# EVALUATING GROUPS OF FEATURES VIA CONSISTENCY, CONTIGUITY, AND STABILITY

**Chaehyeon Kim, Weiqiu You, Shreya Havaldar & Eric Wong**
Department of Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104, USA
{chaenyk,weiqiuy,shreyah,exwong}@seas.upenn.edu

## ABSTRACT

Feature attributions explain model predictions by assigning importance scores to input features. In high-dimensional data such as images, these scores are often assigned to groups of features at a time. There are a variety of strategies for creating these groups, ranging from simple patches to deep-learning-based segmentation algorithms. What makes certain groups better than others for explanations? We formally define three key criteria for interpretable groups of features: consistency, contiguity, and stability. Surprisingly, we find that patch-based groups outperform groups created via modern segmentation tools.

## 1 THE IMPORTANCE OF GROUPS FOR EXPLANATIONS

When explaining high-dimensional data, feature attributions often assign importance scores to groups of features rather than individual features. In images, groups are typically collections of neighboring pixels from the same object. For example, the number of groups in an ImageNet image is much smaller than 50k, the number of pixels. By reducing mental overhead, scores for grouped features can be more accessible for humans to understand than scores for individual features.

Existing strategies for creating groups of features range from simple patches to classic or deep-learning-based algorithms. In this work, we study four techniques for dividing images into groups: classic computer vision approaches (patches (Alexe et al., 2021) and superpixels (Levner & Zhang, 2007)) as well as modern deep-learning approaches (Segment Anything Model (SAM) (Alexander et al., 2023), Archipelago (Michael et al., 2020)). While groups from different approaches can be quite diverse in shape, size, and content, there is no standard benchmark for comparing groups and measuring their qualities.

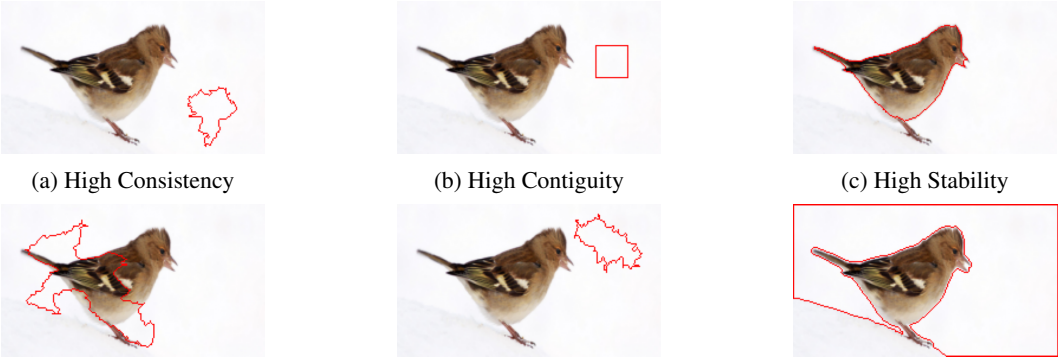
To better understand groups for explanations, we propose three criteria that a group should satisfy to improve the quality of feature attributions. These criteria are designed such that ideal groups result in more consistent, stable, and contiguous attributions. In the following analysis, we find evidence that despite advancements in modern segmentation tools, patch-based groups are more suitable for feature attributions.

## 2 DESIDERATA OF GROUPS FOR FEATURE ATTRIBUTIONS

We first discuss the desiderata of groups for feature attributions. We posit that an ideal group should possess the following characteristics: the group should be consistent and contiguous to be human interpretable, and it also needs to be stable to have a reliable feature attribution. In this section, we elaborate on these properties and show the representative examples in Figure 1.

Formally, we define $x \in \mathbb{R}^n$ be an input such as an image. We also represent a boolean vector group as $\alpha \in \{0, 1\}^n$, where $\alpha_i = 1$ if the $i$th feature is included in the group, and 0 otherwise.

**Consistency.** Consistency is the semantic coherency of features within the same group. Highly consistent groups are easier for humans to understand as such groups do not contain many different



(a) High Consistency (b) High Contiguity (c) High Stability

(d) Low Consistency (e) Low Contiguity (f) Low Stability

Figure 1: Illustrative examples of groups with high and low consistency, contiguity, and stability using patch (b), SAM (c, f), and Archipelago (a, b, e).

concepts. To calculate the consistency, we use the embedding cosine similarity to measure the closeness or similarity between pairs of features. Formally, let $h : \mathbb{R}^n \to \mathbb{R}^d$ be an embedding function of the input. Then, for each pair of $i$-th and $j$-th features,

$$\text{Consistency}(x, \alpha)_{ij} = \frac{h(x_i) \cdot h(x_j)}{\|h(x_i)\| \|h(x_j)\|} \mathbb{1}(\alpha_i \cdot \alpha_j). \tag{1}$$

**Contiguity.** Group contiguity captures how smooth a group is and penalizes the "holes" in the group. Contiguous inputs are more interpretable for humans. Then,

$$\text{ContiguityPenalty}(\alpha) = \frac{\lambda \|K\alpha\|_{2,1}}{\sqrt{\|\alpha\|_1}} \quad \text{where } K = \begin{pmatrix} (1-\rho)\nabla \\ \rho \text{Id}_p \end{pmatrix}. \tag{2}$$

Here, $\nabla$ is the spatial gradient operator, $\text{Id}_p$ is the identity operator, and the parameter $\rho$ balances the sparsity of the input and the spatial gradient.

**Stability.** Stability captures how much the model output changes when adding additional features to a group, a desirable property for feature attributions (Xue et al., 2023). Formally, let $f : \mathbb{R}^n \to \mathbb{R}$ be a predictive model. Then,

$$\text{Stability}(x, \alpha) = \max_i \|f(x \odot \alpha_{+i}) - f(x \odot \alpha)\| \tag{3}$$

where $\alpha_{+i}$ is the group created by adding the $i$-th feature to $\alpha$. Further details of these three properties are discussed in Appendix C.

## 3 Modern Grouping Methods are No Better than Patches

Finally, we analyze how different grouping methods perform across the three desiderata. Specifically, we generate groups using patches, superpixels, SAM, and Archipelago on the ImageNet dataset (Deng et al., 2009) and measured the corresponding consistency, contiguity, and stability of these groups. In this section, we highlight some of the main findings in our analysis.

We find that modern approaches such as SAM and Archipelago are significantly less reliable. These groups have consistency as low as 0.20, about $\times 4$ worse than the average. In contrast, the worst for superpixels is 0.49, which is about $\times 1.6$ worse than the average. Results for all metrics and methods are in Table 1 of Appendix E. In fact, SAM segments have the worst contiguity and stability of all methods, suggesting that SAM segments are the least suitable for feature attributions.

On the other hand, simple patches are the most reliable option for grouping since they have the highest performance across all metrics. We hypothesize that this advantage may be related to the prevalence of Vision Transformers (ViTs) (Alexe et al., 2021), as ViTs process images as sequences of patches. However, superpixels also perform relatively well. Further discussions are in Appendix E.

**Conclusion.** Currently, our results suggest that patches are the more suitable groups over modern approaches. We hope that our desiderata and corresponding analysis can guide future work in developing more reliable groups for explanations and interpretable for humans.

### URM STATEMENT

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2024 Tiny Papers Track.

## REFERENCES

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018.

Julius Adebayo, Michael Muelly, Hal Abelson, and Been Kim. Post hoc explanations may be ineffective for detecting unknown spurious correlation, 2022.

Rishabh Agarwal, Levi Melnick, Nicholas Frosst, Xuezhou Zhang, Ben Lengerich, Rich Caruana, and Geoffrey Hinton. Neural additive models: Interpretable machine learning with neural nets. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=wHkKTW2wrmm.

Kirillov Alexander, Mintun Eric, Ravi Nikhila, Mao Hanzi, Rolland Chloe, Gustafson Laura, Xiao Tete, Whitehead Spencer, C. Berg Alexander, Lo Wan-Yen, Dollár Piotr, and Girshick Ross. Segment anything. In *arXiv preprint arXiv:2304.02643*. 2023.

Dosovitskiy Alexe, Beyer Lucas, Kolesnikov Alexander, Weissenborn Dirk, Zhai Xiaohua, Unterthiner Thomas, Dehghani Mostafa, Minderer Matthias, Heigold Georg, Gelly Sylvain, Uszkoreit Jakob, and Houlsby Neil. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations (ICLR)*. 2021.

Afra Amini, Tiago Pimentel, Clara Meister, and Ryan Cotterell. Naturalistic causal probing for morpho-syntax. *Transactions of the Association for Computational Linguistics*, 11:384–403, 2022. URL https://api.semanticscholar.org/CorpusID:248811730.

Vedaldi Andrea and Soatto Stefano. Quick shift and kernel methods for mode seeking. In *European Conference on Computer Vision*. 2008.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE*, 10, 2015. URL https://api.semanticscholar.org/CorpusID:9327892.

Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. volume 39, pp. 2481–2495. 2017. doi: 10.1109/TPAMI.2016.2644615.

David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Mueller. How to explain individual classification decisions, 2009.

Shahaf Bassan and Guy Katz. Towards formal xai: Formally approximate minimal explanations of neural networks. In *International Conference on Tools and Algorithms for the Construction and Analysis of Systems*, pp. 187–207. Springer, 2023.

Jasmijn Bastings and Katja Filippova. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, 2020. doi: 10.18653/v1/2020.blackboxnlp-1.14. URL http://dx.doi.org/10.18653/V1/2020.BLACKBOXNLP-1.14.

Jasmijn Bastings, Sebastian Ebert, Polina Zablotskaia, Anders Sandholm, and Katja Filippova. "will you find these shortcuts?" a protocol for evaluating the faithfulness of input salience methods for text classification. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 976–991, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.64. URL https://aclanthology.org/2022.emnlp-main.64.

Blair Bilodeau, Natasha Jaques, Pang Wei Koh, and Been Kim. Impossibility theorems for feature attribution. *ArXiv*, abs/2212.11870, 2022. URL https://api.semanticscholar.org/CorpusID:254974246.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models, 2022.

Nadia Burkart and Marco F. Huber. A survey on the explainability of supervised machine learning. *J. Artif. Int. Res.*, 70:245–317, may 2021. ISSN 1076-9757. doi: 10.1613/jair.1.12228. URL https://doi.org/10.1613/jair.1.12228.

Nitay Calderon, Eyal Ben-David, Amir Feder, and Roi Reichart. DoCoGen: Domain counterfactual generation for low resource domain adaptation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7727–7746, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.533. URL https://aclanthology.org/2022.acl-long.533.

Rumelhart David E., Hinton Geoffrey E., and Williams Ronald J. Learning representations by back-propagating errors. In *nature*, volume 5, pp. 533–536. 1986.

Nicola De Cao, Michael Sejr Schlichtkrull, Wilker Aziz, and Ivan Titov. How do decisions emerge across layers in neural models? interpretation with differentiable masking. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020. doi: 10.18653/v1/2020.emnlp-main.262. URL http://dx.doi.org/10.18653/v1/2020.emnlp-main.262.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2009.

Misha Denil, Alban Demiraj, and Nando de Freitas. Extraction of salient sentences from labelled documents. *ArXiv*, abs/1412.6815, 2014. URL https://api.semanticscholar.org/CorpusID:9121062.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4443–4458, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.408. URL https://aclanthology.org/2020.acl-main.408.

Michael Eickenberg, Elvis Dohmatob, Bertrand Thirion, and Gaël Varoquaux. Total variation meets sparsity: statistical learning with segmenting penalties. In *Medical Image Computing and Computer Aided Intervention (MICCAI)*. Oct 2015.

Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. In *Biological Cybernetics*. 1980.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep learning. MIT press, 2016.

Shreya Havaldar, Adam Stein, Eric Wong, and Lyle Ungar. Topex: Topic-based explanations for model comparison, 2023.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. volume 9, pp. 1735–1780. 1997. doi: 10.1162/neco.1997.9.8.1735.

Sara Hooker, D. Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Neural Information Processing Systems*, 2018. URL https://api.semanticscholar.org/CorpusID:202782699.

Sarthak Jain, Sarah Wiegreffe, Yuval Pinter, and Byron C. Wallace. Learning to faithfully rationalize by construction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.409. URL http://dx.doi.org/10.18653/v1/2020.acl-main.409.

Ákos Kádár, Grzegorz Chrupała, and Afra Alishahi. Representation of linguistic form and function in recurrent neural networks. *Computational Linguistics*, 43(4):761–780, December 2017. doi: 10.1162/COLI_a_00300. URL https://aclanthology.org/J17-4003.

Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. Learning the difference that makes a difference with counterfactually-augmented data, 2020.

Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (un)reliability of saliency methods. *Lecture Notes in Computer Science*, pp. 267–280, 2019. ISSN 1611-3349. doi: 10.1007/978-3-030-28954-6_14. URL http://dx.doi.org/10.1007/978-3-030-28954-6_14.

Thibault Laugel, Xavier Renard, Marie-Jeanne Lesot, Christophe Marsala, and Marcin Detyniecki. Defining locality for surrogates in post-hoc interpretablity, 2018.

Ilya Levner and Hong Zhang. Classification-driven watershed segmentation. In *IEEE Transactions on Image Processing*. 2007.

Jiwei Li, Will Monroe, and Dan Jurafsky. Understanding neural networks through representation erasure, 2017.

Xuhong Li, Haoyi Xiong, Xingjian Li, Xuanyu Wu, Xiao Zhang, Ji Liu, Jiang Bian, and Dejing Dou. Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond. *Knowl. Inf. Syst.*, 64(12):3197–3234, dec 2022. ISSN 0219-1377. doi: 10.1007/s10115-022-01756-8. URL https://doi.org/10.1007/s10115-022-01756-8.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015.

Yin Lou, Rich Caruana, and Johannes Gehrke. Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pp. 150–158, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450314626. doi: 10.1145/2339530.2339556. URL https://doi.org/10.1145/2339530.2339556.

Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pp. 623–631, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450321747. doi: 10.1145/2487575.2487579. URL https://doi.org/10.1145/2487575.2487579.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 4768–4777, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. Towards faithful model explanation in nlp: A survey, 2022.

Tsang Michael, Rambhatla Sirisha, and Liu Yan. How does this interaction affect me? interpretable attribution for feature interactions. In *Advances in Neural Information Processing Systems*. MIT Press, 2020.

Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65:211–222, 2017. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2016.11.008. URL https://www.sciencedirect.com/science/article/pii/S0031320316303582.

Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Comput. Surv.*, 55 (13s), jul 2023. ISSN 0360-0300. doi: 10.1145/3583558. URL https://doi.org/10.1145/3583558.

Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2015.

Felzenszwalb Pedro F. and Huttenlocher Daniel P. Efficient graph-based image segmentation. In *International Journal of Computer Vision*. 2004.

Neubert Peer and Protzel Peter. Compact watershed and preemptive slic: On improving trade-offs of superpixel segmentation algorithms. In *ICPR*. 2014.

Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. In *British Machine Vision Conference*, 2018.

Achanta Radhakrishna, Shaji Appu, Smith Kevin, Lucchi Aurelien, Fua Pascal, and Suesstrunk Sabine. Slic superpixels compared to state-of-the-art superpixel methods. In *TPAMI*. 2012.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), April 2018. doi: 10.1609/aaai.v32i1.11491. URL https://ojs.aaai.org/index.php/AAAI/article/view/11491.

Yao Rong, Tobias Leemann, Vadim Borisov, Gjergji Kasneci, and Enkelejda Kasneci. A consistent and efficient evaluation strategy for attribution methods. In *International Conference on Machine Learning*, 2022. URL https://api.semanticscholar.org/CorpusID:249642068.

Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128:336–359, 2016. URL https://api.semanticscholar.org/CorpusID:15019293.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 3145–3153. JMLR.org, 2017.

Sandipan Sikdar, Parantapa Bhattacharya, and Kieran Heese. Integrated directional gradients: Feature interaction attribution for neural NLP models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 865–878, Online, August 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.acl-long.71.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise, 2017.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. Striving for simplicity: The all convolutional net. *CoRR*, abs/1412.6806, 2014. URL https://api.semanticscholar.org/CorpusID:12998557.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, 2017. URL https://api.semanticscholar.org/CorpusID:16747630.

Bhavan Vasu and Chengjiang Long. Iterative and adaptive sampling with spatial attention for black-box model explanations. *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2020. doi: 10.1109/wacv45572.2020.9093576. URL http://dx.doi.org/10.1109/WACV45572.2020.9093576.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 6707–6723, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.523. URL https://aclanthology.org/2021.acl-long.523.

Anton Xue, Rajeev Alur, and Eric Wong. Stability guarantees for feature attributions with multiplicative smoothing, 2023.

Weiqiu You, Helen Qu, Marco Gatti, Bhuvnesh Jain, and Eric Wong. Sum-of-parts models: Faithful attributions for groups of features, 2023.

Jianlong Zhou, Amir H. Gandomi, Fang Chen, and Andreas Holzinger. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, 10(5), 2021a. ISSN 2079-9292. doi: 10.3390/electronics10050593. URL https://www.mdpi.com/2079-9292/10/5/593.

Yilun Zhou and Julie Shah. The solvability of interpretability evaluation metrics. In Andreas Vlachos and Isabelle Augenstein (eds.), *Findings of the Association for Computational Linguistics: EACL 2023*, pp. 2399–2415, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-eacl.182. URL https://aclanthology.org/2023.findings-eacl.182.

Yilun Zhou, Serena Booth, Marco Tulio Ribeiro, and Julie A. Shah. Do feature attribution methods correctly attribute features? *ArXiv*, abs/2104.14403, 2021b. URL https://api.semanticscholar.org/CorpusID:233443847.

Shiping Zhu, Xi Xia, Qingrong Zhang, and Kamel Belloulata. An image segmentation algorithm in image processing based on threshold segmentation. In *IEEE Conference on Signal-Image Technologies and Internet-Based System*. 2007.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1651–1661, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1161. URL https://aclanthology.org/P19-1161.

# A APPENDIX

We include additional information detailing the evaluation of existing groups. Appendix B provides a review of studies related to our work, Appendix C contains a complete description of each metric, Appendix D provides experiment details, and Appendix E shows additional results.

# B RELATED WORK

**Segmentation.** Image segmentation is a crucial task in computer vision and image processing. The primary objective is to divide an image into meaningful groups, which has significant applications in various fields, including medical image analysis. There are several methods available for image segmentation, ranging from simple thresholding (Zhu et al., 2007) to more complex superpixel algorithms such as Felzenswalb's method (Pedro F. & Daniel P., 2004), which is based on graphs, the Quickshift method (Andrea & Stefano, 2008) that approximates kernelized mean-shift, the SLIC method (Radhakrishna et al., 2012) that is based on k-means clustering, and the Watershed method (Peer & Peter, 2014) that is based on the grayscale gradient image.

Modern deep learning methods such as CNN (Fukushima, 1980), Recurrent Neural Networks (RNNs) (David E. et al., 1986), and Long Short-Term Memory (LSTM) (Hochreiter & Schmidhuber, 1997) have yielded a new generation of image segmentation models that exhibit remarkable performance improvements. Fully Convolutional Networks (FCNs) (Long et al., 2015) replace the fully connected layers in convolutional networks with convolutional layers, allowing the network to output spatial heatmaps for segmentation. DeconvNet (Noh et al., 2015) conducts semantic segmentation based on transposed convolution using encoder (Goodfellow et al., 2016) - decoder (Badrinarayanan et al., 2017) based models.

**Feature Attribution.** Extensive surveys discuss various explainability methods (Nauta et al., 2023; Burkart & Huber, 2021; Zhou et al., 2021a; Li et al., 2022; Lyu et al., 2022; Bommasani et al., 2022). Feature attribution methods include post-hoc and built-in attributions. For post-hoc methods, gradient-based attributions include using gradients directly (Selvaraju et al., 2016; Baehrens et al., 2009; Simonyan et al., 2014; Bastings & Filippova, 2020), gradient $\times$ inputs (Sundararajan et al., 2017; Denil et al., 2014; Smilkov et al., 2017) and through propagation (Ribeiro et al., 2018; Springenberg et al., 2014; Bach et al., 2015; Shrikumar et al., 2017; Montavon et al., 2017). Other methods create a surrogate model to approximate the original model (Ribeiro et al., 2016; Lundberg & Lee, 2017; Laugel et al., 2018), or perturb inputs and observe the change in predictions (Petsiuk et al., 2018; Vasu & Long, 2020; Kaushik et al., 2020; Li et al., 2017; Kádár et al., 2017; Ribeiro et al., 2018; De Cao et al., 2020) and use manual (Kaushik et al., 2020) or automatic (Calderon et al., 2022; Zmigrod et al., 2019; Amini et al., 2022; Wu et al., 2021) counterfactual perturbation. While the above methods focus on individual features, Michael et al. (2020); Sikdar et al. (2021) investigates feature interactions. Another line of work has built-in feature attributions including training different modules for each feature (Lou et al., 2012; 2013; Agarwal et al., 2021), or pre-selecting one or multiple groups of features to use for prediction (Jain et al., 2020; You et al., 2023).

**Groups for Explanations.** Feature attribution methods often attribute to groups of features. For images, some works attribute to each individual pixel by randomly sampling masks using patches and interpolating between them (Petsiuk et al., 2018) or attributing to each pixel using gradients (Selvaraju et al., 2016; Baehrens et al., 2009; Simonyan et al., 2014; Bastings & Filippova, 2020). Other works assign scores to patches (Ribeiro et al., 2016; Lundberg & Lee, 2017; You et al., 2023) or segments (Michael et al., 2020; You et al., 2023).

Some feature attribution methods assign scores to a higher level of groups of features. Archipelago (Michael et al., 2020) post-hoc selects groups by merging pairs of interacting features. FRESH (Jain et al., 2020) selects a group of features using top $k$ tokens from an attention mechanism in a pre-trained Transformer. SOP (You et al., 2023) first selects groups of features with external attention trained end-to-end with a pre-trained model. IDG (Sikdar et al., 2021) selects groups based on a hierarchical structure. TopEx (Havaldar et al., 2023) uses topics as groups for explanations.

**Attribution Evaluation.** Various metrics for evaluating feature attribution methods are studied (Nauta et al., 2023; Zhou et al., 2021b; Adebayo et al., 2018; Hooker et al., 2018; DeYoung et al.,

2020; Bastings et al., 2022; Rong et al., 2022; Zhou & Shah, 2023; Adebayo et al., 2022). Some perturbation metrics include insertion and deletion for images (Petsiuk et al., 2018) and comprehensiveness and sufficiency for text (DeYoung et al., 2020). The performance is also evaluated for models with built-in explanations to ensure no considerable degradation from the original model (Jain et al., 2020; You et al., 2023).

There are some debates about what constitutes good explanations. Multiple works have shown the failures of feature attributions (Bilodeau et al., 2022; Sundararajan et al., 2017; Adebayo et al., 2018; Kindermans et al., 2019). Different properties are proposed as important axioms that feature attributions should satisfy, including stability (Xue et al., 2023), minimality (Bassan & Katz, 2023), and faithfulness (You et al., 2023).

## C METRICS

This section describes in detail how we measure group consistency, contiguity, and stability. As mentioned in the main paper, $x \in \mathbb{R}^n$ is input such as an image, and $\alpha \in \{0,1\}^n$ indicates group membership of each feature, where $\alpha_i = 1$ if the $i$th feature is included in the group, and 0 otherwise.

### C.1 GROUP CONSISTENCY

To measure consistency, we use the cosine similarity of embeddings of features within the group. More specifically, we divide the group into patches, compute the embedding similarity between each pair of patches in the group, and take the average for all the pairs in the group. We calculate the group consistency as follows:

$$\forall i, j \text{ such that } i \neq j \quad \text{Consistency}(x, \alpha) = \frac{1}{\|\alpha\|(\|\alpha\| - 1)} \sum_i \sum_j \frac{h(x_i) \cdot h(x_j)}{\|h(x_i)\| \|h(x_j)\|} \mathbb{1}(\alpha_i \cdot \alpha_j) \tag{4}$$

where $h : \mathbb{R}^n \to \mathbb{R}^d$ be an embedding function of the input. To be comparable with other metrics, the consistency is then projected to consistency loss, which ranges from 0 to 1, with 0 as the best value and 1 as the worst:

$$\text{ConsistencyLoss}(x, \alpha) = 1 - \frac{\text{Consistency}(x, \alpha) + 1}{2} \tag{5}$$

When there is only one patch that contains anything in the group, we discard the group when computing the average consistency for a grouping method. This is because a group with only one patch would have the perfect consistency loss of 0, but using the singleton groups will bias the consistency metric to favor smaller groups. Lower consistency loss is more desired than higher consistency loss.

### C.2 GROUP CONTIGUITY

To measure contiguity, we follow the new penalty for segmentation purposes called *sparse variation* (Eickenberg et al., Oct 2015). The sparse variation penalty enforces contiguous zones in the group, and lower sparse variation values can guarantee greater contiguity within a group. We used the normalized form of *sparse variation* to make it independent from the group size. We formalize the group contiguity loss as follows:

$$\text{ContiguityLoss}(\alpha) = \frac{\lambda \|K\alpha\|_{2,1}}{\sqrt{\|\alpha\|_1}} \quad \text{where } K = \begin{pmatrix} (1-\rho)\nabla \\ \rho \text{Id}_p \end{pmatrix}. \tag{6}$$

where $\nabla$ is the spatial gradient operator. For a 2D mask of size $p = H \times W$, where $H$ is the height and $W$ is the width of the image. We have $\nabla \in \mathbb{R}^{2p \times p}$. The matrix $\text{Id}_p$ is the identity operator of size $p \times p$. Finally, $\rho$ is the parameter controlling the trade-off between coordinate sparsity and spatial regularity. We take $\rho = 0.5$ in our calculations. Lower contiguity loss is more desired than higher contiguity loss.

### C.3 GROUP STABILITY

To ensure the stability of a feature group, the addition of more features should not result in a change of predicted class once the group has been selected. To expand the group's feature set, we consider all possible patches with features yet to be chosen by the group. We calculate the stability loss for adding each additional patch by computing the difference in predicted probabilities. For our experiments, we used $16 \times 16$ patches to be consistent with ViT. Then,

$$\text{StabilityLoss}(x, \alpha) = \max_i \|f(x \odot \alpha_{+i}) - f(x \odot \alpha)\| \tag{7}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ be an predictive model, and $\alpha_{+i}$ is the group created by adding the $i$-th patch to $\alpha$. The stability loss ranges from 0 to 1 with 0 being the best and 1 the worst, which matches the direction of other metrics.

## D EXPERIMENT DETAILS

### D.1 DATASET

We use the ImageNet dataset (Deng et al., 2009), a classification task using human-annotated photographs into 1000 classes.

### D.2 MODEL

We use a pre-trained Vision Transformer (ViT) model (Alexe et al., 2021) pre-trained on ImageNet-21k (14 million images, 21,843 classes) at resolution 224x224 and fine-tuned on ImageNet 2012 (1 million images, 1,000 classes) at resolution 224x224 [1].

### D.3 GROUPING METHODS

Here, we describe the four grouping methods we examine.

- Patches (Alexe et al., 2021) divide an image into fixed-size blocks without overlapping regions.
- Superpixels interpret pixel values as local topography and fill areas from initial points until adjacent areas meet at delineated edges. We use watershed segmentation (Levner & Zhang, 2007) to obtain the superpixels.
- Segment Anything Model (SAM) (Alexander et al., 2023) creates groups based on global context and relationships within the image using transformer-based architecture.
- Archipelago (Michael et al., 2020) perturbs different segments and assigns interaction scores to pairs of segments by their influences on the model's prediction.

## E ADDITIONAL RESULTS

We calculate quantitative metrics on the ImageNet dataset in Table 1. To match the number of groups, we take the average number of groups of SAM and archipelago for each example, usually 64, and apply it to the patch and superpixel methods. We then measure quantitative metrics for each group and take the average value for each image. For the consistency loss and stability loss, we use patches of size $16 \times 16$.

### E.1 BEST AND WORSE CASES

To compare the results of each method, we consider the best and worst-case scenarios for each metric. The best cases for all three metrics are all 0, and the worst is all 1. Regarding consistency and stability loss, the worst-case scenario may be calculated using another image. The lower bound for consistency might be calculated by using exactly the same embedding, and for stability, it should

[1] https://huggingface.co/google/vit-base-patch16-224

| Consistency loss | Average | Max | Min | Variance |
|---|---|---|---|---|
| Patch | **0.2013** | 0.5550 | 0.0061 | 0.0102 |
| Superpixel | 0.2185 | 0.5109 | 0.0043 | 0.0090 |
| SAM | 0.2551 | 0.7712 | 0.0028 | 0.0099 |
| Archipelago | 0.2159 | 0.8015 | 0.0021 | 0.0110 |

(a) Consistency Loss

| Contiguity Loss | Average | Max | Min | Variance |
|---|---|---|---|---|
| Patch | **0.0534** | 0.0542 | 0.0526 | 0.0000 |
| Superpixel | 0.0579 | 0.1018 | 0.0529 | 0.0000 |
| SAM | 0.0953 | 0.1118 | 0.0504 | 0.0004 |
| Archipelago | 0.0658 | 0.1118 | 0.0513 | 0.0004 |

(b) Contiguity Loss

| Stability Loss | Average | Max | Min | Variance |
|---|---|---|---|---|
| Patch | **0.4282** | 0.9680 | 0.0111 | 0.0610 |
| Superpixel | 0.4916 | 0.9720 | 0.0016 | 0.0476 |
| SAM | 0.5290 | 0.9680 | 0.0001 | 0.0579 |
| Archipelago | 0.5019 | 0.9756 | 0.0004 | 0.0504 |

(c) Stability Loss

Table 1: Statistics for Consistency, Contiguity, and Stability Loss. All the metrics range from 0 to 1, with the lower the better. We can see that Patch excels in all three metrics.

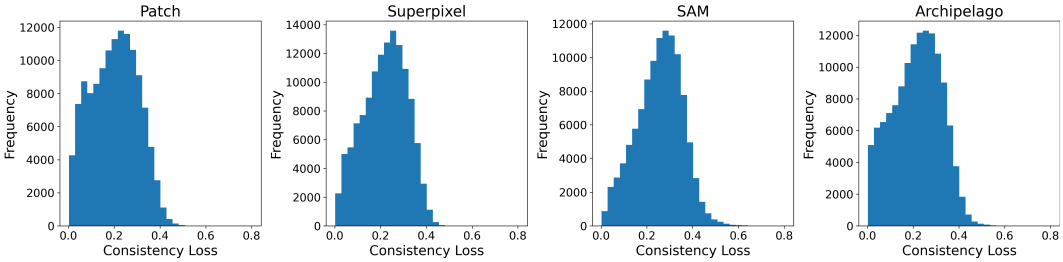
be calculated by having all the pixels in the image. The lower bound for the best possible sparse variation in square shape can be achieved by taking the square root of the number of pixels. On the other hand, the upper bound can be obtained by taking a square like a checkerboard. For instance, if a group of $k$ pixels exists, the best possible sparse variation would be taking the square with one side as the square root of $k$. However, the worst possibility would be taking the square with one side as the square root of $2k$; in such a case, if you choose a pixel to the group, avoid selecting the pixel adjacent to it, just like in a checkerboard.

### E.2 Results and Discussions

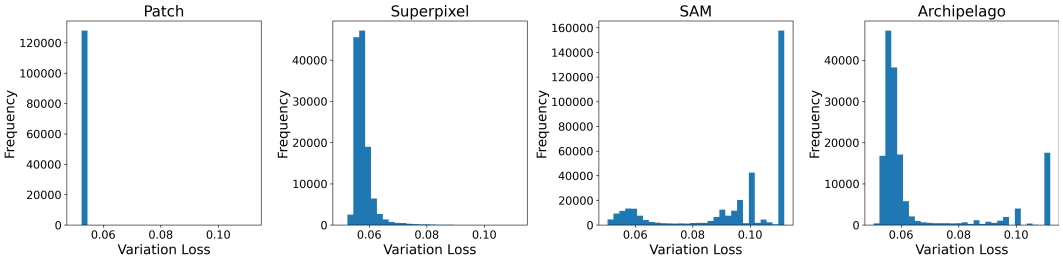
**Consistency.** Regarding the consistency loss plot of merging 1000 ImageNet classes in Figure 2a, all methods appeared similar to a skewed normal distribution with slightly different mean and variance. Table 1a also shows that the average consistency loss is similar across different methods. Both Figure 2a and Table 1a show that different grouping methods have similar embedding similarities on average.

However, Figure 2a shows that SAM and Archipelago have longer tails than patches and superpixels. This can also be verified by the high $\times 3$ to $\times 4$ values in maximum consistency loss for SAM and Archipelago in Table 1a, where patches and superpixels only have around $\times 2.5$ maximum loss compared to the average.

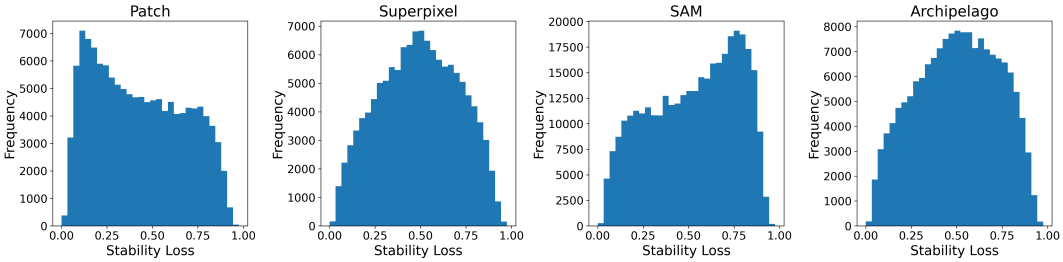
**Contiguity.** Histograms in Figure 2b show that patches and superpixels consistently have low contiguity loss, while SAM consistently has high contiguity loss. Archipelago has a dichotomy of majority low contiguity loss with a smaller peak for high contiguity loss. We can also see from Table 1b that SAM segments have higher average contiguity loss than the other three methods and that patches have lower maximum contiguity loss than the other three methods.




(a) Consistency Loss Plot of 1000 ImageNet Classes




(b) Contiguity Loss Plot of 1000 ImageNet Classes




(c) Stability Loss Plot of 1000 ImageNet Classes

Figure 2: Histograms of consistency, contiguity and stability loss for 1000 ImageNet classes. For each grouping method among patches, superpixels, SAM, and Archipelago, we take one image from each class and generate groups using the grouping method.

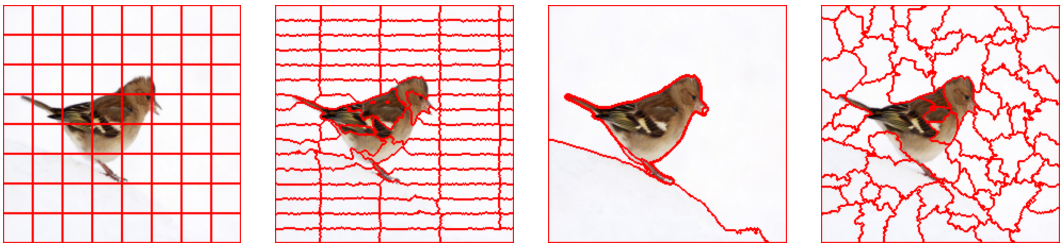
The fact that patches have the best average and maximum contiguity loss is likely because they do not have many edges. Conversely, SAM segments are the worst contiguous for both the average and the worst case, which is due to how it selects semantic segments that are diverse in shape.

**Stability.** Histograms in Figure 2c show diverse distributions of stability for all four methods. Patches skew to 0, while SAM segments skew toward 1. Table 1c also shows that patches have much smaller stability loss than the other three types of groups. Since the value of 0 represents the best-case scenario, patches are the most stable, while SAM segments are the least.

This could be attributed to the fact that the groups in SAM had varying sizes. As shown in Figure 3, some groups found by SAM are small and thin, having little contextual information. Therefore, adding one patch to a small group might result in a higher stability loss. Conversely, the patch groups have uniform sizes, allowing each group to have some contextual information.

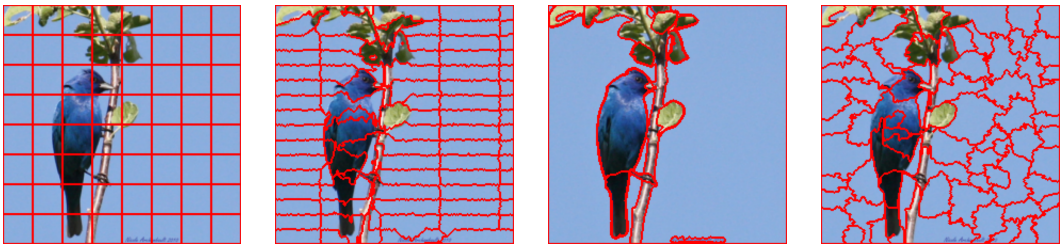
**Summary** The statistics in Table 1 and Figure 2 show that the patches have the minimum loss on average and, in the worst case, for all the properties. This indicates that the patches are the best type of group based on these desiderata.

Also, SAM and Archipelago have worse average stability and contiguity loss and worse worst-case consistency than patches and superpixels. Although SAM and Archipelago's Transformer-based



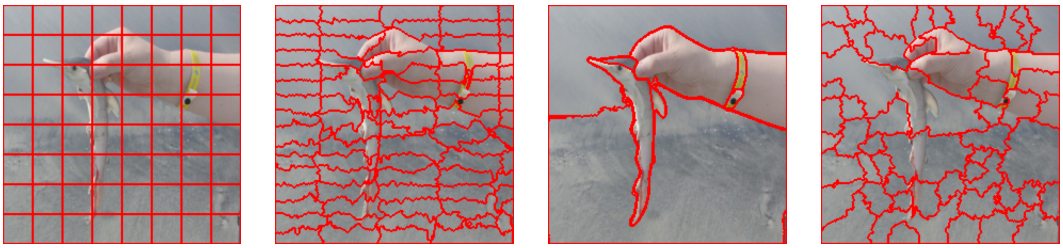
(a) Patch (b) Superpixel (c) SAM (d) Archipelago

(1) Example 1



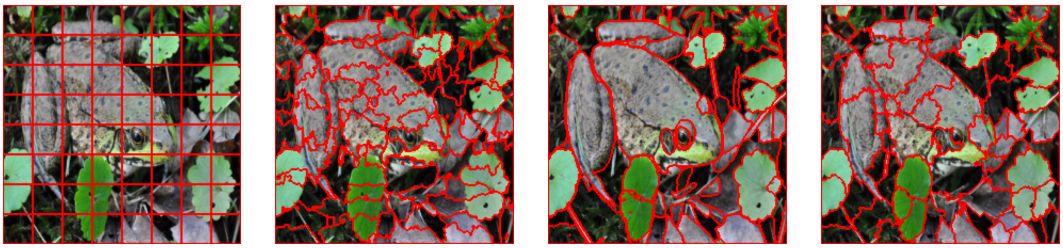
(a) Patch (b) Superpixel (c) SAM (d) Archipelago

(2) Example 2



(a) Patch (b) Superpixel (c) SAM (d) Archipelago

(3) Example 3



(a) Patch (b) Superpixel (c) SAM (d) Archipelago

(4) Example 4

Figure 3: Examples of group shape based on patches, superpixels, SAM, and Archipelago. The examples are from different classes. Groups made of patches are straight, whereas Archipelago patches are crooked.

architectures enable them to comprehend global context effectively, they are not well-suited for grouping when measured with the three desired quantitative metrics.