Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/5/20230711

The advance of neural networks generalization performance

Haoyu Chen[1]

[1]Physical Sciences Division, University of Chicago, Chicago, 60611, USA

hchen99@uchicago.edu

Abstract. A big mystery in deep learning is the promising generalization performance generated by massive neural networks. While over-parameterization increases the tendency of overfitting in other machine learning models, neural networks seem to "magically" overcome this hurdle and achieve minor test errors in various tasks. Researchers are motivated to resolve this enigma through a variety of aspects and methods, both theoretically and empirically. This paper aims to comprehensively review the explanations for the generalization power of deep networks. Firstly, the review compares various types of generalization bounds under PAC-Bayes analysis and non-PAC-Bayesian settings. Then, works of regularizers, both explicit (e.g., dropout) and implicit (e.g., batch normalization), and algorithms-caused regularizations are reviewed in this work. Some researchers also explore networks' generalization ability from other perspectives, and this review talks about works that investigate the relationship between images and generalization performance. Additionally, works of adversarial examples are included in this review, since adversarial attacks have challenged networks' power to generalize well and have become an important field in understanding deep learning. By collecting works from different viewpoints, this paper finally discusses some possible directions in the future.

Keywords: Neural Networks, Generalization, Adversarial Examples, Regularization, Generalization Bounds.

1. Introduction

Massive deep neural networks perform exceptionally well across various application domains. They are capable of memorizing training datasets and, more importantly, generalizing well to test data. Researchers have been analyzing the reasons behind the generalization ability of neural networks. Constructive explanations would help better understand how neural networks work and thus make model designs more appropriate and efficient. Several traditional complexity measures are applied to limit the generalization error, including the VC dimension and Rademacher complexity. However, Zhang et al. challenge these measures in their ability to explain the power of generalization in neural networks [1]. Through experimentation, they claim that over-parameterized networks can memorize every training instance, even on randomized labels. This observation contradicts our traditional view that over-parametrization will lead models to overfit. The community was then fuelled to find explanations for such a phenomenon to understand the mechanism behind deep neural networks. In this review, works from different perspectives were separated into four main groups: generalization bounds, regularization, images and data distribution, and adversarial examples.

DOI: 10.54254/2755-2721/5/20230711

2. Generalization bounds

To comprehend neural networks' generalization capability, a direct approach is to derive tight generalization bounds. Besides calculating the bounds directly from some quantities of the training set, the notion of complexity measure is predominantly utilized as a core component in tightening bounds. These measures include both theoretical (e.g., VC dimension and norms of weights) and empirical ones (e.g., sharpness and Fisher-Rao norm) [2, 3]. Many works are defined under the PAC-Bayesian setting, while some bounds are characterized through other analyses like the covering number [4].

2.1. Non-PAC-Bayesian analysis

The predominant role of norm or margin is reflected by many works in the field. The Lipschitz constant, the result of multiplying every layer's weights in spectral norm, is widely used to derive generalization bounds, but it continues to grow even after the excess risk stabilizes [4]. To overcome this issue, Bartlett et al. derive a bound based on Lipschitz constant divided by the margin, which is the difference between predictions and the actual target variable. This bound is distinct from previous Lipschitz-sensitive and margin-based bounds in its dependence on the spectral norm of the weights. Based on the proposed bounds, Bartlett et al. further examined the margin distributions of the MNIST, CIFAR-100, and CIFAR-10 datasets. The first key finding is that CIFAR-10 is "harder" than MNIST because MNIST can be easily fitted by many models. This finding is consistent with the common view of these two datasets because MNIST contains only images of digits, while CIFAR-10 has greater complexity and variety. Secondly, the observation suggests that randomizing MNIST and randomizing CIFAR-10 result in similar hardness. Another interesting phenomenon is that the randomization of labels in CIFAR-10 makes it almost as hard as the original CIFAR-100 dataset, and it is known that these two datasets share the same images, but CIFAR-10 has a reduced number of categories. The researchers also claim that input randomization makes datasets harder to learn than random labeling [4].

Researchers have explored other norm-based bounds. Neyshabur et al. introduce the definition of path norm in their work on optimizing Stochastic Gradient Descent (SGD) [5]. The path norm calculates the weight summed along a route between the input unit and the output unit. After that, Neyshabur et al. investigated several measures based on different norms, including l_1 norm, l_2 path norm, l_1 path norm, and spectral norm [6]. The results show that these measures can all capture the gaps between random labels and real data, and the differences increase as the training set enlarges. l_2 norm and l_2 path norm seem to have the greatest gaps between true labels and random labels, and their sensitivities to the number of training instances also appear to be the most significant. Kawaguchi et al. further leverage the path norm under a different problem setting: they contend that determining whether a model generalizes effectively to new data requires evaluating its performance on the validation set [7].

Another novel norm, Fisher-Rao, is defined by Liang et al. in their work to analyze the relationship between geometry and complexity measures [3]. Inspired by information geometry, the Fisher-Rao metric corroborates the geometric invariances with norm and flatness, and such a measure can tackle some remained questions from previous research. The invariance is aimed to reflect that a network's ability to generalize should not be affected by a particular parametrization [3]. Also, previous works have failed to include invariance in their definitions of flatness for loss functions; hence, Liang et al. insist that this void needs to be filled to prove that flatness is related to the generalization performance. The Fisher-Rao norm is intended to represent the bottom bound of the path norm, but Jiang et al. suggest that it demonstrates inferior performance than the path norm since it poorly reflects the correlation between depth and generalization [8]. Besides, Jiang et al. reveal Fisher-Rao's inability to reflect the interactions between different hyperparameters and state that they can only reveal the change in one hyperparameter [8].

2.2. PAC-Bayesian analysis

Over the years, many works have been conducted under the PAC-Bayesian setting. Neyshabur et al. test the role of sharpness in explaining generalization ability [6]. Sharpness indicates the robustness of empirical risks to perturbations in the parameter space. Neyshabur et al. state that sharpness is variant

DOI: 10.54254/2755-2721/5/20230711

to the scale of parameters and does not perform well on smaller neural networks; thus, sharpness solely is not enough to measure the capacity of a model [6]. However, combinations of sharpness and norms are suggested to be effective in capturing generalization performance and can be used to derive tighter bounds [6, 9]. The work of Dziugaite and Roy also notes the relationship between sharpness and capacity control [9]. They obtain a non-vacuous bound by optimizing the PAC-Bayes bound across Gaussian distributions.

Neyshabur et al. also propose generalization bound scales by the multiplication of spectral norms, which is similar to the one provided by Bartlett et al..[10]. The distinction is that Neyshabur et al. verify their bound using the PAC-Bayesian analysis, and the bound is dependent on the Frobenius norm in each layer, which is at first better than the bound in Bartlett et al.'s work. However, after the initial preprint of Bartlett et al.'s work, they improve their result by replacing the l_1 norm with the $l_{2,1}$ norm, which is strictly tighter than the Frobenius norm used by Neyshabur et al. [10].

Motivated by the failures of some bounds on improving generalization performance empirically, some researchers derive a non-vacuous bound that requires the compressibility of neural networks [11]. Compared to the work of Dziugaite and Roy, this bound is non-vacuous not only on the MNIST dataset but also on the ImageNet dataset. Zhou et al. also take robustness into account since they assert that networks are often insensitive to weights [12]. However, a limitation of this work is its requirement for deep compressible networks. Arora et al. expand the notion of compressibility to a framework for verifying generalization bounds [13]. Built on top of this framework, they propose a bound that depends on the noise stability of a network that is proven to be effective in practice [14]. But their achievements are also confined to compression techniques.

2.3. Limitations

Many existing generalization bounds suffer from several common issues: sensitivity to data complexity, vacuousness, dependency on network size, robustness, computational expensiveness, and constraints on algorithms. Although each bound has tackled one or several of the limitations, like non-vacuous bounds [9,11] and the size-independent bound [15], none of them can escape from all the restrictions. More importantly, the existing generalization bounds are still not tight enough. With that being said, advanced developments and further studies are still needed.

3. Regularization

3.1. Explicit regularization

To overcome the tendency of overfitting, regularizers have been widely utilized to control networks' effective capacity. Naturally, regularizations are regarded as part of the reasons that make deep networks generalization well. However, experiments conducted by Zhang et al. reveal that explicit regularizers (dropout, weight decay, and data augmentation) can only help to improve the generalization errors to a limited extent and should not be considered as the explanation for the story of generalization [1]. This statement is consistent with Bartlett et al.'s work: there is no visible enhancement of margin distributions under regularizations [4]. Arpit et al. carry out experiments and indicate that explicit regularizers impede deep networks from memorizing noise but do not hinder their learnability in real data [16]. However, Liu et al. take another viewpoint on this problem: they analyze the role of explicit regularizers in search dynamics instead of as the stabilizer to fluctuations of examples [11]. Suggested by experiments on several datasets (CIFAR-10, CIFAR-100, CINIC10, and restricted ImageNet), the authors conclude that explicit regularizations (data augmentation and l_2 regularization) play a very important role in generalization because they will drive SGD away from bad minima in the searching process [10].

3.2. Implicit regularization

Zhang et al. show in practice that implicit regularizations like early stopping and batch normalization are insufficient to express the generalization power [1]. But they have discovered the implicit regularization induced by SGD, followed by many researchers continuing to explore such implicit bias.

DOI: 10.54254/2755-2721/5/20230711

Zhang et al. examine linear models to see if there are pertinent conclusions that can be extended to networks. Through a theoretical approach, they formulate that SGD often converges to the solutions with the minimum norm [1]. Shah et al. study the implicit bias in linear regression [14]. Even though their results are consistent with Zhang et al.'s finding that SGD does converge to the minimum norm solutions, they also claim the adaptive algorithms converge to a different solution that is better than the minimum norm solutions. Similar observations are detected in neural networks, but no clear conclusion is drawn in this work.

The implicit bias of gradient descent is also evaluated in logistic regression models. Soudry et al. observe that logistic loss and exponential loss tend to find the max-margin solution on separable data [17]. Besides, the experimental results show that the models converge at a very slow rate, which explains why optimization after obtaining a very small loss is beneficial. The same examination is applied to a network in a constrained setting. Although the results are similar, a defined conclusion cannot be stated until more unrestricted experiments are performed [17]. Arora et al. take a step further toward understanding the implicit regularization in deep networks [18]. To mitigate the data requirements in previous works, Arora et al. exploit the randomness of the matrix completion problem and investigate the solution of matrix factorization which can be viewed as a two-layer network. Their core observation is that the tendency toward low-rank answers is positively correlated with the depth of a network. Moreover, the authors put a call on the community that existing expressions of regularization cannot fully assess the implicit tendency induced by learning algorithms [18]. In summary, many efforts are put into characterizing the implicit regularization originating from learning functions, but we still lack a satisfying answer in the field of neural networks, and we need additional research to better understand the system.

4. Images and data distribution

Other than direct perspectives in explaining why networks generalize well, researchers also attempt to understand the process from other standings. Some of them explore the significance of data distribution in the story of generalization. Feldman employs the notion of long-tail distribution to elaborate on the memorization of neural networks [19]. Modern datasets appear to be long-tail distributed since the frequencies of rare instances are long-tailed, as shown in Figure 1. Feldman insists that the main obstacle for networks to learn well is not the noise from labels but rather the scarcity of the amounts of atypical examples. The results also demonstrate that a model built for learning from inadequate subpopulations may generalize almost perfectly [19]. Such insights can be further analyzed to link models with recent datasets and help to better design networks.

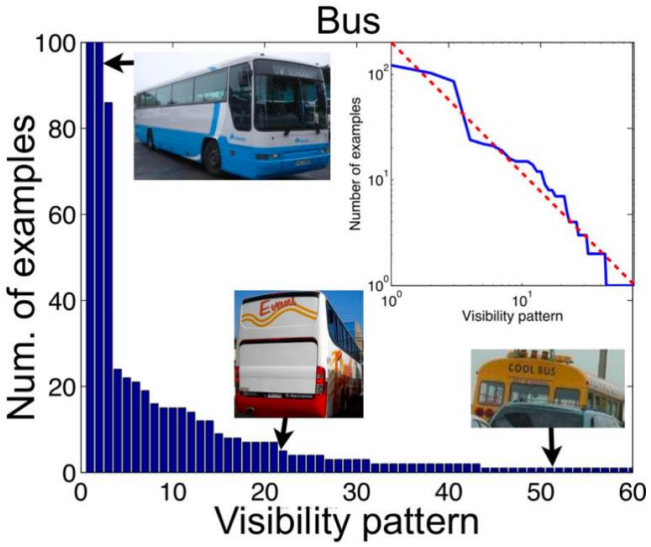

Figure 1. Example of dataset with long tail distribution [19].

DOI: 10.54254/2755-2721/5/20230711

The nature of images is also appraised as a necessary factor to minimize generalization errors. Wang et al. discover the differences between networks and human eyes in perceiving images: networks can capture the high-frequency components in images while humans cannot [20]. This finding is verified on datasets, and one example is selected from the work. As shown in Figure 2, the model successfully classifies the image with only high-frequency components in the third column while failing to predict the low-frequency image in the second column [20]. However, the second image is much easier to classify for humans because we can detect the semantic component of images. This perceptional disparity between humans and networks helps to explain why networks fit into random labels. The roles of some heuristics (e.g., batch normalization and mix-up) should thus be reconsidered since the observations suggest that these techniques stimulate models to capture high-frequency components [20]. The authors also leverage their findings into the study of adversarial examples, which is an increasingly significant field in deep learning and is further discussed in the next part.

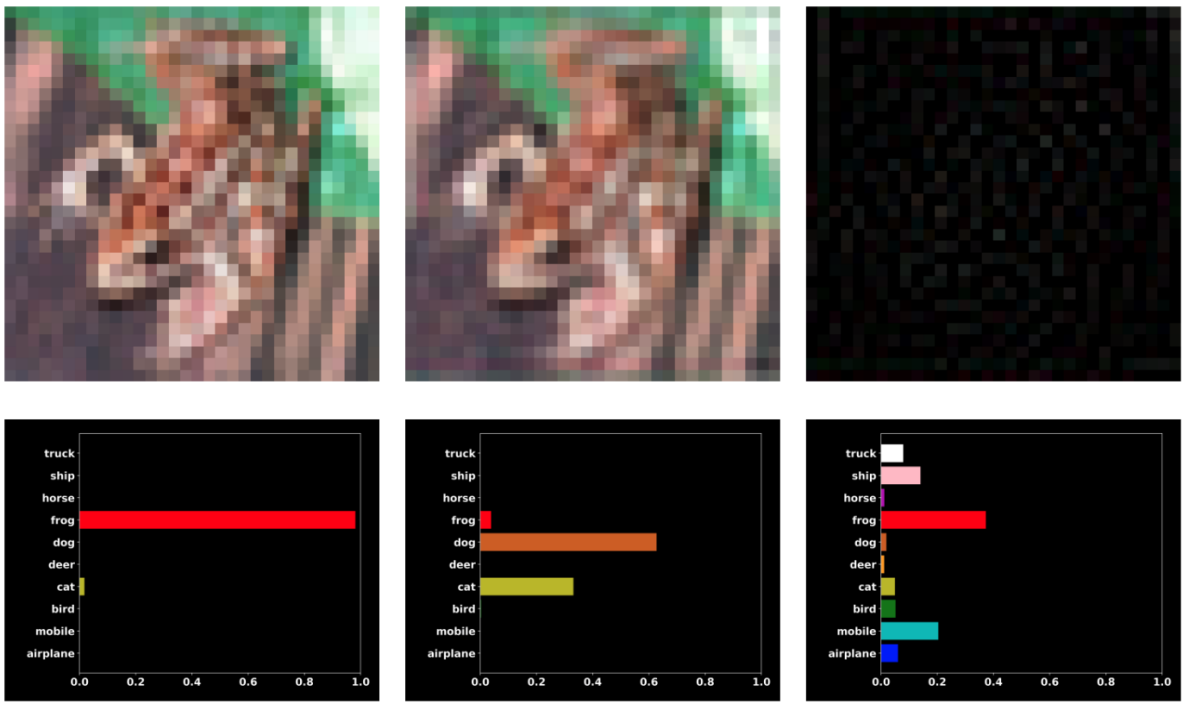

Figure 2. Example FROM CIFAR-10 that Explains High-Frequency Components [20].

5. Adversarial examples

Despite the great success that neural networks have achieved, researchers have discovered that neural networks suffer from small perturbations in input data, namely adversarial examples [21]. Even though imperceivable to human eyes, adversarial examples lead state-of-the-art models to misclassifications. Such vulnerability reveals blind spots in our understanding of neural networks' generalization performance. Many efforts have then been paid to overcome adversarial attacks, such as adversarial training, preprocessing-based methods, and generative-model-based frameworks. Among these approaches, Adversarial training [22, 23] has emerged as the most effective strategy for defending against adversarial cases and enhancing model robustness [24]. The fundamental concept of adversarial training is to incorporate image perturbations during training loops. Based on this idea, researchers follow up to improve and expand adversarial training from various perspectives. Another crucial area for deriving the generalization effectiveness is the model robustness against adversarial instances. Gao et al. analyze the reasons behind the success of adversarial training and support that wider models are required for more robust performance [24]. Wu et al. further the study about network width: they claim that width is positively correlated to natural accuracy, but wider networks will reduce perturbation

DOI: 10.54254/2755-2721/5/20230711

stability [25]. Other works have taken different views on understanding adversarial examples and adversarial training [22].

Even though Adversarial Training has shown effective improvements in networks' performance on adversarial attacks, researchers reveal some limitations of this method. A trade-off discovered is that adversarial training can improve adversarial accuracy but hurt standard accuracy (accuracy on clean data) [26]. Although many works focus on explaining this phenomenon, researchers have not yet reached a consensus. Besides, researchers also observe the time-consuming nature of adversarial training, owing to the generation of perturbations and extra training epochs needed [23]. Through systematic experiments, Yang et al. claim that new approaches are needed to achieve robust and accurate performance because existing methods, including adversarial training, might not be sufficient .Overall, adversarial attacks have questioned networks' ability to generalize well and draw more attention to the overfitting problem of deep networks.

6. Discussion

Indeed, many other works cannot be simply grouped into these sections. Other perspectives have been explored, such as model capacity, random initialization, and gradients. To conclude, researchers have made much progress in both theoretical understanding and empirical studies on the generalization of networks. While there is no universal answer yet, these findings are still constructive and insightful. However, many works fail to include networks' performance in adversarial examples, and the definition of generalization is often confined to test performance on unperturbed data only, which leaves the performance gap in adversarial training ill-understood. Also, a better understanding of overfitting in neural networks is needed after the existence of adversarial attacks, since many traditional regularization techniques do not help much in reducing adversarial overfitting. Furthermore, evaluating how adversarial examples impact traditional machine learning models may generate parallel insights to help understand deep networks, but the community lacks analyses in this area. In summary, the generalization performance of large neural networks is very complicated, and its scope needs to be expanded. A comprehensive understanding of deep networks needs to combine works across areas and from different time periods. The community looks forward to seeing more developments in this field and hopes the mystery can be fully explained one day.

7. Conclusion

In this paper, works related to explanations of neural networks' generalization performance are reviewed. In the first part, recent progress in calculating tight generalization bounds is separated into two groups and further compared by their properties. Both differences and similarities among these bounds are discussed. Although constructive, there exist some limitations in these state-of-the-art generalization bounds, such as vacuousness, sensitivity to network width and depth, sensitivity to data, robustness, computing cost, and algorithm restrictions. Both implicit and explicit regularizers have proven to be a minor factor of generalization, but implicit regularization originated from learning algorithms is revealed to play an important role, and this paper reviews some works that investigate the implicit bias of different models. Also, works on the role of data are discussed, including characteristics of images and distributions of data. Besides, this paper summarizes progress in adversarial attacks and defenses, and thus presents several questions and potential orientations in the study of neural networks' generalization ability: the definition of generalization should take account of adversarial performance and should be expanded; it is necessary to well-understand neural network's tendency to overfit in adversarial attacks; more research of conventional machine learning models' performance against adversarial examples are needed; the community can benefit from incorporating multi-dimensional and multi-perspective works to understand the story of networks' generalization comprehensively.

References

[1] Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. Communications of the ACM, 64(3), 107-115.

DOI: 10.54254/2755-2721/5/20230711

[2] Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P 2016 On large-bat ch training for deep learning: Generalization gap and sharp minima Preprint arXiv:1609.0483 6
[3] Liang, T., Poggio, T., Rakhlin, A., & Stokes, J. (2019, April). Fisher-rao metric, geometry, and c omplexity of neural networks. In The 22nd international conference on artificial intelligence and statistics (pp. 888-896). PMLR.
[4] Bartlett, P. L., Foster, D. J., & Telgarsky, M. J. (2017). Spectrally-normalized margin bounds fo r neural networks. Advances in neural information processing systems, 30.
[5] Neyshabur, B., Salakhutdinov, R. R., & Srebro, N. (2015). Path-sgd: Path-normalized optimizati on in deep neural networks. Advances in neural information processing systems, 28.
[6] Neyshabur, B., Bhojanapalli, S., McAllester, D., & Srebro, N. (2017). Exploring generalization i n deep learning. Advances in neural information processing systems, 30.
[7] Kawaguchi, K., Kaelbling, L. P., and Bengio, Y 2017 Generalization in deep learning Preprint a rXiv:1710.05468
[8] iang, Y., Neyshabur, B., Mobahi, H., Krishnan, D., and Bengio, S 2019 Fantastic generalizatio n measures and where to find them Preprint arXiv:1912.02178
[9] Dziugaite, G. K., and Roy, D. M 2017 Computing nonvacuous generalization bounds for deep (s tochastic) neural networks with many more parameters than training data Preprint arXiv:170 3.11008
[10] Neyshabur, B., Bhojanapalli, S., and Srebro, N 2017 A pac-bayesian approach to spectrally-nor malized margin bounds for neural networks Preprint arXiv:1707.09564
[11] Liu, S., Papailiopoulos, D., and Achlioptas, D 2020 Advances in Neural Information Processing Systems 33 8543-8552
[12] Zhou, W., Veitch, V., Austern, M., Adams, R. P., and Orbanz, P 2018 Non-vacuous generalizati on bounds at the imagenet scale: a PAC-bayesian compression approach Preprint arXiv:1804. 05862
[13] Arora, S., Ge, R., Neyshabur, B., and Zhang, Y 2018 In International Conference on Machine L earning (pp. 254-263) PMLR
[14] Shah, V., Kyrillidis, A., & Sanghavi, S. (2018). Minimum norm solutions do not always general ize well for over-parameterized problems. stat, 1050, 16.
[15] Golowich, N., Rakhlin, A., and Shamir, O. 2018 In Conference On Learning Theory (pp. 297-29 9) PMLR
[16] Arpit, D., Jastrzębski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M. S., ... and Lacoste-Juli en, S 2017 In International conference on machine learning (pp. 233-242) PMLR
[17] Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S., and Srebro, N 2018 The Journal of Machi ne Learning Research 19(1) 2822-2878
[18] Arora, S., Cohen, N., Hu, W., and Luo, Y 2019 Advances in Neural Information Processing Sys tems 32
[19] Feldman, V 2020 In Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing (pp. 954-959)
[20] Wang, H., Wu, X., Huang, Z., & Xing, E. P. (2020). High-frequency component helps explain th e generalization of convolutional neural networks. In Proceedings of the IEEE/CVF Confere nce on Computer Vision and Pattern Recognition (pp. 8684-8694).
[21] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R 201 3 Intriguing properties of neural networks Preprint arXiv:1312.6199
[22] Goodfellow, I. J., Shlens, J., and Szegedy, C 2014 Explaining and harnessing adversarial examp les Preprint arXiv:1412.6572
[23] Das, N., Shanbhogue, M., Chen, S. T., Hohman, F., Chen, L., Kounavis, M. E., and Chau, D. H 2017 Keeping the bad guys out: Protecting and vaccinating deep learning with jpeg compress ion Preprint arXiv:1705.02900

DOI: 10.54254/2755-2721/5/20230711

[24] Gao, R., Cai, T., Li, H., Hsieh, C. J., Wang, L., and Lee, J. D 2019 Advances in Neural Information Processing System, 32
[25] Wu, B., Chen, J., Cai, D., He, X., and Gu, Q 2021 Advances in Neural Information Processing Systems 34 7054-7067
[26] Liao, F., Liang, M., Dong, Y., Pang, T., Hu, X., and Zhu, J 2018 Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1778-1787)