Final Project: Visualizing and Understanding Diffusion Models

Mingjian Lu mxl1171@case.edu
Sixu Li sxl2199@case.edu

Abstract

This project investigates the working principles of diffusion models through multiple visualization approaches. We visualize the intermediate results of each diffusion step, compare feature maps of U-Net architectures in both diffusion models and image segmentation tasks, and conduct ablation studies on key architectural components. Our findings provide insights into how diffusion models progressively denoise images, how U-Net extracts features differently across tasks, and which architectural elements contribute most significantly to model performance. These results enhance our understanding of diffusion models' inner workings and provide practical guidance for optimizing model design.

1 Introduction and Background

Diffusion models have emerged as powerful generative models, achieving state-of-the-art performance in image synthesis tasks. Despite their impressive capabilities in generating high-quality outputs, these models are often regarded as "black boxes," making it difficult to understand their internal representations and generation processes. Our project explores the interpretability of diffusion models by analyzing their denoising dynamics, feature representations, and architectural components.

Our work is based on Improved Denoising Diffusion Probabilistic Models (IDDPM), which enhanced the original DDPM by introducing residual blocks and self-attention mechanisms [1]. Understanding these models is essential for domains requiring trust and transparency, and can help researchers improve model performance when unsatisfactory outputs are generated. Through visualization and ablation studies, we aim to find how different components contribute to the model's effectiveness.

2 Methods and Results

2.1 Visualizing the Sampling Steps

When generating an image, diffusion models progressively remove noise from an initially random input [2]. To visualize this process, we extracted intermediate results at various denoising steps using a pre-trained stable diffusion model. Figure 1 shows the progression from noise to a coherent image.

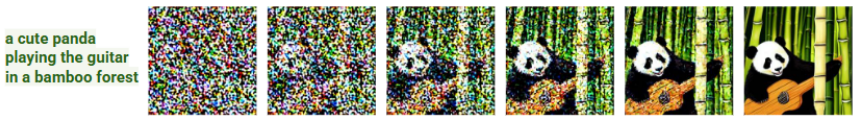

Figure 1: Intermediate images generated by stable diffusion model, showing the progressive denoising process from pure noise (left) to final image (right).

This visualization reveals distinct phases in the generation process: early steps establish global structure and rough object positions; middle steps develop basic forms and layouts; and final steps

refine textures, edges, and fine details. Only at the final step does the image become fully sharp and clean.

2.2 Comparison of Feature Maps Across Tasks

The U-Net serves as the backbone network for diffusion models but was originally designed for biomedical image segmentation tasks [3]. A large number of feature maps were generated through experimental procedures, from which a representative subset has been selected. The images below illustrate the intermediate representations captured across different layers of the U-Net. We compared feature maps from U-Nets in both contexts to understand how the same architecture learns different representations based on the task objective. By examining their similarities and differences, we aim to achieve a deeper understanding of the underlying mechanisms of U-Nets.

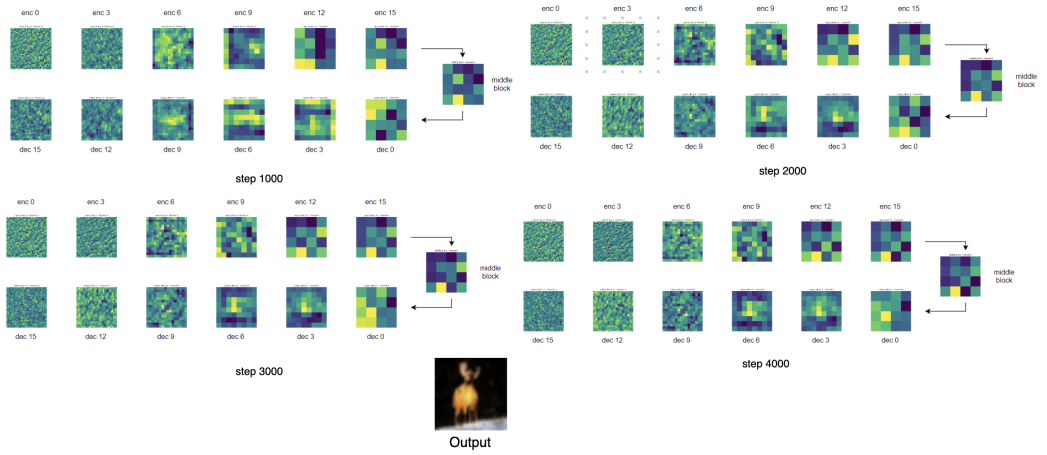

Figure 2: Feature maps from diffusion model U-Net (step 4000, 3000, 2000, 1000; channel 0), showing "mosaic-like" patterns that reflect the model's focus on predicting noise rather than semantic content.

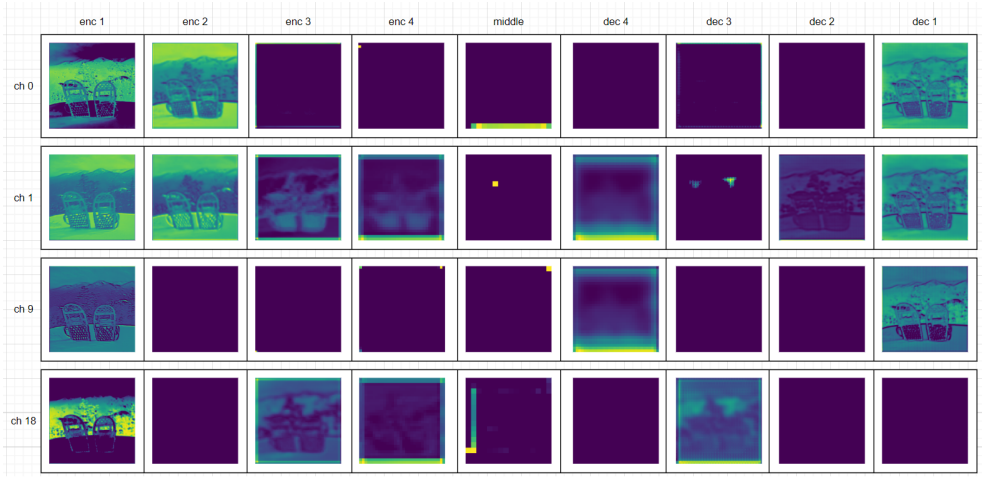

Figure 3: Feature maps from a U-Net trained for image segmentation across different channels and layers, showing clear boundary detection and semantic region identification that contrasts with diffusion model patterns.

By examining the feature maps across different layers of the model (Figures 2, 3), we can intuitively observe how U-Net operates. As the depth of the encoder increases, the extracted features become progressively more abstract. During the decoding process, skip connections help compensate for

spatial information lost during encoding, assisting the decoder to reconstruct the details of the target. As the decoder progresses through its layers, the spatial resolution of the feature maps gradually increases, recovering the original image size.

Our analysis revealed a fundamental difference between diffusion model U-Nets and segmentation U-Nets: while segmentation U-Nets learn to identify boundaries and semantic regions (Figure 3), diffusion model U-Nets exhibit "mosaic-like" patterns (Figure 2 because they are trained to predict noise distributions across the entire image rather than focusing on specific semantic features. This explains why diffusion feature maps don't show recognizable outlines of the generated content. The objective of diffusion model U-Net aims to generate the noise and the network tries to detect the noisy pixel in every corner. It is expected that the visualized feature maps display "mosaic images".

Key findings from our feature map comparison:

- Segmentation U-Nets learn edge information and discriminative features
- Diffusion U-Nets distribute attention across the entire image to model noise
- Skip connections play different roles in each context, but remain essential for information preservation
- The optimization objective fundamentally shapes what representations are learned

2.3 Ablation Studies

To systematically evaluate which architectural components contribute most to diffusion model performance, we conducted ablation studies by varying three key parameters:

1. **Width** (channels_32 vs. channels_64)
2. **Depth** (res_blocks_0 vs. res_blocks_1 vs. res_blocks_2)
3. **Attention** (with vs. without self-attention mechanisms)

We measured performance using both training loss (MSE) and Fréchet Inception Distance (FID), training each variant for 100,000 steps. Table 1 summarizes the results.

Table 1: Performance comparison of different model architectures after 100,000 training steps

Model Configuration	Loss	Loss Reduction (%)	FID	Relative Improvement
channels_32 (baseline)	0.00418	99.58	84.32	0%
channels_64	0.00086	99.91	62.75	+25.6%
no_attention	0.00103	99.90	118.65	-40.7%
res_blocks_0	0.00296	99.70	102.51	-21.6%
res_blocks_1	0.00359	99.64	93.78	-11.2%

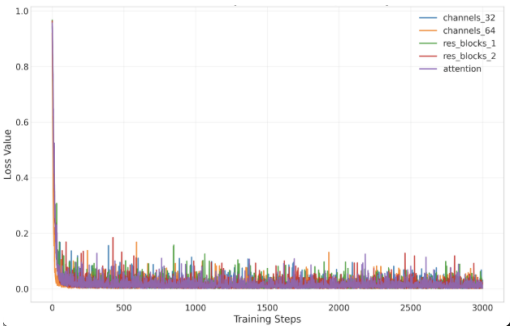

Figure 4: Loss curves comparison for first 3000 training steps.

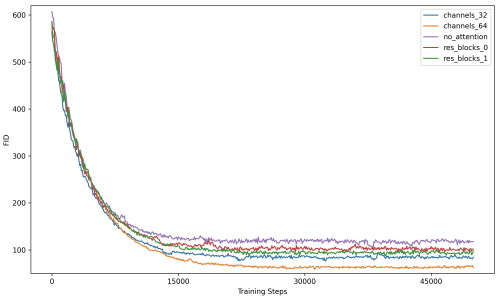

Figure 5: FID curves for first 50,000 training steps.

Our experiments revealed several critical insights about diffusion model architecture:

Impact of Channel Width: Doubling the number of channels from 32 to 64 produced the most significant improvement, reducing FID by 25.6%. The wider model achieved faster convergence and consistently maintained lower FID scores throughout training (Figure 5), indicating that feature capacity is crucial for effective noise estimation.

Importance of Attention Mechanisms: Removing attention layers caused severe performance degradation, increasing FID by 40.7%. Interestingly, the no-attention model still achieved low MSE loss values despite poor perceptual quality (Figure 4), highlighting a disconnect between pixel-wise error and global coherence. This confirms that attention's role in modeling long-range dependencies is essential for generating coherent images.

Effect of Residual Blocks: Reducing residual blocks from 2 to 0 increased FID by 21.6%. Models with fewer residual blocks initially learned faster but plateaued earlier, suggesting that deeper networks provide better asymptotic performance at the cost of slower initial convergence.

3 Conclusions and Future Work

Our visualizations and ablation studies provide valuable insights into the inner workings of diffusion models. We demonstrated how these models progressively denoise images, how the U-Net architecture learns different representations when applied to different tasks, and which architectural components contribute most significantly to model performance.

Our findings suggest several design principles for diffusion model architectures:

- U-Net's effectiveness stems from its multi-scale processing and skip connections, which align with the hierarchical nature of image formation
- When facing computational constraints, increasing model width should be prioritized over depth
- Attention mechanisms are essential components despite their small parameter footprint
- Residual blocks can be adjusted based on computational budget, with diminishing returns beyond two blocks

Future work could explore more fine-grained interpretability techniques, investigate the role of timestep embeddings in the diffusion process, and analyze how different architectural choices affect specific aspects of image quality such as coherence, diversity, and fidelity [4].

References

[1] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171, 2021.

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851, 2020.

[3] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241, 2015.

[4] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.