A Comprehensive Survey of Evaluation Metrics for Domain Gap Minimization

Author: Mr. Sandip Omprakash Patil, Johndeere | Co-author: Girish Kumar Bellenavar, Johndeere

Abstract

In recent years, domain gap reduction has emerged as a crucial challenge in the field of computer vision. Domain gap reduction techniques are being employed in verity of computer vision applications like object detection and segmentation, Image style transfer and simulation to real etc. As the success of many vision models hinges on their generalization capabilities across different domains hence quantifying the performance of domain adaptation techniques is essential to assess their effectiveness in reducing the domain gap. In this paper, we present a systematic literature review on different techniques used for domain adaptation and analyze various methods used for quantifying domain gap. We commence the review with introduction and applications of domain adaptation and literature survey on different techniques used for domain gap minimization. Subsequently we discuss the methodology and evaluation techniques used to quantify domain gap followed by the merits and drawbacks associated with each approach.

Through our comprehensive survey we reveal the progress made in quantifying domain gap reduction enabling researchers to make informed decisions when selecting appropriate evaluation technique for domain gap minimization.

Introduction

In recent times, we have seen how artificial intelligence (AI) and machine learning (ML) have made it possible for intelligent systems to be used in various real-world applications. However, there is a significant challenge that needs to be overcome in implementing these systems - the domain gap. This refers to the difference between the data used for training, which is simulated, and the real-world data that the model is expected to work with. It is important to reduce this domain gap in order to ensure that AI systems are reliable and can withstand real-world scenarios. Even though we are making progress in understanding the difference between different domains, there is still a lack of research when it comes to thoroughly analyzing the evaluation metrics that are specifically designed to minimize this domain gap. The purpose of this paper is to fill this gap by conducting a detailed analysis of various evaluation metrics. The reason why we are conducting this analysis is because we want to have a deep understanding of how well metrics work, so that researchers and practitioners can make smart choices when trying to create strong AI models. The main goal of this research is to carefully examine the evaluation metrics that are currently being used to minimize the differences between different domains. We intend to evaluate these metrics in a variety of contexts to expose their limitations and benefits to provide insightful data that will improve performance in domain gap reduction.

Background and Motivation

The concept of domain gap adaptation has come up as an important field of study and research. Domain adaptation refers to the task of adjusting a model that was trained in one domain so that it can work well in another related domain. This usually comes into play when there is difference between original data where the model was trained and the new domain where it needs to perform properly [1]. The primary objective of domain adaptation is to reduce the disparity between these two domains, enabling the model to demonstrate its efficacy in the new domain. The crucial skill of efficiently bridging the gap between diverse domains and adjusting models and algorithms to unfamiliar data holds immense significance in numerous applications, like computer vision, natural language processing, and healthcare [2][3][4]. As the organizations are striving to leverage the power of ML and AI, the need for robust domain adaptation techniques has becoming increasingly evident [5][6]. By minimizing domain gap, we can enhance performance of the models when applied on data that diverges from their original training domain. This is especially important in scenarios where the target domain has different features than the source domain. Domain gap adaptation holds substantial importance as it allows the application and utilization of ML models and algorithms across the different domains.[6][7][8]. Many domain adaptation methods use adversarial training to improve the robustness of a model by employing adversarial examples. In recent years generative adversarial networks (GANs) are admired for their excellent generative capabilities. Goodfellow et al. [9] proposed this technique in 2014. Since then, countless papers published on using the GANs for verity of syntenic data generation applications [10][11]. One of the major applications is using these generative models for applications like Sim2Real.Many researchers have showcased ability of using GANs for such applications [12][13][14][15]. One key challenge that is consistent for all these methods is to quantify the model performance. Most common metrics which are mainly used to assess the GANs is Fréchet Inception Distance (FID) and Kernel Inception Distance (KID) [16][17]. Both the metrics are used to quantify the quality of generated samples by GANs. We are utilizing FID and KID as foundational metrics to assess the performance of GAN models. These metrics serve as key benchmarks in our evaluation process.

We have selected three distinct implementations of FID and KID i.e., Torchmetrics, clean FID and Pytorch FID [18][19][20], and we will be conducting a comprehensive quantitative evaluation of these implementations. This will encompass a systematic and thorough assessment of their performance, ensuring a data-driven analysis of the capabilities and limitations of each implementation.

Domain Gap Minimization

Domain gap minimization is a technique that aims to reduce the difference, between the source domain and the target domain. As depicted in figure 1 there is a shift between the source and target domains, which leads to lower performance of the source classifier on the target domain. By minimizing this gap, we can align the source and target domains resulting in improved performance of domain classifier. In this study our goal is to measure the disparity between real data domains. We aim to enhance the quality of data by reducing the domain gap between simulation and real data. There are many approaches that can be used to accomplish this objective.

Domain Adaptation: This approach aims to align the model with the target domain by minimizing the discrepancy between source and target data domains. Adversarial training and domain adversarial neural networks are some of the examples of domain adaptation.

Transfer learning: This approach includes refines the model that has already been trained on the source domain and applies it to the target domain. The model leverages the knowledge acquired in the source domain and endeavors to adjust to the target domain.

Augmentation: During training process, various transformations are applied to simulated data, including geometric transformations such as scaling, rotation and color adjustments. These transformations make the model more robust and able to handle a variety of real-world changing scenarios.

Self-supervised learning: Model trained with this technique predicts certain properties of data without explicit labels. Such model

leverages the learnings to learn some useful representations that can be transferable to target domain.

We will be conducting comprehensive analysis on evaluation metrics used for this by focusing on performance evaluation of different evaluation metrics used in general.

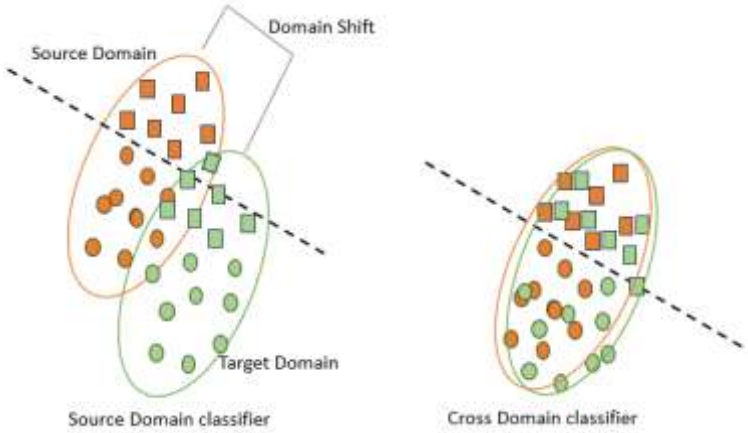

Figure 1. Domain Gap Minimization: Aligning the source and target domain features improves cross domain classification.

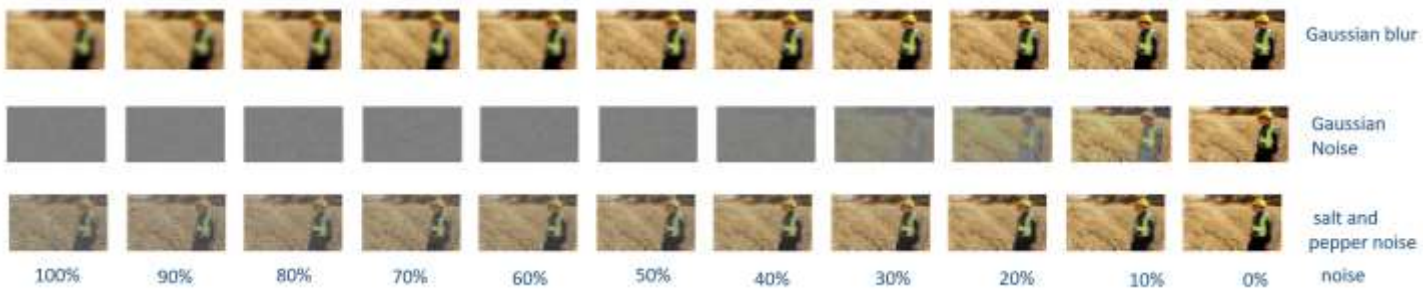

Figure 2 Systematic noise variations in the image: Gaussian noise, Gaussian blur and Salt and paper noise

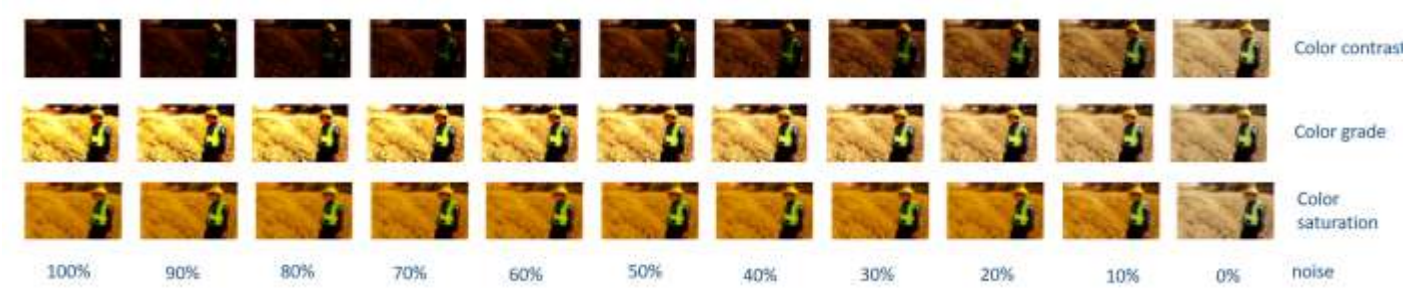

Figure 3 Systematic color variations in the image: color contrast, color grade and color saturation

Method

Overview

The objective of this study is to rigorously evaluate the effectiveness of domain gap minimization techniques, with a particular focus on the quantitative measures of FID and KID. To achieve this goal, a comprehensive image dataset representative of the target domain is utilized, and stringent preprocessing procedures, such as standardization and normalization, are implemented. Our experimental design involves the systematic application of three distinct types of noise Gaussian noise, Salt and Pepper noise, and Gaussian blur across a range of intensities, and also manipulating color-related parameters, such as contrast, saturation, and color grades, to shed light on their impact on quantifying domain gap by these metrics. To minimize the impact of potential biases, a randomization protocol is implemented. Evaluation metrics, including FID and KID, are rigorously calculated for each experimental condition, establishing benchmark metrics.

Data Preparation

The experimental setup is based on a factorial arrangement that systematically integrates different noise levels and color variations. To minimize the impact of potential bias. Only a few data samples are selected from the real data distribution in this case it is from roboflow dataset [22]and as mentioned in the methodology, noise and color variations are systematically added to the data. Refer figure 2 and figure 3. In addition, this advanced data is used to evaluate the performance of FID and KID by recording the variation of the metrics according to the different levels of noise and color variation.

Evaluation

Performance evaluation of the selected metrics is conducted by plotting the behavior of FID and KID implementations in response to noise and color variation. This evaluation offers valuable insights into the sensitivity and robustness of these metrics in the presence of image distortions. The impact of noise on FID and KID scores can indicate whether the metric is capable of capturing dissimilarities between generated and reference images. Ultimately, determining the thresholds at which FID and KID scores begin to exhibit significant deviations helps to establish the limits of noise tolerance for these metrics.

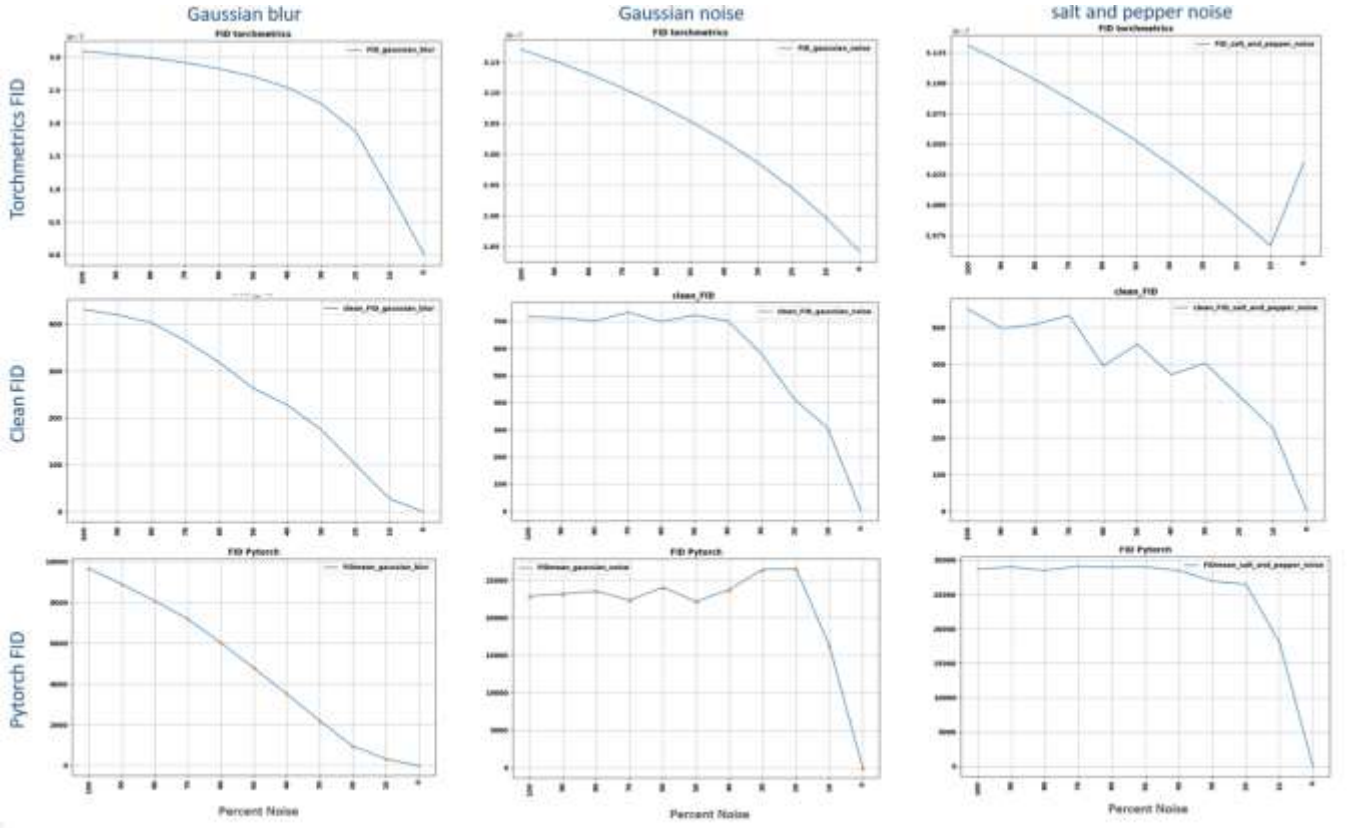

Figure 4 Torchmetrics FID, Clean FID and Pytorch FID implementation performance plots for noise variations

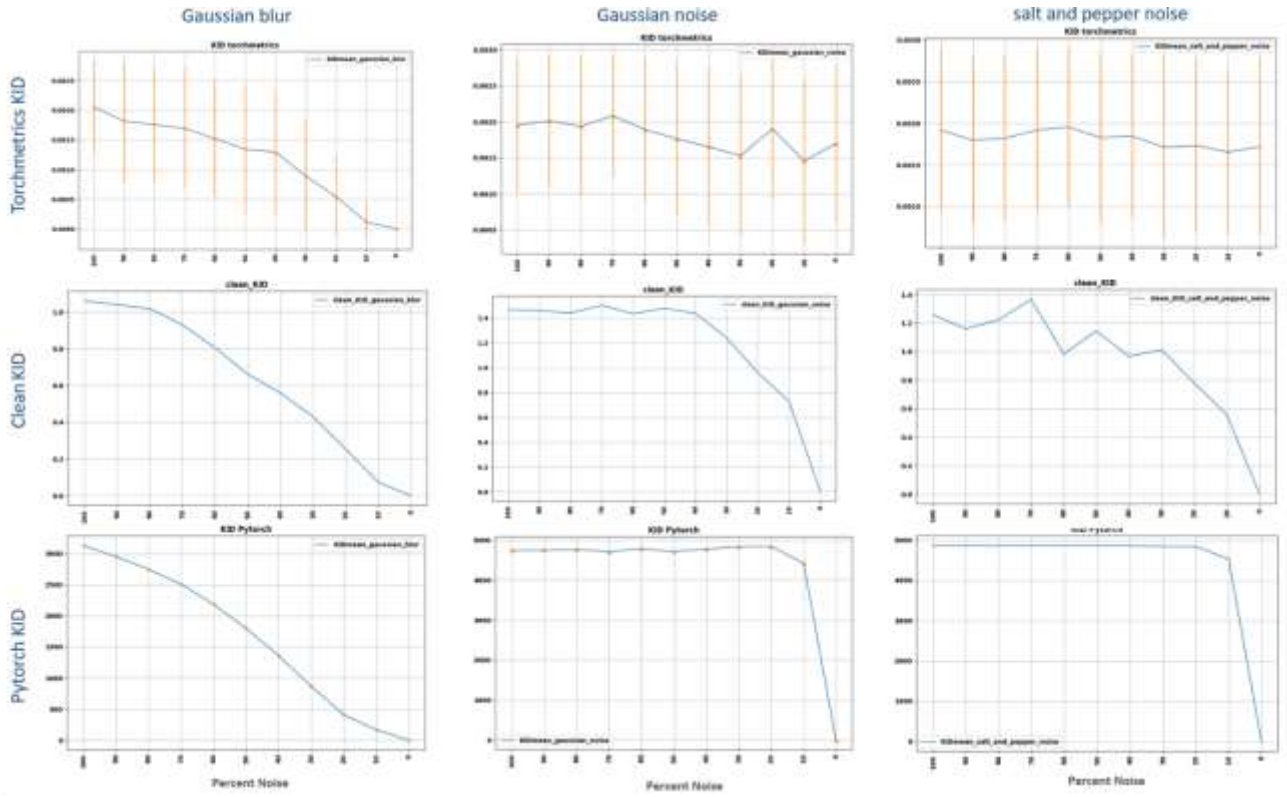

Figure 5 Torchmetrics KID, Clean KID and Pytorch KID implementation performance plots for noise variations

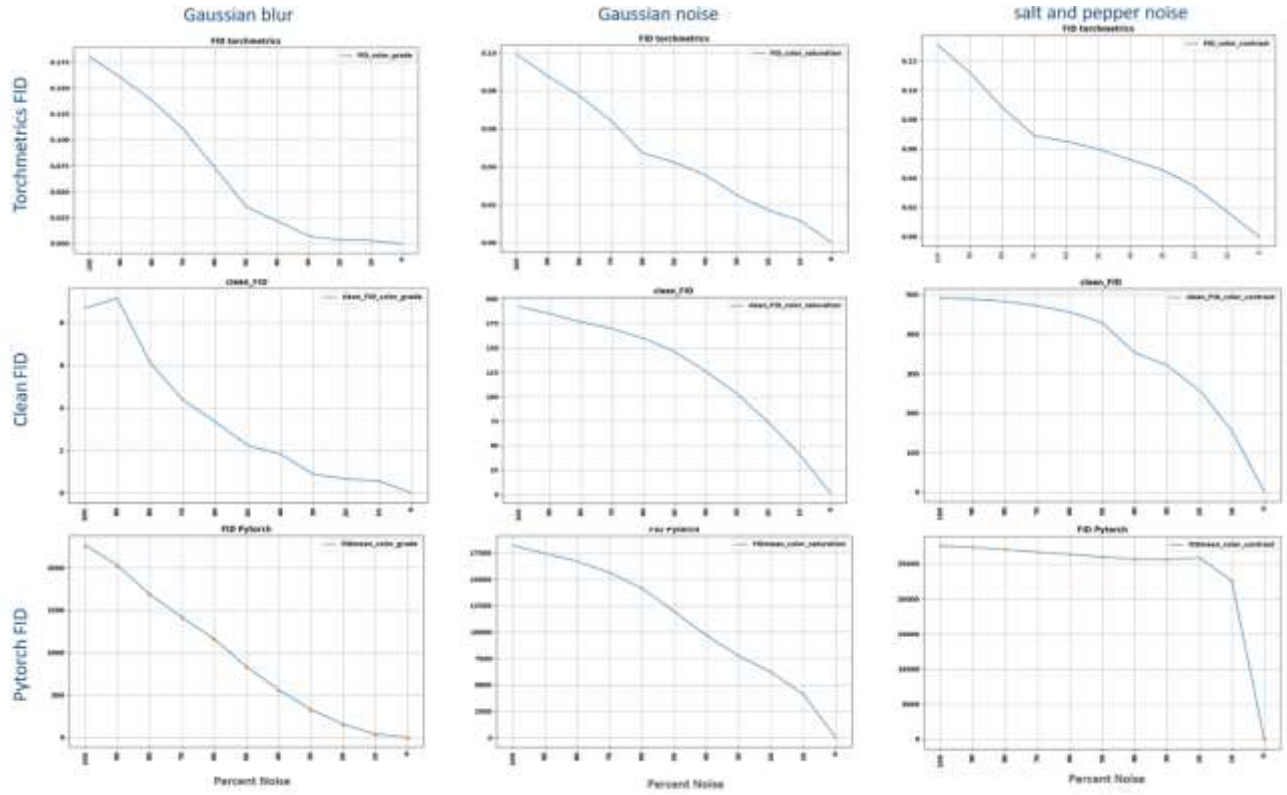

Figure 6 Torchmetrics FID, Clean FID and Pytorch FID implementation performance plots for color variations

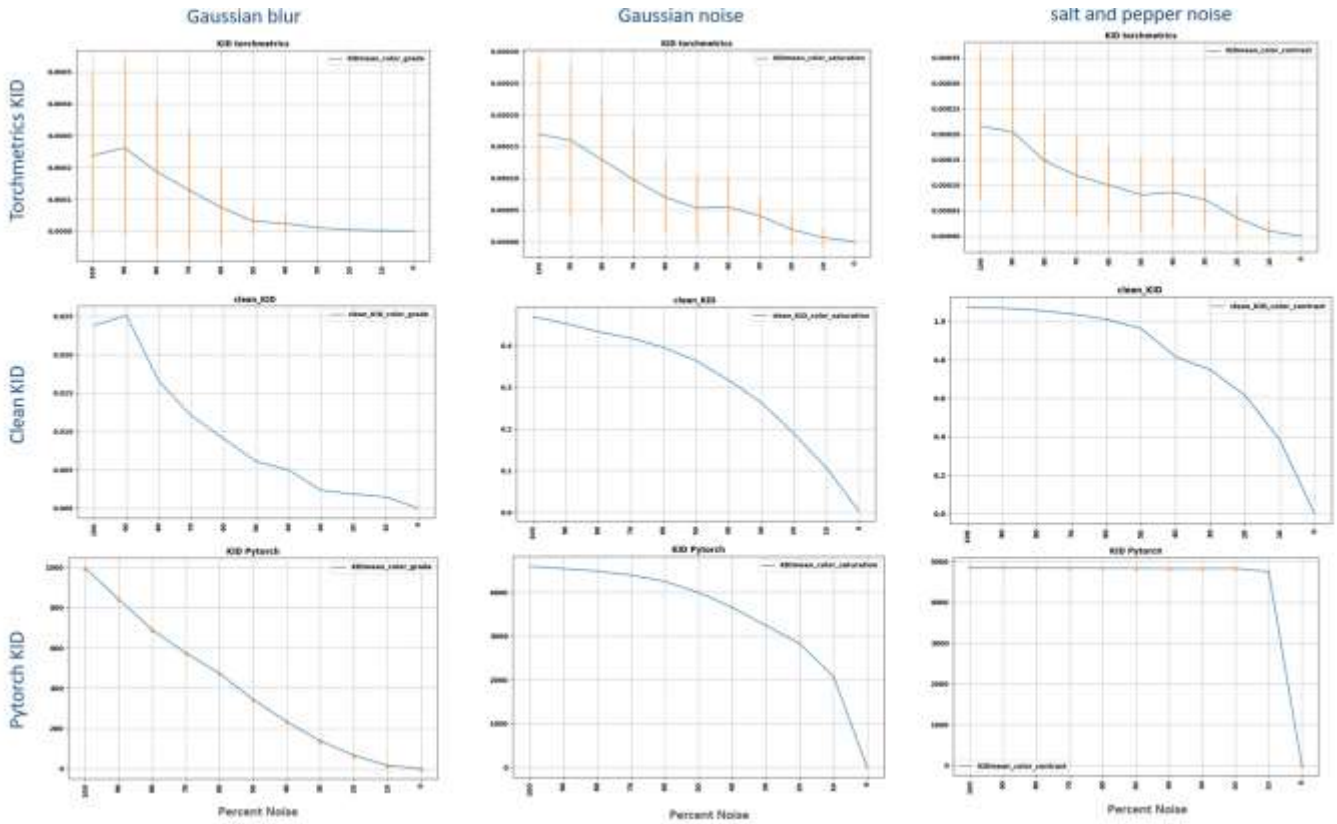

Figure 7 Torchmetrics KID, Clean KID and Pytorch KID implementation performance plots for color variations

Results and Analysis

FID: It is clear from the figure 4,5 below that the Torchmetrics FID outperforms the other implementations in capturing noise levels at each level for all types of noise and color variations introduced. The consistently increasing slope of the Torchmetrics scores confirms its sensitivity to noise and color variation in output. In contrast, the Pytoch FID implementation is poor and less sensitive to noise variations beyond 20% except for gaussian blur, indicating inconsistent behavior. The clean FID behavior is close to Torchmetrics FID and captures all noise and color variations with some deviation for higher levels of noise.

KID: Regarding the KID, while it achieves unbiased estimates, it suffers from high variance and is sensitive to sample size. Varying the sample size may lead to unreliable KID scores. The Torchmetrics KID also has high variance, resulting in an unreliable score. However, the clean KID performs better than the other two implementations.

Based on these findings, we hypothesize that the Torchmetrics FID is robust enough to capture various noise and color variations in images, making it suitable for applications such as Sim2Real and minimizing the gap between simulation and reality. In the validation section, we will experiment with sim2Real and use the same metrics to assess model performance.

Validation

To validate our hypothesis, we established an experimental setup for Sim2Real using Enhancing Photorealism Enhancement (EPE) GAN. This GAN implementation uses intermediate representations produced by conventional rendering pipelines called G-Buffers to enhance the images. Our objective was to employ the EPE GAN module to minimize the domain gap and utilize the FID and KID implementations to assess the model's performance. We trained our model using approximately five thousand synthetic and real images respectively. The model was trained for a total of 180k iterations, with a validation interval of 3k iterations. At each validation interval, we calculated the FID scores for all implementations and plotted the results. The figure 4 depicts that Torchmetrics FID effectively captures the progress of simulating to real gap minimization as compared to the other two implementations. Furthermore, we validated the domain-adapted synthetic images generated by the EPE GAN [21] model by predicting them through an image segmentation model that was trained on real images only. The segmentation model's performance was improved on the converted synthetic images, confirming the domain gap minimization, and the model was able to align the source synthetic domain to the target real domain.

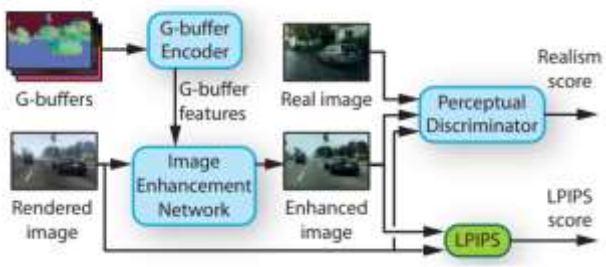

Figure 8 EPE GAN block diagram: Image Enhancement network uses G-buffers to enhance the rendered image.

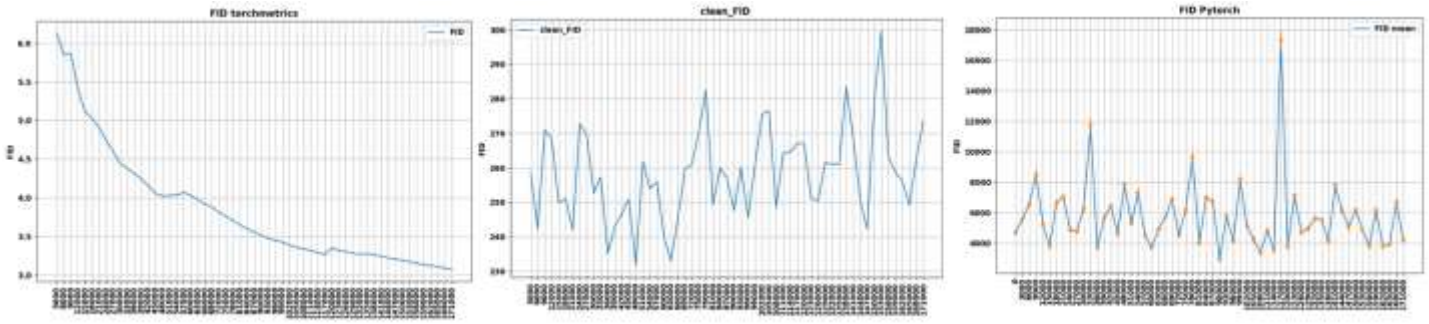

Figure 9 Validating EPE GAN using Torchmetrics, clean and Pytorch FID: Torchmetrics FID plot confirms the hypothesis made above the smooth downward slope captures the domain gap minimization effectively as compared to other two implementations.

Summary

In assessing the efficacy of various metrics in determining dissimilarities between visual representations, it has been observed that Kernel Inception Distance (KID) provides unbiased evaluations; however, it is characterized by a considerable degree of variability. The measurements obtained from Fréchet Inception Distance (FID) and KID exhibit inconsistent outcomes depending on the quality and dimensions of the images under consideration. Among the various implementations of Clean FID, one has been noted for its capacity to seamlessly accommodate disparate outcomes without experiencing substantial alterations.

When it comes to accurately quantifying domain shift phenomena, Torchmetrics FID shows itself to be a strong performer when capturing domain shift in both simulation and real-world scenarios. An interesting avenue for future research is to apply FID and KID metrics to closely matching features instead of doing comprehensive image comparisons. This refined strategy may help clarify minute details in features and textures in transformed images, which would improve the fineness of domain gap analysis and increase the accuracy of evaluation techniques.

References

1. S. J. Pan, I. W. Tsang, J. T. Kwok and Q. Yang, "Domain Adaptation via Transfer Component Analysis," in IEEE Transactions on Neural Networks, vol. 22, no. 2, pp. 199-210, Feb. 2011, doi: 10.1109/TNN.2010.2091281.
2. M. B. Bejiga and F. Melgani, "Gan-Based Domain Adaptation for Object Classification," IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 2018, pp. 1264-1267, doi: 10.1109/IGARSS.2018.8518649.
3. H. Guan and M. Liu, "Domain Adaptation for Medical Image Analysis: A Survey," in IEEE Transactions on Biomedical Engineering, vol. 69, no. 3, pp. 1173-1185, March 2022, doi: 10.1109/TBME.2021.3117407.
4. P. Singhal, R. Walambe, S. Ramanna and K. Kotecha, "Domain Adaptation: Challenges, Methods, Datasets, and Applications," in IEEE Access, vol. 11, pp. 6973-7020, 2023, doi: 10.1109/ACCESS.2023.3237025.
5. A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang and R. Webb, "Learning from Simulated and Unsupervised Images through Adversarial Training," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 2242-2251, doi: 10.1109/CVPR.2017.241.
6. K. Bousmalis, N. Silberman, D. Dohan, D. Erhan and D. Krishnan, "Unsupervised Pixel-Level Domain Adaptation with Generative Adversarial Networks," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 95-104, doi: 10.1109/CVPR.2017.18.
7. M. Schwonberg et al., "Survey on Unsupervised Domain Adaptation for Semantic Segmentation for Visual Perception in Automated Driving," in IEEE Access, vol. 11, pp. 54296-54336, 2023, doi: 10.1109/ACCESS.2023.3277785.
8. Wilson, Garrett & Cook, Diane. (2020). “A Survey of Unsupervised Deep Domain Adaptation.” In ACM Transactions on Intelligent Systems and Technology. doi:11. 1-46. 10.1145/3400066.
9. Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative adversarial nets." Advances in neural information processing systems 27 (2014).
10. Mirza, Mehdi, and Simon Osindero. "Conditional generative adversarial nets." arXiv preprint arXiv:1411.1784 (2014).
11. Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434. 2015 Nov 19.
12. B. T. Imbusch, M. Schwarz and S. Behnke, "Synthetic-to-Real Domain Adaptation using Contrastive Unpaired Translation," 2022 IEEE 18th International Conference on Automation Science and Engineering (CASE), Mexico City, Mexico, 2022, pp. 595-602, doi: 10.1109/CASE49997.2022.9926640.
13. Park, Taesung, et al. "Contrastive learning for unpaired image-to-image translation." Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16. Springer International Publishing, 2020.
14. Zhang, Charles Y., and Ashish Shrivastava. "AptSim2Real: Approximately-Paired Sim-to-Real Image Translation." arXiv preprint arXiv:2303.12704 (2023).
15. J. -Y. Zhu, T. Park, P. Isola and A. A. Efros, "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 2242-2251, doi: 10.1109/ICCV.2017.244.

16. Heusel, Martin, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. "Gans trained by a two time-scale update rule converge to a local nash equilibrium." Advances in neural information processing systems 30 (2017).
17. Bińkowski, M., Sutherland, D.J., Arbel, M. and Gretton, A., 2018. Demystifying mmd gans. arXiv preprint arXiv:1801.01401.
18. "Frechet Inception Distance (FID)¶." Frechet Inception Distance (FID) - PyTorch-Metrics 1.2.1 documentation. Accessed December 16, 2023. https://lightning.ai/docs/torchmetrics/stable/image/frechet_inception_distance.html.
19. G. Parmar, R. Zhang and J. -Y. Zhu, "On Aliased Resizing and Surprising Subtleties in GAN Evaluation," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022, pp. 11400-11410, doi: 10.1109/CVPR52688.2022.01112.
20. Seitzer, Maximilian. 'Pytorch-Fid: FID Score for PyTorch', August 2020. https://github.com/mseitzer/pytorch-fid.
21. S. R. Richter, H. A. Alhaija and V. Koltun, "Enhancing Photorealism Enhancement," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 2, pp. 1700-1715, 1 Feb. 2023, doi: 10.1109/TPAMI.2022.3166687.
22. Construction Site Safety Object Detection Dataset (v27, YOLOv8s) by Roboflow Universe Projects, https://universe.roboflow.com/roboflow-universe-projects/construction-site-safety/dataset/27, 2023.

Contact Information

Author: Sandip Omprakash Patil
Co-author: Girish Kumar Bellenavar
Email: sandip.patil824@gmail.com

Definitions/Abbreviations

AI	Artificial Intelligence
ML	Machine Learning
GANs	Generative Adversarial Networks.
FID	Fréchet Inception Distance
KID	Kernal Inception Distance
EPE	Enhancing Photorealism Enhancement