Evaluating Performance Boost in 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

We explore the impact of data augmentation using generative models, specifically diffusion models and Generative Adversarial Networks(GANs) on Masked Autoencoders (MAE) performance. MAEs are unsupervised models that reconstruct masked portions of input images, learning efficient representations without labels. We trained four MAE models: the first 2 on a dataset of 200 RGB and a separate 200 thermal images from the KAIST Multispectral Pedestrian Dataset, and the other two on an augmented dataset comprising the same 200 RGB images plus 200 diffusion-generated synthetic images and the same 200 thermal images plus 200 CyleGan-generated synthetic images. Based on Denoising Diffusion Probabilistic Models (DDPMs), the diffusion model was used to generate high-quality, diverse synthetic images that simulate various environmental and lighting conditions. Both MAE models utilized a mask ratio of 75%, adhering to standard implementation practices. Performance was evaluated using mean squared error (MSE) between reconstructed and original images, as well as the standard deviation of the loss. Results show that the diffusion-augmented dataset did not significantly improve reconstruction accuracy, with a 42.51% increase in average loss (from 1.05e-05 to 1.49e-05%). On the contrary, using images generated by a CycleGAN reduced the training loss of MAE by over 9%. These findings highlight the potential of using generative data augmentation in low-data regimes and suggest that including synthetic images may improve performance. A more robust study must be done in order to check the full potential of combining MAEs with generative models to enhance unsupervised learning. All the models, generated images and graphs can be accessed at https://github.com/richidubey/Improving-Multispectral-Object-Detection/tree/main.

1. Introduction

We aim to enhance the performance of Masked Autoencoders (MAE) by augmenting the training dataset with images generated by diffusion models and CycleGANs [11]. MAEs are trained to reconstruct missing patches of images, making them a suitable testbed for evaluating data diversity's role in unsupervised learning. The goal is to determine whether augmenting the dataset with generated data could improve MAE's generalization and reconstruction capabilities compared to using only original data. If successful, this approach could significantly improve the robustness of models in low-data regimes. Applications include medical imaging, autonomous driving, and other fields with scarce labelled data. The findings also highlight the potential of generative models like diffusion as powerful tools for data augmentation.

2. Related Work

2.1. Denoising Diffusion Probabilistic Models (DDPMs) [6]

This paper proposed a powerful probabilistic model for image generation and transformation. This model operates by gradually adding noise to an image and then reversing the process to recover the original image, enabling the generation of high-quality images. Specifically, DDPMs learn through the use of Gaussian noise and employ a reverse process to reconstruct the image, resulting in a stable generation process. In our project, we plan to use this probabilistic denoising model for creating new RGB images from our dataset. The strong image generation capabilities of DDPMs can be applied to this transformation process. DDPM's ability to generate high-quality images may be more suitable than traditional GAN-based methods. Thus, using DDPMs to convert images, followed by further analysis through thermal-based object detection models, is expected to play a significant role in the success of our approach.

2.2. A Comparative Analysis Between GAN and Diffusion Models in Image Generation

This paper is a literature review comparing GANs and diffusion models in image generation. While it acknowledges that GANs perform well in generating realistic images, it also notes that diffusion models offer better stability and controllability [8]. GANs are an easier model to turn out results are unstable. The author suggest that future research should focus on combining the advantages of both approaches.

2.3. Masked Autoencoders are Robust Data Augmentors [1]

The paper investigates using masked autoencoders for data augmentation to enhance the robustness and performance of deep neural networks in image recognition tasks. They mask parts of input images and use a lightweight masked autoencoder to reconstruct them. The reconstructed images are used as augmentations for training. The method improves accuracy across supervised, semi-supervised, few-shot, fine-grained, and long-tail classification tasks. However, our work involves using diffusion models to improve performance of MAEs on downstream tasks, rather than using MAEs for data augmentation.

2.4. Diffusion Models as Masked Autoencoder [10]

The authors introduce DiffMAE, a novel approach integrating masked autoencoding into diffusion models. This method involves conditioning diffusion models on masked inputs to approximate the pixel distribution of masked regions based on visible areas. DiffMAE demonstrates strong performance in initialization for Recognition Tasks, Image inpainting and video classification. This is interesting way to improve performance of diffusion models, but out work focuses on improving the baseline performance of MAEs.

2.5. MAEDiff: Masked Autoencoder-enhanced Diffusion Models for Unsupervised Anomaly Detection in Brain Images [9]

The authors introduce a novel approach, MAEDiff, which combines masked autoencoders and diffusion models to address challenges in unsupervised anomaly detection in brain MRI images. They embed an MAE mechanism into the diffusion U-Net architecture to improve conditioning on visible regions for more accurate reconstructions of the noised patches. Their work demonstrates superior anomaly detection and image reconstruction performance across three public datasets: IXI, BraTS21, and MSLUB. This is an interesting use of integrating MAEs in diffusion models to improve downstream performance of diffusion models and is different from our work that aims to improve the baseline MAEs performance.

3. Technical Approach

- **Strategy and Approach:** The objective of this project is to evaluate the effectiveness of diffusion-based and GAN data augmentation in improving the reconstruction capabilities of Masked Autoencoders (MAEs). The strategy involves training MAEs under two different configurations:

  1. **Baseline Training:** There are two baseline MAEs. One is trained on a dataset of 200 RGB images the other on 200 thermal images with both sets extracted from the KAIST Multispectral Pedestrian Dataset. This serves as the baseline for comparison.
  2. **Augmented Training:** The two augmented MAEs are trained on a dataset the same 200 dataset images and 200 generated synthetic images. The diffusion model augments the dataset by introducing realistic and diverse variations of the original RGB images, simulating different environmental and lighting conditions. The CycleGan uses similar augments to convert RGB images to thermal.

  Both configurations use the same model architecture, optimizer, and evaluation metrics to ensure fair comparisons. The training pipeline involves masking random patches of the input images and tasking the MAE to reconstruct them, using Mean Squared Error (MSE) as the loss function to evaluate reconstruction accuracy.

- **Diffusion Model** [7] Diffusion models operate by progressively introducing noise into an image over multiple steps and then learning to predict and remove this noise to regenerate the original image. This process involves a forward diffusion step, where Gaussian noise with zero mean and unit covariance is added iteratively [6], and a reverse denoising step, where the model is trained to reconstruct the image by approximating the reverse of the noise addition process. The stochastic nature of the noise introduction enables the generation of novel, high-quality images by sampling from the learned latent distribution and reversing the noise process.

  The forward process is shown in Figure 1 and the reverse process is shown in Figure 2.

  The standard architecture used for diffusion models is a U-Net, incorporating downsampling and upsampling blocks with residual connections, which efficiently learn the latent representation of an image for complete reconstruction. A sample UNet architecture is shown in figure 4.

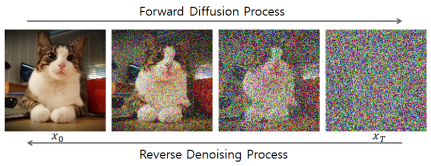

Figure 1. Forward process in diffusion

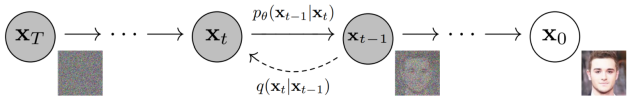

Figure 2. Reverse (denoising) step in diffusion

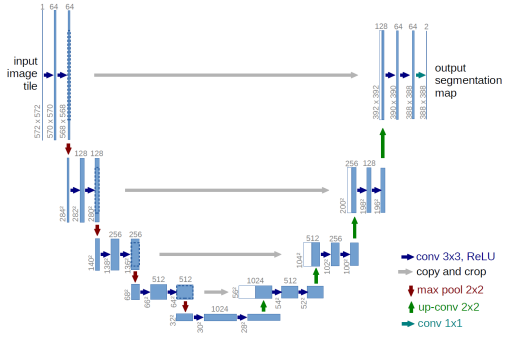

Figure 3. Sample UNet architecture

- **Mask Auto Encoder:** [4] [5]

  An image is worth 16x16 words [3]: Transformers for image recognition at scale. In ICLR, 2021.

  We use the MAE architecture introduced by He et al. [5] which employs a Vision Transformer (ViT) base [3] as the encoder. The encoder processes only the visible (non-masked) patches of the input image, while the masked patches are ignored at this stage. Subsequently, mask tokens are introduced, and the decoder takes the encoded visible tokens along with the mask tokens to reconstruct the original image. The architecture is illustrated in Figure 4

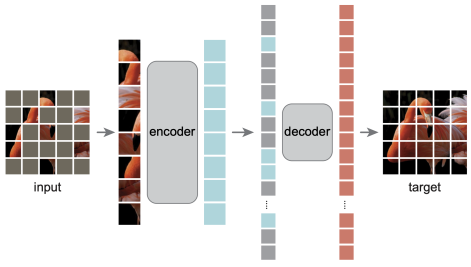

Figure 4. MAE architecture

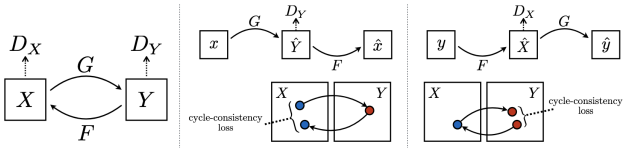

Figure 5. Style transfer with CycleGAN

- **CycleGAN:**

  CycleGAN [11] is a Generative Adversarial Network (GAN) designed for style transfer. It employs two generators (G_1 and G_2) and two discriminators (D_X and D_Y) to perform bidirectional style transfer between two domains, X and Y. The generator G_1 learns to create images in domain Y that resemble images from domain X, while the discriminator D_Y attempts to differentiate between real images from domain Y and the synthesized images produced by G_1. Similarly, G_2 and D_X handle the reverse transformation from domain Y to domain X. A cycle consistency loss is used to prevent mode collapse by ensuring that translating an image from one domain to the other and back again (e.g., $x \rightarrow G(x) \rightarrow F(G(x))$) should reproduce the original image ($F(G(x)) \approx x$).

- **Dataset:** The KAIST Multispectral Pedestrian Dataset provides aligned RGB and thermal image pairs captured under diverse traffic and lighting conditions. To augment the dataset, 200 additional synthetic images are generated using a pre-trained diffusion model. These images are designed to enrich the training data with diverse variations, enhancing the robustness of the MAE.

- **Reasoning for Method Effectiveness:** The proposed method leverages the complementary strengths of MAEs and diffusion models:

  - **Masked Autoencoders (MAEs):** MAEs are highly effective for unsupervised representation learning, as they focus on reconstructing masked patches, enabling the model to learn global and local features without requiring labeled data.
  - **Diffusion and CycleGAN Models:** These models provide high-quality, realistic synthetic images that introduce meaningful diversity into the dataset. By simulating varied traffic and lighting conditions, these models help overcome the limitations of small datasets.
  - **Combined Approach:** Training the MAE with a combination of original and augmented images enriches the dataset, reducing overfitting and improving generalization, particularly in low-data regimes.

This approach ensures that the MAE can effectively reconstruct masked patches in diverse conditions, demonstrating the utility of generative augmentation in unsupervised learning.

- **Metrics for Evaluation:** The performance of the MAE is evaluated using the following metrics:
  - **Mean Squared Error (MSE):** The primary evaluation metric, MSE measures the pixel-wise difference between the reconstructed and original images. Lower MSE indicates better reconstruction performance.
  - **Training Loss:** The training loss is tracked over epochs to monitor convergence and detect overfitting. Loss curves provide insights into the stability and effectiveness of the training process.
  - **Qualitative Analysis:** Reconstructed images are visually inspected to assess the quality of reconstructions and the impact of diffusion-augmented data.

4. Results

4.1. Generating Synthetic RGB Images

4.1.1 Denoising Diffusion Probabilistic Models (DDPMs)

To augment our dataset and improve the robustness of our pedestrian detection model, we generated synthetic RGB images using Denoising Diffusion Probabilistic Models (DDPMs). DDPMs have demonstrated superior performance in image synthesis tasks compared to traditional Generative Adversarial Networks (GANs) [2]. They are a class of generative models that learn to generate data by reversing a gradual noising process [6]. Starting from pure Gaussian noise, the model iteratively denoises the image over several timesteps, producing high-fidelity images that resemble the training data distribution.

We used a UNet architecture with 5 layers of downsampling and equal layers of upsampling with a spatial self-attention block. We trained our model for 50 epochs on a dataset of 1k images (out of 23k).

After training, we generated synthetic RGB images by sampling from the noise prior and iteratively denoising using the trained DDPM model. The sample images in our dataset are shown in and the generated images are shown in Figure 7.

4.2. Generating Synthetic RGB To Thermal Images

4.2.1 Vision Transformer (base-sized model)

Originally, a diffusion model using a pre-trained model (google/vit-base-patch16-224) as an encoder was the starting method for synthesizing RGB to thermal images. Using a pre-trained model was suppose to address time constraints in model training. ViT was chosen for its ability to capture spatial and structural information, and its potential for fine-tuning on image-to-image translation tasks with limited paired data [3].

The model employs an encoder-decoder architecture, using the pre-trained ViT as the encoder and a custom decoder. It uses Adam optimizer with a learning rate of 1e-4 and Mean Squared Error (MSE) as the loss function and trained for 50 epochs. However despite steadily increased epochs and constantly changing the parameters in training as well

Figure 6. Original images in dataset

Figure 7. Synthetic images generated by using a diffusion model

as the encoders, the model only produced the same fuzzy images as shown in Figure 8. With the deadline approaching and training taking longer with the 50 epochs a separate CycleGan model was created which eventual produced usable images further discussed in section 4.2.2. Since the CycleGan model was ready and produced images quickly the diffusion model for thermal was drop to test the MAE on results made by the CycleGan.

Figure 8. ViT RGB to Thermal Model Epoch 10

4.2.2 CycleGAN

As discussed subsection 2.2 GANs excel at model generation and produced usable results faster. For the sake of time the thermal diffusion model was drop and we continued with the GAN.

The CycleGAN model comprises two U-Net generators and two convolutional discriminators, using Adam optimizers with a learning rate of $2e^{-4}$ and $beta_1$ of 0.5. It employs binary cross-entropy loss and cycle consistency loss. Trained for 50 epochs on the KAIST dataset, the model generates fake images and updates components based on adversarial and cycle consistency losses. As shown in Figure 9, the model produces fairly realistic images but with missing consistency expected of GANs models [8].

Figure 9. CycleGAN Model Epoch 50

4.3. Mask Auto Encoder (MAE)

Four MAEs were created for this project using HuggingFace's pretrained model called "facebook/vit-mae-base" [4]. The MAE's were split evenly into two groups. One group only had thermal images as an input the other half only had RGB images. The first MAE of each group is our experiments control which only contains 200 images from the KAIST Multispectral Pedestrian Dataset. The second group contained the same 200 images from the KAIST Multispectral Pedestrian Dataset as the first along with an additional 200 generated images. The second MAE RGB had images generated from the diffusion model discussed in the subsection 4.1.1. The second MAE Thermal had images generated from the CycleGan model discussed in section 4.2.2. All models were evaluated using the same preset parameters and test sets to ensure a fair comparison. The results reveal significant differences in performance between the two approaches.

The following subsections will evaluate the models performances via their reconstruction loss. The reconstruction loss is a critical metric for evaluating MAE performance, with lower values indicating better reconstruction capabilities.

4.3.1 MAE RGB Evaluation

Both MAE RGB models utilized a mask ratio of 75.00%, adhering to the standard MAE implementation. However, the similarities between the models stop there.

Figure 10 and Figure 11 show the training loss for the model using the only the database (MAE RGB Database) and the model using both the database and generated images (MAE RGB Database + Generative). The MAE RGB Database had an average loss of $1.0470873486511811e^{-05}$ with a standard deviation of $5.100347822663716e^{-06}$. The MAE RGB Database + Generative meanwhile had an average loss of $1.49218631734654e^{-05}$ with a standard deviation of $8.972070497143067e^{-06}$.

Between the two models the MAE RGB Database + Generative average loss increased by 42.51% compared to the MAE RGB Database alongside a 75.91 % with the standard deviation. Contrary to our expectations, the addition of generative images to the RGB dataset resulted in higher average loss and increased variability demonstrating a lower performance to its counterpart. This coincides with the graphs. As you can see Figure 11 has more random loss spikes compared to Figure 10. These findings may suggests that the generative images may have introduced noise or complexity that the model struggled to handle effectively. The inclusion of generative images in the training process may not led to more robust and generalizable feature learning, we initially hypothesized. However this is only one small experiment. The two models had a small dataset with the highest being 400, trained on 50 epochs with no varied testing done to "facebook/vit-mae-base" parameters due to time constraints. A more rigorous study needs to be done in order to thoroughly disprove our hypothesis.

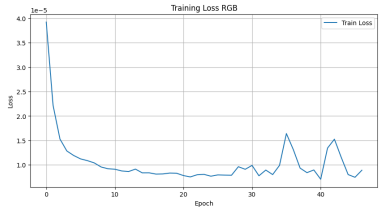

Figure 10. MAE Loss RGB Database

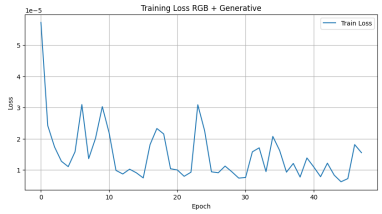

Figure 11. MAE Loss RGB Database + Generative

4.3.2 MAE Thermal Evaluation

Similar to the MAE RGB, both MAE Thermal models utilized a mask ratio of 75.00%, also adhering to the standard MAE implementation. Continuing the similarities with MAE RGB the models begin to differ when evaluating the loss.

Figure 12 and Figure 13 show the training loss for the model using the only the database (MAE Thermal Database) and the model using both the database and generated images (MAE Thermal Database + Generative). The MAE Thermal Database had an average loss of $7.317142422544832e^{-06}$ with a standard deviation of $1.3844919993640968e^{-05}$. The MAE Thermal Database + Generative meanwhile had an average loss of $6.609162300524726e^{-06}$ with a standard deviation of $8.418641833272377e^{-06}$.

Contrary to the RGB MAE's evaluation, the MAE Thermal Database + Generative average loss decreased by 9.68% compared to the MAE RGB Database alongside a decrease of 39.19% with the standard deviation. This decrease can be seen in the fiugres with Figure 13 having one big loss spike compared to Figure 12 multiple. Contrary to expectations set by the previous section, the additional generative images in the Thermal dataset resulted in lower average loss and variability showcasing a better performance though by a low percentage. This result is also unexpected when taking into account the generative thermal images used a CycleGan model which is suppose to be more varied and unstable compared to diffusion [8]. However, this decrease may be caused by the thermal images being gray-scale. With two less dimension there's less noise to process lower the loss. However, as stated in the section prior section, this is a small scale experiment and a bigger study needs to be done in order to be more conclusive.

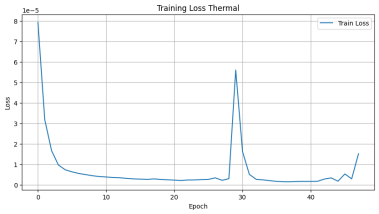

Figure 12. MAE Loss Thermal

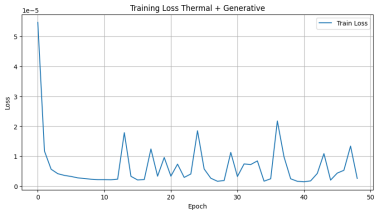

Figure 13. MAE Loss Thermal + Generative

5. Conclusion

In conclusion, our study demonstrates the potential benefits of augmenting Masked Autoencoder (MAE) training with images generated by generative deep models. The MAE trained on the combined dataset of original and synthetic images showed an increase in the average reconstruction loss in the case of augmenting with diffusion models' output but a reduction of 9% in the case of augmenting with CycleGAN's output. These results show the potential benefits of generative models in increasing dataset size in applications that are often plagued with low-regime data, like medical imaging. However, this is a small study of generative models, and future work could explore the scalability of this method with larger datasets, investigate its effectiveness in other domains, and examine how the quality and diversity of diffusion-generated and CycleGAN generated images impact MAE performance. Additionally, combining this approach with other data augmentation techniques or applying it to different architectures could yield further insights into improving unsupervised learning models.

6. Contribution Table

We all contributed equally to the project.

Group Member	Contribution
Emmanuel	Modified, fine-tuned, and trained the vit-mae model for thermal, rgb, rgb and generated rgb, and thermal and generated thermal images datasets. Produced graphs and loss data for each model to be used for evaluation and the results section. Contributed to References and Citations in the report, and provided guidance for the direction of the project.
Richi	Created the diffusion model, ran experiments, guided the group on directions (Project objective, use of MAE, loss functions and graphs), adding said analysis to results/conclusion/abstract section. Contributed to Abstract, Introduction, Related works, Technical Approach and Results sections.
Hyun Soo	Created, fine-tuned, and trained the vit-mae model for 200 RGB only datasets. Contributed to Abstract, Introduction, Technical Approach sections of the report. Took charge of designing and printing the poster.
Sidney	creating the RGB to Thermal model, analyzing data (calculating Mask ratio, comparing loss of MAE models), adding said analysis to results/conclusion/abstract section, writing sections: 4.2, 4.3.1, 4.3.1 2.2

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, 5

[5] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, 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, 4

[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, 5, 6

[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

[11] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks, 2020. 1, 3