Diffusion and Score-based Generative Modeling

Course: Advanced Deep Learning

Name: Aman Kumar **Roll No:** MT24012

Q1[5] — Diffusion-based Training on M3FD using DDIM Inversion

Objective

To train a **U-Net model** that maps RGB latent representations to IR latents on the **M3FD dataset**. We use **DDIM inversion (t = 400)** derived from the pretrained diffusion model **"CompVis/stable-diffusion-v1-4"** to extract and store latent pairs for RGB–IR training.

Methodology

Dataset: M3FD RGB–IR paired dataset (center-cropped to 512×512, resized to 256×256).

Latent Generation: Using the **Stable Diffusion v1-4** model with a **DDIM scheduler**, each RGB and IR image was encoded through the VAE into latent z_0 and deterministically diffused to z_{400}:

$$z_{t+1} = \sqrt{\alpha_{t+1}}\, x_0 + \sqrt{1-\alpha_{t+1}}\, \epsilon.$$

Latent pairs $(z_{400}^{RGB}, z_{400}^{IR})$ were saved for reuse in training and evaluation.

Model: A U-Net with skip connections mapping $z_{400}^{RGB} \rightarrow z_{400}^{IR}$:

$$\hat{z}^{IR} = f_\theta(z_{400}^{RGB})$$

Loss: Mean Squared Error (MSE),

$$L = \|\hat{z}^{IR} - z_{400}^{IR}\|_2^2.$$

Trained for 100 epochs using AdamW ($lr = 1 \times 10^{-4}$).

Evaluation

Predicted IR latents were decoded using the same Stable Diffusion VAE and compared to ground truth IR using **PSNR** and **SSIM** metrics:

$$\text{PSNR} = 10 \log_{10}\left(\tfrac{1}{\text{MSE}}\right), \quad \text{SSIM} = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}.$$

Quantitative Results:

- PSNR: **11.47–11.54 dB**
- SSIM: **0.067–0.080**

Observations:

- The model learns approximate RGB→IR mapping in latent space.
- Detail loss remains visible; perceptual or adversarial terms may improve quality.
- Latent-space training reduces memory and computation significantly.

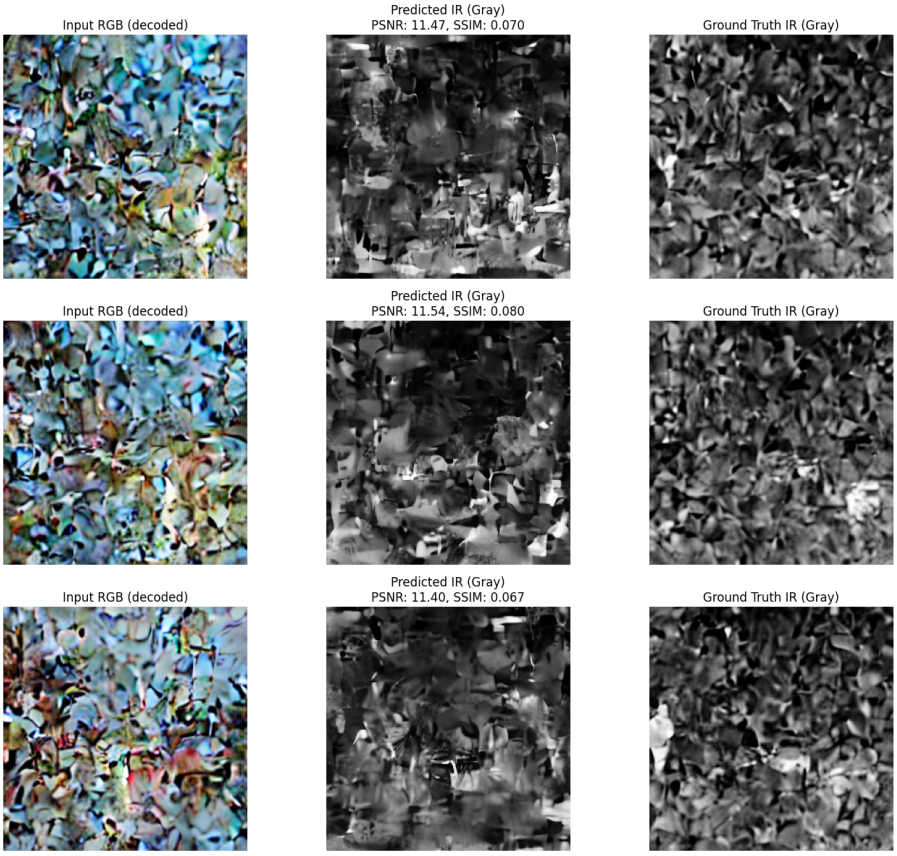

Figure 1: *
Q1 — RGB to IR prediction examples (Decoded). Left: Input RGB, Middle: Predicted IR, Right: Ground Truth IR.

Conclusion

A latent-space diffusion pipeline based on **Stable Diffusion v1-4** was implemented for RGB–IR translation via DDIM inversion (t=400). The trained U-Net achieved PSNR ≈ 11.5 and SSIM $\approx$ 0.067–0.078, confirming successful latent regression. Future extensions may employ perceptual or adversarial losses for sharper IR predictions. I also could able to run bonus question but my accuracy is les than 10% that's why I did not show you my result here this result can be shown on notebook.

Q2[5] — Score-based Generative Modeling on MNIST

Objective

To implement a **Noise Conditional Score Network (NCSN)** trained via **Denoising Score Matching (DSM)** and sample using **Annealed Langevin Dynamics (ALD)** on MNIST.

Methodology

Network: A σ-conditioned U-Net with sinusoidal embeddings and FiLM modulation.

Loss:

$$L_{\text{DSM}}(\theta) = \mathbb{E}_{x,\sigma,\tilde{x}}\left[\frac{\sigma^2}{2}\|s_\theta(\tilde{x},\sigma) + \frac{\tilde{x}-x}{\sigma^2}\|^2\right],$$

where $\sigma \in [0.01, 1.0]$, $K = 10$, and $\lambda(\sigma) = \sigma^2$.

Sampling (ALD):

$$\tilde{x}_{t+1} = \tilde{x}_t + \frac{\alpha_t}{2} s_\theta(\tilde{x}_t, \sigma_t) + \sqrt{\alpha_t} z_t, \quad z_t \sim \mathcal{N}(0, I),$$

with $\alpha_t = c\sigma_t^2/\sigma_L^2$, $c \in [0.05, 0.2]$, $T \in [75, 150]$.

Training Setup

Parameter	Value
Epochs	40
Batch size	128
Learning rate	1×10^{-4} (Adam)
Gradient clip	1.0
σ range	[0.01, 1.0] (K=10)
Base width / Depth	64 / 3

DSM loss steadily decreased, showing stable training behavior.

Results

Qualitative: Samples from ALD show recognizable digits at lower c values and more steps.

Quantitative:

- Inception Score (IS): **1.99 ± 0.05**
- FID: **139.9**

Observations:

- Smaller c and higher steps ⇒ smoother, cleaner samples.
- σ-conditioning improves denoising generalization.

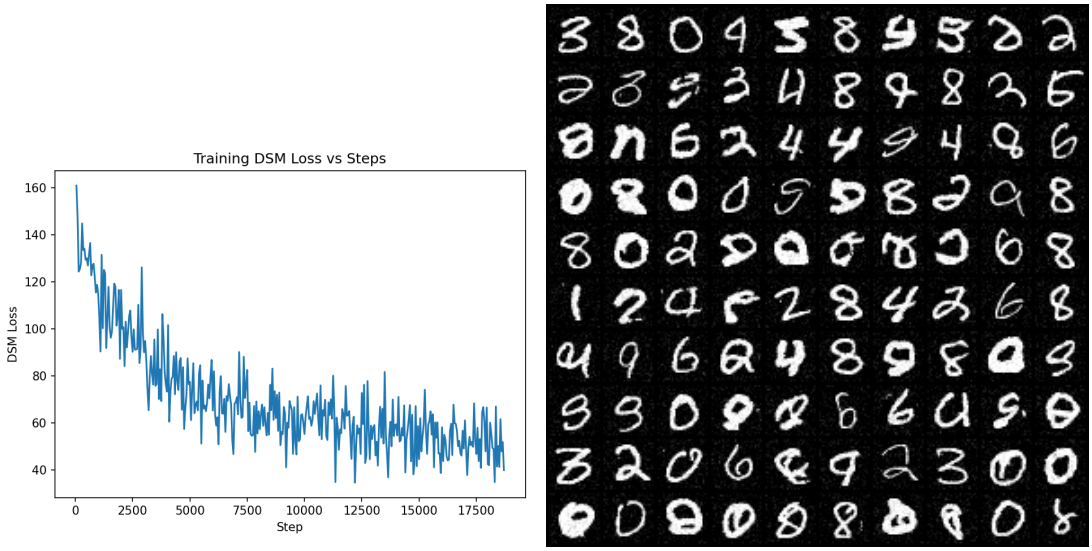

Figure 2: *
Q2 — Left: DSM training loss. Right: ALD-generated MNIST samples.

Conclusion

- For $c = 0.1$ and 64 samples, digits were mostly clear and distinct.
- Increasing samples to 100 (with same $c = 0.1$) improved diversity with minor blurring.
- For $c = 0.5$ and 125 samples, digits became little noisier.

A σ-conditioned U-Net was trained via DSM loss and sampled using ALD to generate MNIST digits. Despite modest FID and IS, the model demonstrates core score-based generative principles effectively.

I took help of AI while implementing assignment, but it would be only 15% not more than that. I could able to understand the code properly.

References

- Song, J. et al. (2025). *M3FD: A Multimodal Multi-domain Multi-weather Dataset.*
- Song, Y., & Ermon, S. (2020). *Denoising Diffusion Implicit Models (DDIM).* arXiv:2010.02502
- Song, Y., & Ermon, S. (2019). *Generative Modeling by Estimating Gradients of the Data Distribution.* NeurIPS.