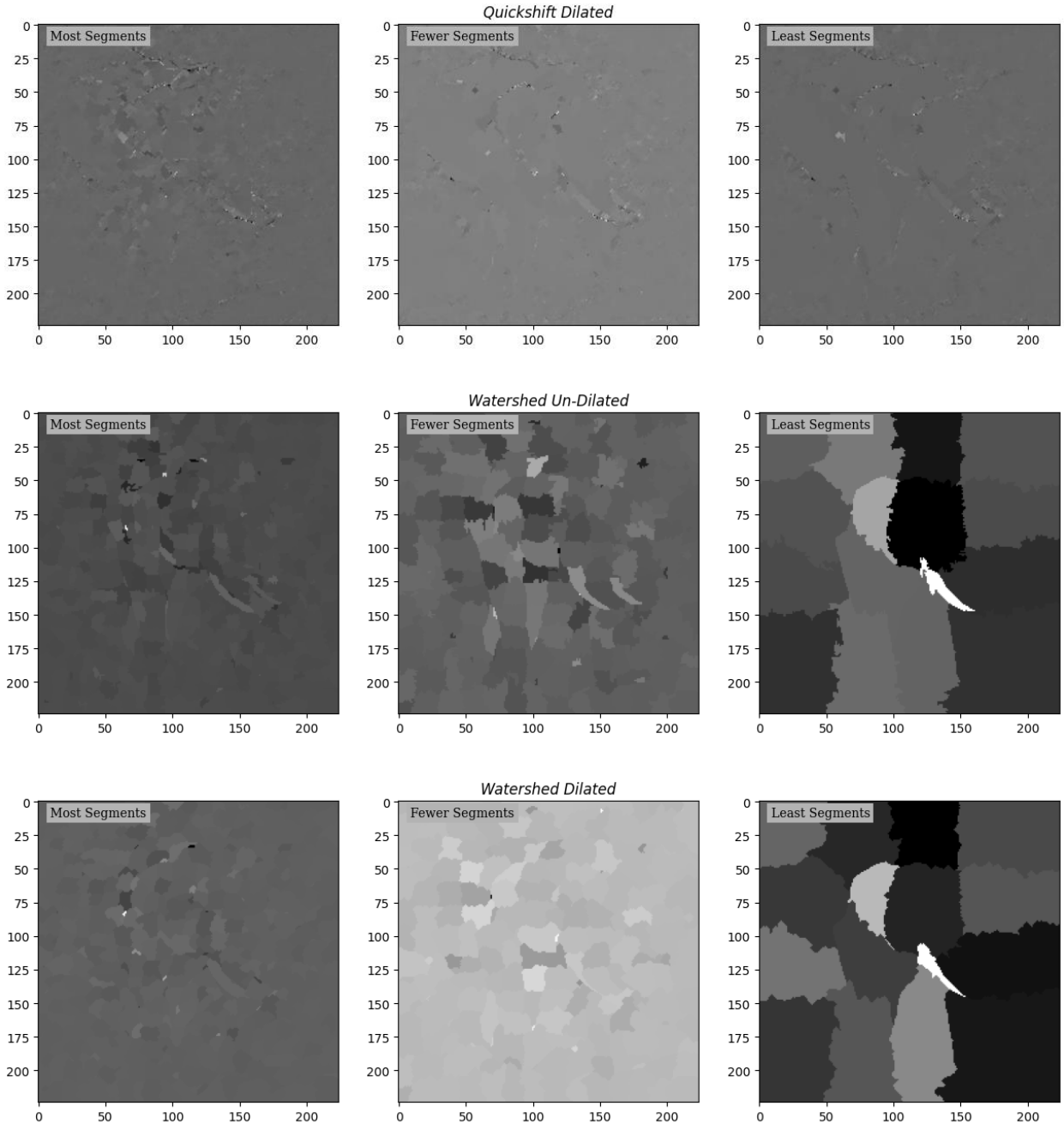
7. TU Chemnitz, "Superpixels," Department of Electrical Engineering and Information Technology, Chemnitz University of Technology. [Online]. Available: https://www.tu-chemnitz.de/etit/proaut/en/research/superpixel.html. [Accessed: Apr. 16, 2023]
8. P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient Graph-Based Image Segmentation," International Journal of Computer Vision, vol. 59, no. 2, pp. 167-181, Sep. 2004.
9. Joseph Rocca, "Ensemble Methods: Bagging, Boosting, and Stacking," Towards Data Science, 2021. [Online]. Available: https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205. [Accessed: 16-April-2023].

APPENDIX

Example of each attributed segmentation. Brighter segments have a higher attribution.




Slic Un-Dilated
Most Segments
Fewer Segments
Least Segments
Slic Dilated
Most Segments
Fewer Segments
Least Segments
Quickshift Un-Dilated
Most Segments
Fewer Segments
Least Segments


Quickshift Dilated
Most Segments
Fewer Segments
Least Segments
Watershed Un-Dilated
Most Segments
Fewer Segments
Least Segments
Watershed Dilated
Most Segments
Fewer Segments
Least Segments