Evaluating Performance Boost in Base Masked Autoencoders by Enhancing Data Augmentation with Generative Models

Richi Dubey(richidubey@gatech.edu), Sidney Wise(swise30@gatech.edu), Emmanuel Ebhohimen(eebhohimen3@gatech.edu),
Hyun Soo Kim(hkim3100@gatech.edu)

ABSTRACT

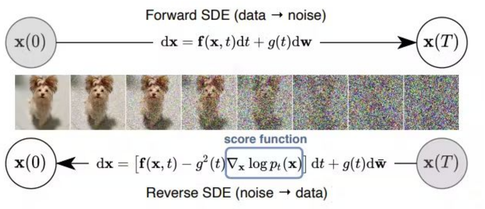
This project investigates the potential of data augmentation using generative models to improve the performance of Masked Autoencoders (MAEs) in unsupervised learning tasks. MAEs reconstruct masked portions of images, learning efficient representations without labeled data. The study compares four configurations: a baseline MAE trained on 200 RGB images from the KAIST Multispectral Pedestrian Dataset, an enhanced MAE trained on the same dataset augmented with 200 synthetic images generated by a Denoising Diffusion Probabilistic Model (DDPM) and a baseline MAE trained on 200 thermal images from the KAIST Multispectral Pedestrian Dataset, an enhanced MAE trained on the same dataset augmented with 200 synthetic images generated by CycleGAN. Results show 45% increased loss when data augmentation is used with diffusion models but a 9% decrease in training loss when CycleGAN's are employed.

INTRODUCTION

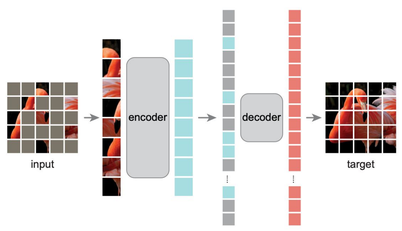
Masked Autoencoders (MAEs) are a powerful tool for unsupervised learning, particularly in scenarios where labeled data is limited. By reconstructing randomly masked patches of input images, MAEs can learn robust representations of image features. However, their performance is often constrained by the diversity and quantity of available training data. This project explores whether introducing synthetic images generated by diffusion models can mitigate this limitation. Diffusion models, known for their ability to create high-quality and diverse synthetic data, were used to augment the training dataset. By comparing the performance of MAEs trained on original and augmented datasets, this study aims to establish a scalable methodology for improving model robustness and generalization.

BACKGROUND INFORMATION

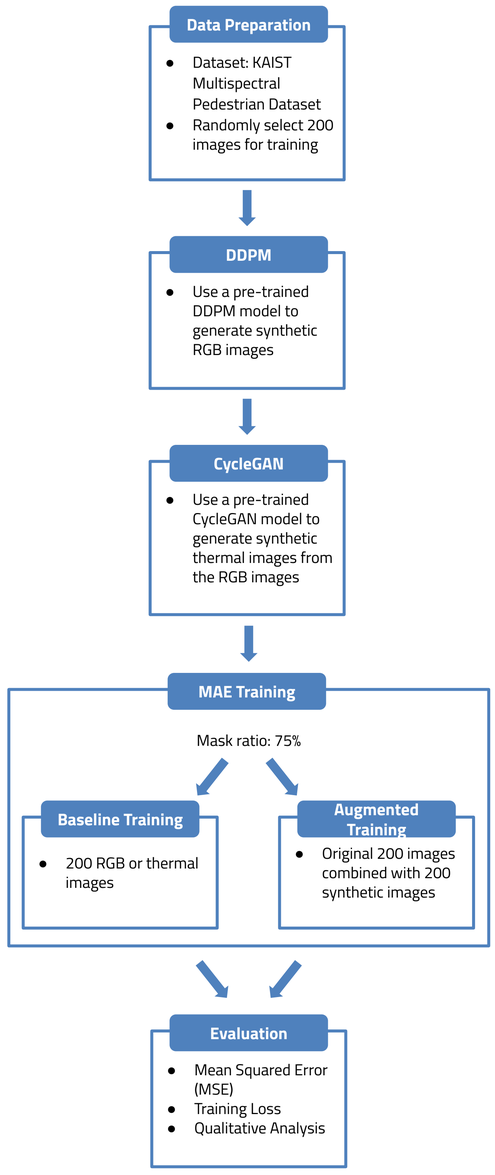
What Is Denoising Diffusion Probabilistic Model (DDPM)?



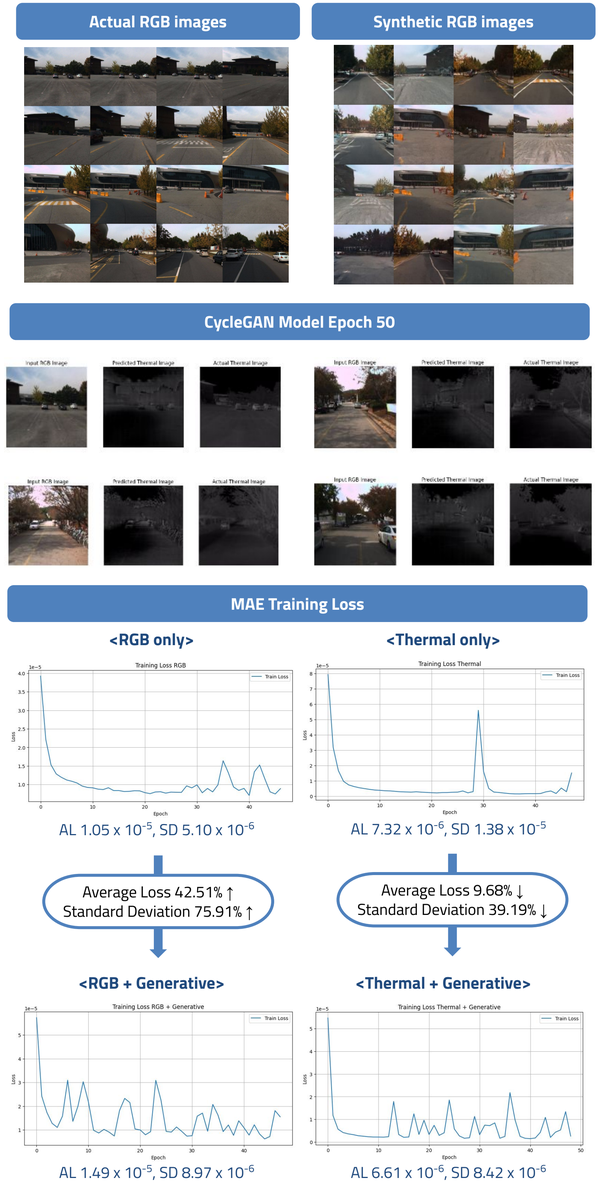
What Is Masked Autoencoders (MAE)?



METHODOLOGY



RESULTS



CONCLUSION

This study aimed to evaluate the impact of generative data augmentation on Masked Autoencoders (MAEs) by incorporating synthetic RGB and thermal images generated by DDPM and CycleGAN models.

The results indicate mixed outcomes: while the inclusion of synthetic images improved the performance of the thermal dataset, it increased the loss and variability in the RGB dataset. These results suggest that the impact of generative data augmentation depends on the dataset and input modality.

- RGB Dataset: The addition of generative images likely introduced excessive complexity or noise that hindered the model's ability to learn robust features effectively.
- Thermal Dataset: Generative images enhanced performance, potentially due to the simplicity of grayscale images, which introduced less noise during training.

This study demonstrates the potential of generative data augmentation to improve MAE performance in specific conditions, particularly in low-data regimes for thermal imaging. However, the challenges observed in the RGB dataset highlight the importance of careful dataset design and the quality of synthetic data.

FUTURE DIRECTIONS

- Conduct experiments with larger datasets and varied augmentation techniques to mitigate noise and improve generalization.
- Investigate the quality and diversity of synthetic images generated by diffusion and GAN models, as these factors may significantly influence performance.
- Explore the integration of domain-specific priors or hybrid augmentation approaches to improve results across diverse input modalities.

REFERENCES

[1] Hangbo Bao, Li Dong, Wenhui Piao, Haitao Xu, Xiaodong Song, and Jianfeng Gao. Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2206.04846, 2022. 2

[2] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021. 4

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020. 3, 4

[4] Hugging Face. facebook/vit-mae-base, n.d. 3, 4

[5] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Doll´ar, and Ross Girshick. Masked autoencoders are scalable vision learners. arXiv preprint arXiv:2111.06377, 2021. 3

[6] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In Advances in Neural Information Processing Systems, pages 6840–6851, 2020. 1, 2,

[7] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1505.04597, 2015. 2

[8] Yingying Peng. A comparative analysis between gan and diffusion models in image generation. Transactions on Computer Science and Intelligent Systems Research, 5:189–195, 08 2024. 2, 4

[9] Author(s) (replace this with actual authors if available). Masked autoencoders for scalable representation learning. arXiv preprint arXiv:2401.10561, 2024. 2

[10] Chen Wei, Karttikeya Mangalam, Po-Yao Huang, Yanghao Li, Haoqi Fan, Hu Xu, Huiyu Wang, Cihang Xie, Alan Yuille, and Christoph Feichtenhofer. Diffusion models as masked autoencoders. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 16284–16294. IEEE, 2023. 2