NoFADE: Analyzing Diminishing Returns on CO2 Investment

Andre Fu [*†] **Justin Tran** [*] **Andy Xie** [*] **Jonathan Spraggett**[*] **Elisa Ding** [*]

Chang-Won Lee[*] **Kanav Singla**[*] **Mahdi S. Hosseini**[‡] **Konstantinos N. Plataniotis**[*]

Abstract

Climate change continues to be a pressing issue that currently affects society at-large. It is important that we as a society, including the Computer Vision (CV) community take steps to limit our impact on the environment. In this paper, we (a) analyze the effect of diminishing returns on CV methods, and (b) propose a *"NoFADE"*: a novel entropy-based metric to quantify model–dataset–complexity relationships. We show that some CV tasks are reaching saturation, while others are almost fully saturated. In this light, NoFADE allows the CV community to compare models and datasets on a similar basis, establishing an agnostic platform.

1 Introduction

We are currently experiencing a surge in the development and implementation of neural architectures in the computer vision (CV) community. This is accompanied by exponential increases in both model complexity and accuracy, which are consistently correlated with performance dependent on computer power [1]. The compute required for training AI models follows an exponential trend with a 3.4 month doubling period [2], as state of the art (SOTA) models (eg. NAS, AlphaZero) utilize over $\times 0.3$M the compute of previous benchmarks (eg. AlexNet). The community is heavily focused on surpassing SOTA results. However, this will require further increases in compute power and downstream implications cannot be overlooked, particularly CO2 output of deep learning workloads.

The Intergovernmental Panel on Climate Change (IPCC) provides transparent reports on the effects of climate change to policy makers and the scientific community. In the 2019 IPCC report, the IPCC highlighted that the $1.5°$C above pre-industrial levels is required to mitigate extreme climate related events. In 2016, there was 52 GtCO2 with an expected 52-58 GtCO2 by 2030, with an IPCC recommended 25-30 GtCO2 [3] in the same time frame. Recently, the sixth IPCC report was released where they stressed that "global warming of $1.5°$C and $2.0°$C will be exceeded in the 21st century unless deep reductions in CO2 and other greenhouse gases occur in the coming decades" [3]. The climate crisis is a pressing issue that faces society at-large, and as such everyone, even those within the deep-learning community have their role to play to curb the climate crisis. It is crucial our community gains a better understanding of how our deep learning models affect the climate crisis, our responsibilities within the crisis, and improvements in our methodologies to include awareness of the crisis.

There has been no previous analysis on the topic within the deep learning and CV communities; therefore we propose and investigate the following questions:

[*]University of Toronto

[†]Corresponding Author: andre.fu@utoronto.ca

[‡]University of New Brunswick, mahdi.hosseini@unb.ca

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

1. How much CO2 is emitted compared to evaluation metrics for any model? Is there a saturation occurring within popular fields?
2. How can we gain a better understanding of the relationships between models and datasets on the basis of CO2 investment?

In answering these questions, we find that (1) there exists a diminishing return on increasing computation and equivalently CO2 emission and (2) we develop a metric to characterize the relationship between a model and the dataset's difficulty of learning while normalizing for complexity.

2 Diminishing Returns of Increasing Computation

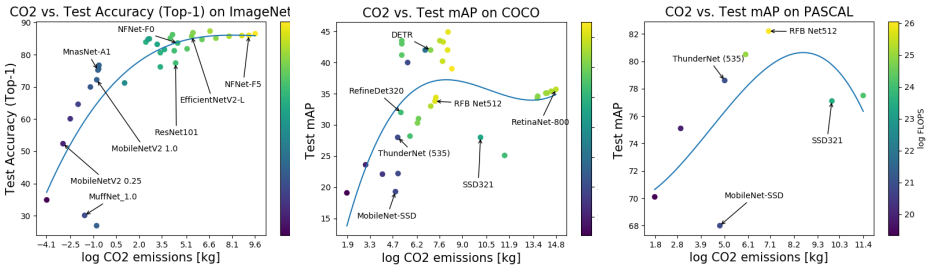
As deep learning grows in popularity, there is corresponding competition to improve upon SOTA methods. In doing so, the community often overlooks the implicit costs of chasing SOTA such as the increasing computational burdens [1, 4] which in turn have severe environmental consequences. These models necessitate growing computational footprints during training but also during their lifetime [4] as deployed systems. Here, we analyze the trade-offs between increasing computational cost in Watt-hours and environmental cost in CO2 emissions.

Inspired by [1, 4] methodologies, we begin by generating a corpus of computer vision models within the three most common tasks (a) classification, (b) segmentation and (c) detection. We surveyed 13 classification papers [5–17], 22 segmentation papers [18–39] and 10 detection papers [40–49]. For every model, we extracted the Top-1 Test-accuracy, mAP or mIOU, FLOPS, GPU hours and GPU type.

Using GPU type, the Watt-to-FLOPS was calculated using the model's FLOPS as f as ω = Watt/f. Then, the GPU Watt-to-FLOPS are denoted as ω_g and the CPU's Watt-to-FLOPS as ω_c [4]. We obtain a model's power draw, P_m over training by multiplying the model's flops, f by the sum of the Watt-to-FLOPS ratios. This quantity would be multiplied by the total GPU hours to train [4], seen below:

$$P_m[\text{Wh}] = f \times (\omega_g + \omega_c) \times \text{GPU hours}. \tag{1}$$

The final CO2 emissions are calculated by multiplying the power draw from a model by the EPA's Wh to CO2 measurement 0.707×10^{-3} metric tonnes/kWh, CO2 = $P_m \cdot 0.707 \times 10^{-3}$.



(a) ImageNet (b) COCO (c) PASCAL

(d) CityScapes (e) PASCAL VOC

Figure 1: (a) Classification task ImageNet [50], (b, c) Sematic Segmentation tasks MS COCO [51] and PASCAL [52], (d,e) Detection tasks CityScape [53] and PASCAL VOC [52]

Within classification, the ImageNet dataset [50] has the most distinct saturation as seen in Figure 1(a). This implies a clear diminishing return on 'investment' of CO2 emissions for gained accuracy. The classification task is the most mature, combined with long-term interest prior to deep-learning methods. To that end, it is justified that ImageNet reaches saturation as the field reaches optimal performance with minimal improvements in recent years. In fields such as Segmentation and Detection that are still developing, the saturation curve is observable but less well defined, as in Figure 1(b) and (d). Notice in all the subfigures the models with more FLOPS generally emit significantly more CO2, yet these high-FLOPS models have the same or lower performance in comparison to low-FLOPS models. On MS COCO, PASCAL, and PASCAL VOC, we see that it is possible to trade-off computational complexity to achieve SOTA performance by reducing the FLOPS constraint. For example, implementing ThunderNet(535) in lieu of SSD321 would achieve better test performance with less FLOPS, thus reducing carbon emissions without sacrificing results.

3 NoFADE Development

In the pursuit of increasing SOTA at a diminishing computational cost, we find that there currently is no concrete analysis of the dataset learning space in relationship to models. To identify if a model is learning well on a dataset, the difficulty of learning must be determined, and weighed with the relationship to test-accuracy. We begin by creating a collection of the unique model i and dataset j pairs $(\text{Model}_i, \text{Dataset}_j)$. For any two model-dataset pairs, it becomes difficult to compare their relative performances while taking into account difficulty of learning and computational complexity of a model. In this work, we attempt to characterize a novel metric, *NoFADE*: Normalized FLOPS for Accuracy-Dataset Entropy which allows the CV community to compare models in an effective way.

We use Shannon entropy as the underlying measure of difficulty of learning [54]. We begin by converting colour images to greyscale and using $H(\text{Image}) = -\sum_{i=1}^{K} p_i \cdot \log p_i$ to determine the entropy. The Image being analyzed has a frequency of pixel intensity p_i, with K unique intensities.

Segmentation and Detection Segmentation (Cityscapes, PASCAL) and Detection (MS COCO, PASCAL) datasets are comprised of a set of images with their reference true labels. As entropy represents the randomness within an image, we can assess the total randomness of the dataset. Therefore we apply Shannon Entropy to each image in the dataset to generate a normal distribution. As shown in [54] measuring a dataset's entropy distribution allows us to gain a better understanding of image complexity and most importantly the difficulty of learning.

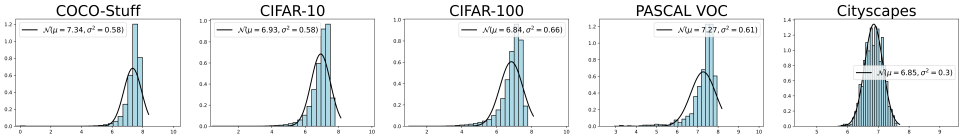


Figure 2: Entropy Histograms for Cityscapes, Pascal VOC, CIFAR10, CIFAR100 and MS COCO

Classification: Classification difficulty is inherently related not only to image complexity, but also multi-distribution learning as the classes are themselves distributions. To resolve this, we compute the entropy using Shannon Entropy for every class within CIFAR10, CIFAR100 and ImageNet. Each class distribution now represents the complexity of that class, but to gain an informational relationship about the dataset as a whole, we can measure the distances between classes. To do this we employ the Jensen-Shannon distance [55], a statistical measure that assesses the distance between two distributions in a symmetric and finite manner. Based on the Kullback-Leibler divergence, Jensen-Shannon distance allows us to gain understanding of mutual information between the two distributions.

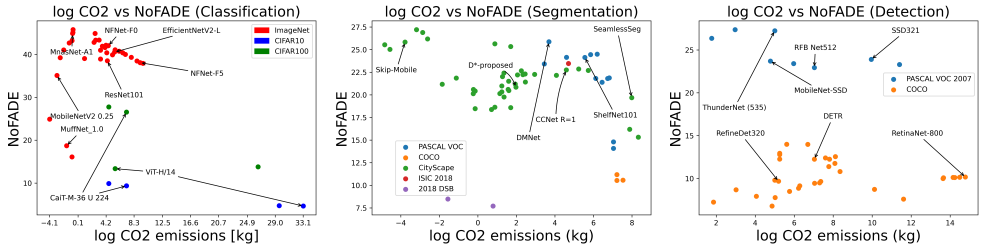
$$JSD(P\|Q) = \sqrt{\frac{1}{2}D_{KL}(P\|M) + \frac{1}{2}D_{KL}(Q\|M)}. \tag{2}$$

Here, $M = \frac{1}{2}(P + Q)$ and D_{KL} refers to the KL-divergence. Using Equation 2 we determine the distance between each class distributions, which represents the difficulty of learning between the two classes. Therefore, the sum of the Jensen-Shannon distances for the unique pairs of classes, provides

a measure of the complexity of the dataset as a whole, with higher values representing more difficult learning. In order to normalize the data to a similar scale as the entropy calculations, we take the log of the summation.

NoFADE attempts to characterize a relationship between models and datasets while normalizing for computational complexity. To do this, we propose multiplying the test-accuracy gained by the entropy of the dataset. As a dataset's difficulty of learning is correlated to it's entropy or JS-distance, we are effectively increasing the amount of "metric points" a model is gaining in proportion to the difficulty of learning on that dataset. For example a 1% increase on ImageNet is "better" in the community than a 1% increase on CIFAR10. To that end, we also recognize that the informal rule of "deeper performs better" may skew these results. To mitigate this risk, we normalize the "metric points" by the $\log$ FLOPS, as most of the FLOPS are tightly grouped the $\log$ allows us to exploit sensitivity. Doing so lets the metric understand that higher FLOP models have a constraint and thus deserve a normalization to allow for comparisons.

$$\text{NoFADE} = \frac{\text{test-accuracy} \times \text{Entropy or JS-dist}}{\log \text{FLOPS}} \tag{3}$$



(a) Classification (b) Segmentation (c) Detection

Figure 3: $\log$ CO2 vs. NoFADE, (a) Classification with ImageNet, CIFAR10, CIFAR100 (b) Segmentation with COCO, PASCAL, Cityscape, ISIC 2018 and DSB 2018, (c) Detection with MS COCO and PASCAL VOC

In Figure 3 we can compare methods in a models–dataset–complexity–agnostic manner, building the foundation for comparisons. For example, because the ViT transformer models have such large FLOPS they do perform extremely well but at a significant complexity cost which is reflected in the lower NoFADE score, their training times are also reflected with their extremely large CO2 footprints. Therefore, implying the transformer models don't perform well on these key dimensions. Similarly, some of the best models like MNasNet perform extremely well in all three categories, having low CO2 while maintaining a high NoFADE score.

4 Conclusion

In this paper, we demonstrate how our choices as the CV community have an impact on global climate change. Specifically, the selection of various architecture and dataset pairs were explored and how dataset complexity affects a model's efficiency and performance, resulting in increased CO2 emissions. We propose a new metric to quantify a dataset's complexity and validated as a reasonable method to track CO2 emissions arising from model training. We hope that this proposal will help the AI community be more environmentally conscious and to be as efficient as possible when training their models and utilize the novel tool to measure model–dataset–CO2 relationship to gain a better understanding of their place in the field. In addition, this measurement aid could facilitate model and dataset selection for performance and efficiency trade-off. Overall, with this new metric in hand, we anticipate researchers to start trying to reduce their environmental impact and hopefully find new ways to do so as well.

We'd like to thank Michal Fishkin for her worthwhile contributions without whom this work would not have taken place.

References

[1] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in nlp," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3645–3650, 2019.

[2] D. Amodei, "Ai and compute," Jun 2021.

[3] M. V., "Ipcc, 2021: Summary for policymakers. in: Climate change 2021: The physical science basis. contribution of working group i to the sixth assessment report of the intergovernmental panel on climate change," *Intergovernmental Panel on Climate Change, Working Group I Contribution to the IPCC Sixth Assessment Report (AR6)*, 2021.

[4] A. Fu, M. S. Hosseini, and K. N. Plataniotis, "Reconsidering co2 emissions from computer vision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2311–2317, 2021.

[5] A. Brock, S. De, S. L. Smith, and K. Simonyan, "High-performance large-scale image recognition without normalization," *arXiv preprint arXiv:2102.06171*, 2021.

[6] M. Tan and Q. V. Le, "Efficientnetv2: Smaller models and faster training," *arXiv preprint arXiv:2104.00298*, 2021.

[7] H. Chen, M. Lin, X. Sun, Q. Qi, H. Li, and R. Jin, "Muffnet: Multi-layer feature federation for mobile deep learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 0–0, 2019.

[8] T. Ridnik, H. Lawen, A. Noy, and I. Friedman, "Tresnet: High performance gpu-dedicated architecture. arxiv 2020," *arXiv preprint arXiv:2003.13630*.

[9] Z. Gao, L. Wang, and G. Wu, "Lip: Local importance-based pooling," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3355–3364, 2019.

[10] N. Ma, X. Zhang, and J. Sun, "Funnel activation for visual recognition," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pp. 351–368, Springer, 2020.

[11] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[12] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "Eca-net: efficient channel attention for deep convolutional neural networks, 2020 ieee," in *CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE*, 2020.

[13] M. Tan, B. Chen, R. Pang, V. Vasudevan, M. Sandler, A. Howard, and Q. V. Le, "Mnasnet: Platform-aware neural architecture search for mobile," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2820–2828, 2019.

[14] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 116–131, 2018.

[15] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 510–519, 2019.

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[17] H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, and H. Jégou, "Going deeper with image transformers," *arXiv preprint arXiv:2103.17239*, 2021.

[18] X. Li, Z. Zhong, J. Wu, Y. Yang, Z. Lin, and H. Liu, "Expectation-maximization attention networks for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9167–9176, 2019.

[19] H. Park, Y. Yoo, G. Seo, D. Han, S. Yun, and N. Kwak, "C3: Concentrated-comprehensive convolution and its application to semantic segmentation," *arXiv preprint arXiv:1812.04920*, 2018.

[20] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "Enet: A deep neural network architecture for real-time semantic segmentation," *arXiv preprint arXiv:1606.02147*, 2016.

[21] X. Hu and H. Wang, "Efficient fast semantic segmentation using continuous shuffle dilated convolutions," *IEEE Access*, vol. 8, pp. 70913–70924, 2020.

[22] B. Olimov, K. Sanjar, S. Din, A. Ahmad, A. Paul, and J. Kim, "Fu-net: fast biomedical image segmentation model based on bottleneck convolution layers," *Multimedia Systems*, pp. 1–14, 2021.

[23] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation. arxiv 2015," *arXiv preprint arXiv:1505.04597*, 2019.

[24] L. Porzi, S. R. Bulo, A. Colovic, and P. Kontschieder, "Seamless scene segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8277–8286, 2019.

[25] M. Gamal, M. Siam, and M. Abdel-Razek, "Shuffleseg: Real-time semantic segmentation network," *arXiv preprint arXiv:1803.03816*, 2018.

[26] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 325–341, 2018.

[27] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2337–2346, 2019.

[28] S. Mehta, M. Rastegari, A. Caspi, L. Shapiro, and H. Hajishirzi, "Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation," in *Proceedings of the european conference on computer vision (ECCV)*, pp. 552–568, 2018.

[29] L. Wang, D. Li, Y. Zhu, L. Tian, and Y. Shan, "Dual super-resolution learning for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3774–3783, 2020.

[30] H. Li, P. Xiong, H. Fan, and J. Sun, "Dfanet: Deep feature aggregation for real-time semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9522–9531, 2019.

[31] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnet: Criss-cross attention for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 603–612, 2019.

[32] N. Vallurupalli, S. Annamaneni, G. Varma, C. Jawahar, M. Mathew, and S. Nagori, "Efficient semantic segmentation using gradual grouping," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 598–606, 2018.

[33] X. Chen, Y. Wang, Y. Zhang, P. Du, C. Xu, and C. Xu, "Multi-task pruning for semantic segmentation networks," *arXiv preprint arXiv:2007.08386*, 2020.

[34] J. He, Z. Deng, and Y. Qiao, "Dynamic multi-scale filters for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3562–3572, 2019.

[35] J. Zhuang, J. Yang, L. Gu, and N. Dvornek, "Shelfnet for fast semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 0–0, 2019.

[36] C. Kaul, N. Pears, H. Dai, R. Murray-Smith, and S. Manandhar, "Focusnet++: Attentive aggregated transformations for efficient and accurate medical image segmentation," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 1042–1046, IEEE, 2021.

[37] P. Li, X. Dong, X. Yu, and Y. Yang, "When humans meet machines: Towards efficient segmentation networks.," in *BMVC*, 2020.

[38] Y. Li, Y. Chen, X. Dai, D. Chen, M. Liu, L. Yuan, Z. Liu, L. Zhang, and N. Vasconcelos, "Micronet: Towards image recognition with extremely low flops," *arXiv preprint arXiv:2011.12289*, 2020.

[39] T. He, C. Shen, Z. Tian, D. Gong, C. Sun, and Y. Yan, "Knowledge adaptation for efficient semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 578–587, 2019.

[40] Z. Qin, Z. Li, Z. Zhang, Y. Bao, G. Yu, Y. Peng, and J. Sun, "Thundernet: Towards real-time generic object detection on mobile devices," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6718–6727, 2019.

[41] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision*, pp. 213–229, Springer, 2020.

[42] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, pp. 91–99, 2015.

[43] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[44] Y. Li and F. Ren, "Light-weight retinanet for object detection," *arXiv preprint arXiv:1905.10011*, 2019.

[45] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

[46] Y. Chen, T. Yang, X. Zhang, G. Meng, X. Xiao, and J. Sun, "Detnas: Backbone search for object detection," *Advances in Neural Information Processing Systems*, vol. 32, pp. 6642–6652, 2019.

[47] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[48] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Single-shot refinement neural network for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4203–4212, 2018.

[49] S. Liu, D. Huang, *et al.*, "Receptive field block net for accurate and fast object detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 385–400, 2018.

[50] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[51] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*, pp. 740–755, Springer, 2014.

[52] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[53] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

[54] A. A. Rahane and A. Subramanian, "Measures of complexity for large scale image datasets," in *2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pp. 282–287, IEEE, 2020.

[55] B. Fuglede and F. Topsoe, "Jensen-shannon divergence and hilbert space embedding," in *International Symposium onInformation Theory, 2004. ISIT 2004. Proceedings.*, p. 31, IEEE, 2004.