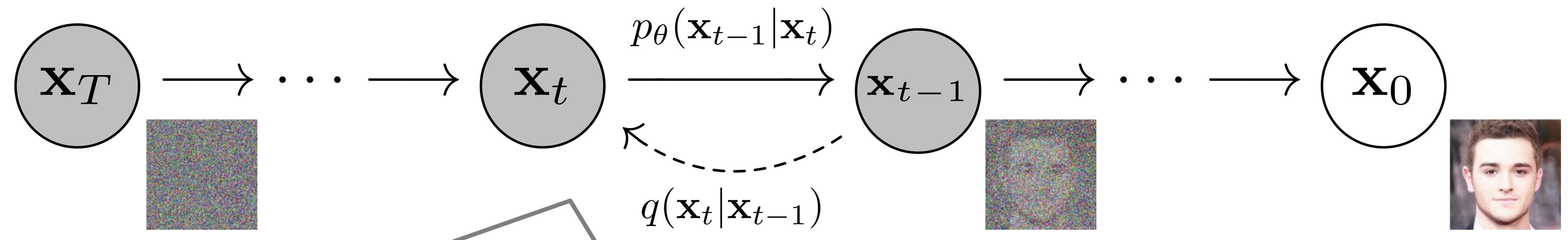
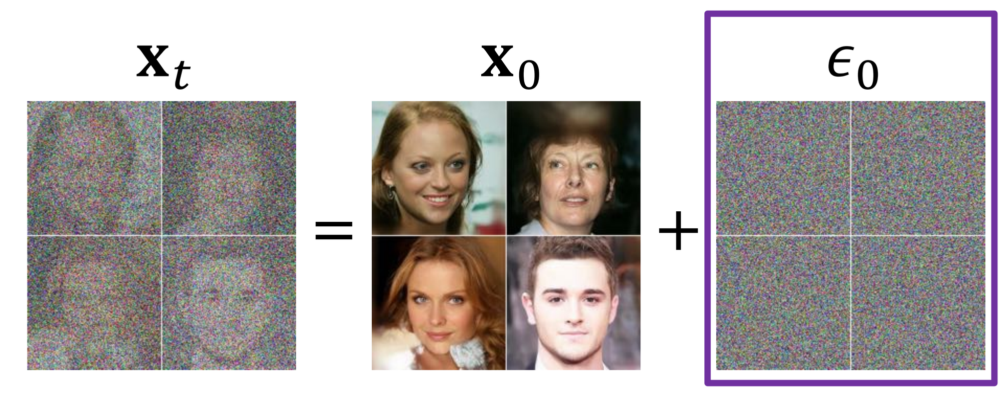
Positive-Unlabeled Diffusion Models for Preventing Sensitive Data Generation

Hiroshi Takahashi, Tomoharu Iwata, Atsutoshi Kumagai, Yuuki Yamanaka, Tomoya Yamashita
NTT Corporation

1. Diffusion Model

- Diffusion process gradually adds noise to data until it becomes Gaussian noise, and denoising process removes the noise using the estimator $\epsilon_\theta(\mathbf{x}_t, t)$, thereby generating data from noise [1].





The noise estimator $\epsilon_\theta(\mathbf{x}_t, t)$ is trained to minimize the squared error between the predicted noise and the true noise ϵ_0:

$$\ell_\theta(\mathbf{x}_0) \equiv \|\epsilon_0 - \epsilon_\theta(\mathbf{x}_t, t)\|^2$$

2. Problem, Setting and Goal

- **Problem**: If the dataset contains sensitive data, diffusion models may reproduce them.
- **Setting**: Removing all sensitive data is impractical since the dataset is large, but we can prepare a small set of such data.
- **Goal**: preventing sensitive data generation using unlabeled data along with a small amount of sensitive data.

3. Positive-Unlabeled (PU) Learning Assumption

- Let be $p_\mathcal{N}$, $p_\mathcal{S}$, and $p_\mathcal{U}$ be normal, sensitive, and unlabeled data distributions. Then, $p_\mathcal{N}$ can be rewritten as follows:

$$p_\mathcal{U}(\mathbf{x}) = \underbrace{\beta}_{\text{hyperparameter}} p_\mathcal{S}(\mathbf{x}) + (1-\beta)p_\mathcal{N}(\mathbf{x}) \Leftrightarrow p_\mathcal{N}(\mathbf{x}) = \frac{1}{1-\beta}\{p_\mathcal{U}(\mathbf{x}) - \beta p_\mathcal{S}(\mathbf{x})\}$$

4. Our Approach

- We want to minimize loss for normal data and maximize it for sensitive data. However, we cannot access normal data.
- Based on PU learning, we approximate training on normal data by using unlabeled data and sensitive data [2].

Minimize loss for normal data

$$\min_\theta \mathbb{E}_{p_\mathcal{N}}[\ell_\theta(\mathbf{x}_0)]$$

Maximize loss for sensitive data

$$\max_\theta \mathbb{E}_{p_\mathcal{S}}[\ell_\theta(\mathbf{x}_0)]$$

Intractable in practice

$$p_\mathcal{N}(\mathbf{x}) = \frac{1}{1-\beta}\{p_\mathcal{U}(\mathbf{x}) - \beta p_\mathcal{S}(\mathbf{x})\}$$

$$\min_\theta \mathbb{E}_{p_\mathcal{N}}[\ell_\theta(\mathbf{x}_0)] = \min_\theta \frac{1}{1-\beta}\{\mathbb{E}_{p_\mathcal{U}}[\ell_\theta(\mathbf{x}_0)] - \beta\mathbb{E}_{p_\mathcal{S}}[\ell_\theta(\mathbf{x}_0)]\}$$

5. Results

Task: Generate a generic middle-aged man (not Brad Pitt)
Data: 80 unlabeled (20% sensitive) & 20 labeled sensitive

Unlabeled Data | Sensitive Data | Supervised [3] | PU (Ours)

Ours blocked sensitive data with **99%** success vs. 84% for existing methods.

1. Ho et al. "Denoising diffusion probabilistic models." NeurIPS 2020.
2. Kiryo et al. "PU learning with non-negative risk estimator." NeurIPS 2017.
3. Heng & Soh. "Continual learning via selective forgetting in generative models." NeurIPS 2023.