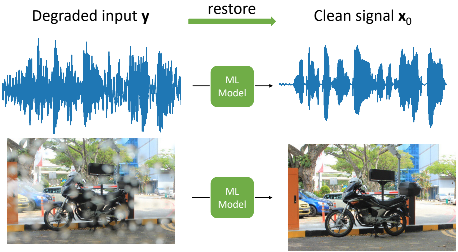
RestoreGrad: Signal Restoration Using Conditional Denoising Diffusion Models with Jointly Learned Prior

Ching-Hua Lee, Chouchang Yang, Jaejin Cho, Yashas Malur Saidutta, Rakshith Sharma Srinivasa, Yilin Shen, Hongxia Jin

Artificial Intelligence Center - Mountain View, Samsung Electronics
Mountain View, CA, USA

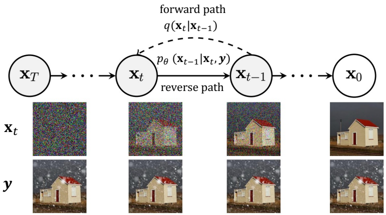
Signal Restoration with DDPMs

- Signal restoration problems: Restoring clean signal $\mathbf{x}_0$ from degraded input $\mathbf{y}$



- **Denoising Diffusion Probabilistic Models (DDPMs)** demonstrate promising results for generating high-fidelity data with strong generalizability

- Can be used for signal restoration tasks, by *conditioning* the DDPM model θ on the observed degraded signal $\mathbf{y}$ to recover $\mathbf{x}_0$



Inefficiency in Existing DDPM Modeling

- However, most existing DDPMs assume **a standard Gaussian prior** for simplicity, ignoring the correlation between the degraded and clean signals, leading to inefficiencies in both training and inference
- We propose RestoreGrad to improve modeling efficiency of DDPMs
  - The main idea is to jointly learn the diffusion prior with the conditional DDPM model

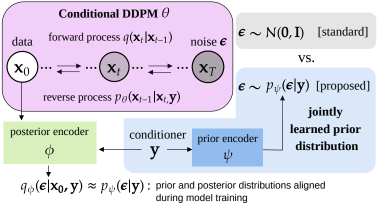


Figure: Proposed vs. standard methods.

Figure: Diffusion processes visualized.

Jointly Learnable Diffusion Priors: Combine with VAE

- To recover a clean signal $\mathbf{x}_0$ given a noisy signal $\mathbf{y}$ with a model parameterized by θ:

$$\max_{\theta} \; \log p_\theta(\mathbf{x}_0|\mathbf{y}) : \text{ conditional data log-likelihood} \tag{1}$$

- Evidence lower bounds (ELBOs) for solving (1):
  - Conditional Variational Autoencoders (VAEs):

$$\geq \mathbb{E}_{q_\phi(\boldsymbol{\epsilon}|\mathbf{x}_0,\mathbf{y})}\left[\log p_\theta(\mathbf{x}_0|\mathbf{y},\boldsymbol{\epsilon})\right] - D_{\mathrm{KL}}\left(q_\phi(\boldsymbol{\epsilon}|\mathbf{x}_0,\mathbf{y})||p_\theta(\boldsymbol{\epsilon}|\mathbf{y})\right) \tag{2}$$

  - Conditional DDPMs:

$$\geq \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\left[\log \frac{p_\theta(\mathbf{x}_{0:T}|\mathbf{y})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\right] \tag{3}$$

  - **RestoreGrad (ours)**:

$$\geq \mathbb{E}_{q_\phi(\boldsymbol{\epsilon}|\mathbf{x}_0,\mathbf{y})}\left[\mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\left[\log \frac{p_\theta(\mathbf{x}_{0:T}|\mathbf{y},\boldsymbol{\epsilon})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\right]\right] - D_{\mathrm{KL}}\left(q_\phi(\boldsymbol{\epsilon}|\mathbf{x}_0,\mathbf{y})||p_\psi(\boldsymbol{\epsilon}|\mathbf{y})\right) \tag{4}$$

  Embraces the Best of Both Worlds: generative power (DDPM) for improved restoration quality, and modeling efficiency (VAE) for faster training/sampling

RestoreGrad Learning Framework

- The **new ELBO** that integrates DDPM into VAE:

$$\mathbb{E}_{q_\phi(\boldsymbol{\epsilon}|\mathbf{x}_0,\mathbf{y})}\Bigg[\underbrace{\mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\left[\log\frac{p_\theta(\mathbf{x}_{0:T}|\mathbf{y})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\right]}_{\text{conditional DDPM }\theta}\Bigg] - D_{\mathrm{KL}}\Big(\underbrace{q_\phi(\boldsymbol{\epsilon}|\mathbf{x}_0,\mathbf{y})}_{\text{Posterior Net }\phi}\ \Big\|\ \underbrace{p_\psi(\boldsymbol{\epsilon}|\mathbf{y})}_{\text{Prior Net }\psi}\Big) \quad (5)$$

leads to the joint optimization framework:

$$\min_{\theta,\phi,\psi}\ \mathcal{L}(\theta,\phi,\psi) = \eta\mathcal{L}_{\mathrm{LR}} + \mathcal{L}_{\mathrm{DM}} + \lambda\mathcal{L}_{\mathrm{PM}}, \quad \eta>0,\ \lambda>0 \quad (6)$$

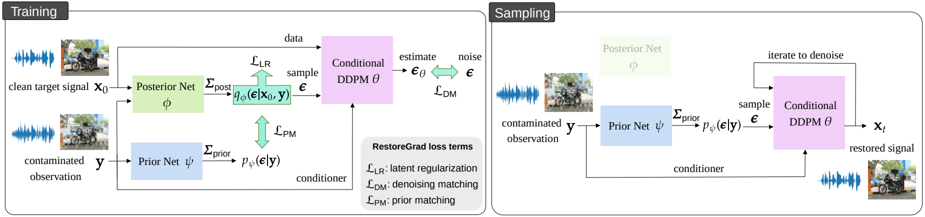


Figure: During training, the conditional DDPM θ, Prior Net ψ, and Posterior Net ϕ are jointly optimized by (6). During inference, the DDPM θ samples the latent noise $\boldsymbol{\epsilon}$ from the jointly learned prior distribution to synthesize the clean signal.

Key Difference from Related Work

- Our idea was motivated by existing work on using data-dependent diffusion priors. E.g., PriorGrad:
  - S.-g. Lee et al., "PriorGrad: Improving conditional denoising diffusion models with data-dependent adaptive prior," in *ICLR* 2022

- On top of that, we have introduced the idea of jointly learnable priors by employing the prior and posterior encoders, ψ and ϕ:
  - PriorGrad:

$$\min_{\theta} \|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \mathbf{y}, t)\|^2_{\boldsymbol{\Sigma}_y^{-1}}, \text{ where } \boldsymbol{\Sigma}_y = f(\mathbf{y}) \tag{7}$$

    and $f(\cdot)$ is a *pre-defined* function that maps $\mathbf{y}$ into the prior distribution.

  - **RestoreGrad (ours):**

$$\begin{aligned} \min_{\theta,\phi,\psi} \ & \eta\big(\bar{\alpha}_T\|\mathbf{x}_0\|^2_{\boldsymbol{\Sigma}^{-1}_{\text{post}}} + \log|\boldsymbol{\Sigma}_{\text{post}}|\big) + \|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \mathbf{y}, t)\|^2_{\boldsymbol{\Sigma}^{-1}_{\text{post}}} \\ & + \lambda\big(\log\frac{|\boldsymbol{\Sigma}_{\text{prior}}|}{|\boldsymbol{\Sigma}_{\text{post}}|} + \text{tr}(\boldsymbol{\Sigma}^{-1}_{\text{prior}}\boldsymbol{\Sigma}_{\text{post}})\big). \end{aligned} \tag{8}$$

    By utilizing encoders for the prior, we bypass the manual search process for a suitable mapping function $f(\cdot)$, which requires certain domain knowledge given a specific task. Our framework is thus applicable to **more modalities**.

Improved Model Learning Efficiency

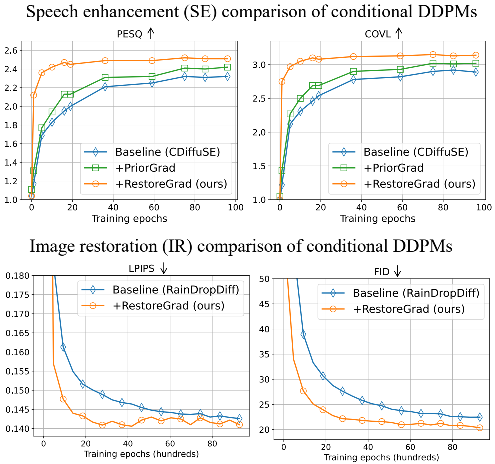


Figure: (Top) In speech domain, RestoreGrad outperforms PriorGrad, a recently proposed improvement to baseline DDPM (CDiffuSE) that leverages handcrafted data-dependent priors. (Bottom) In image domain, RestoreGrad provides a paradigm to improve DDPM baseline (RainDropDiff).

Inference Efficiency and Restoration Examples

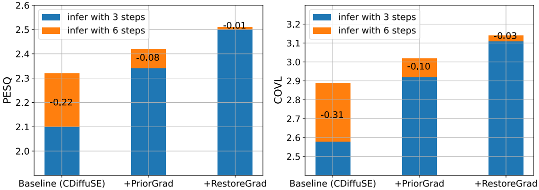


Figure: Robustness to the reduction in reverse sampling time steps for inference.

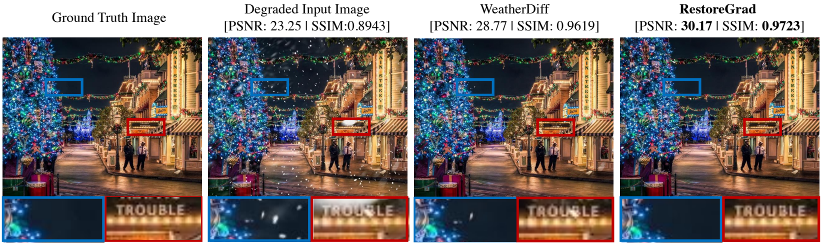


Figure: Image restoration examples. RestoreGrad was trained for 2 times fewer steps than WeatherDiff.

Conclusion

- New diffusion-based signal restoration through integrating conditional DDPMs with VAEs
- Leveraging jointly learnable diffusion priors to achieved improved restoration quality and faster training and sampling
- General and applicable to multiple modalities (audio, images, ...)