Label-Noise Robust Diffusion Models

Byeonghu Na[1], Yeongmin Kim[1], HeeSun Bae[1], Jung Hyun Lee[2], Se Jung Kwon[2], Wanmo Kang[1], Il-Chul Moon[1,3]

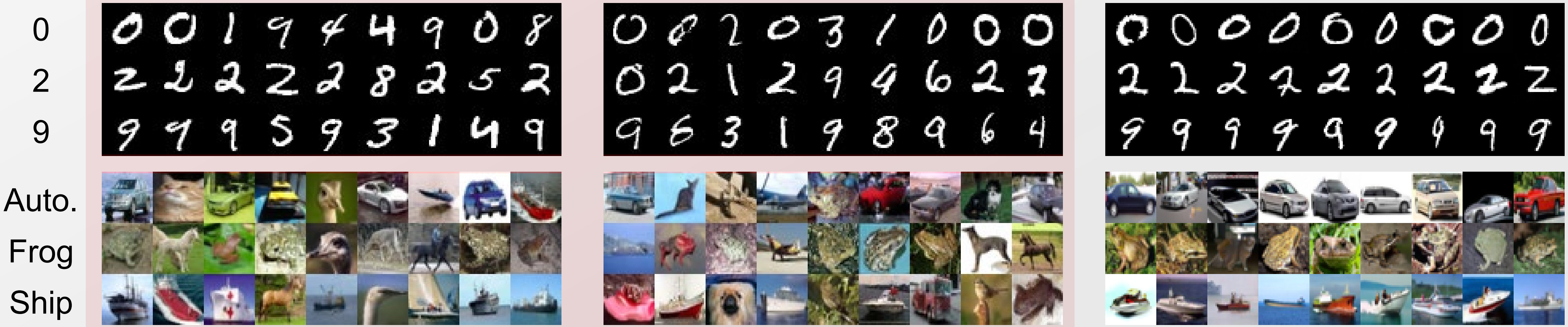
1 KAIST

2 NAVER Cloud

3 summary.ai

Motivation

- Diffusion models have gained significant interest for their high-quality sample generation.
- However, training diffusion models requires large-scale datasets, which often contain data instances with noisy labels.
- Noisy labels leads to condition mismatch and quality degradation of generated data.
- Although the problem of learning with noisy labels has been extensively studied in supervised learning, there are only a few studies on generative models.

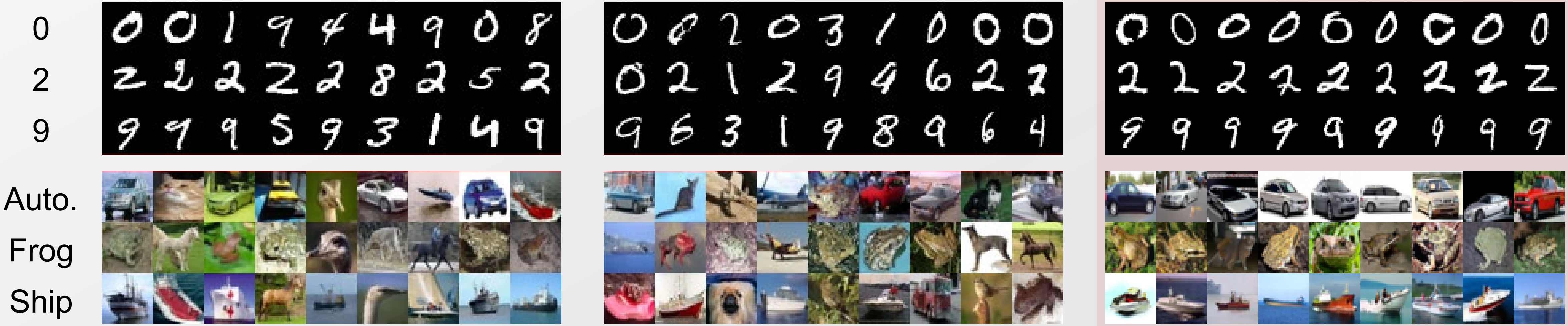


(a) Dataset with noisy labels (b) Baseline (c) Ours

Label-Noise Robust Diffusion Models

- We propose a method for training conditional diffusion models with noisy labels.
- We propose a training objective of diffusion models under label noise, called **Transition-aware weighted Denoising Score Matching (TDSM)** objective.

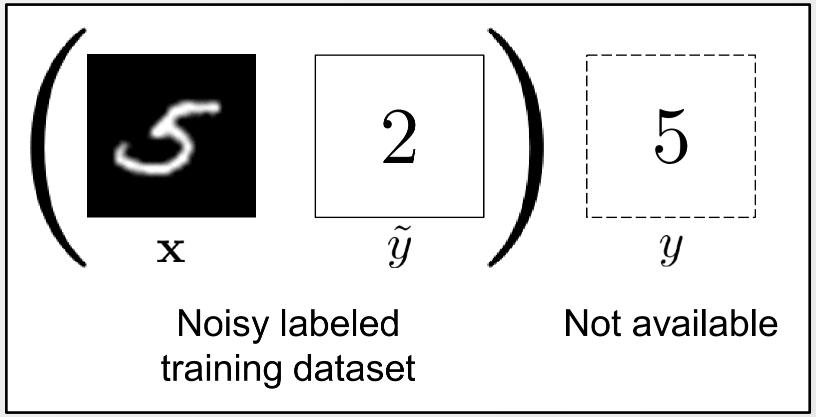
$$\mathcal{L}_{\mathrm{TDSM}}(\boldsymbol{\theta}; \tilde{p}_{\mathrm{data}}(\mathbf{X}, \tilde{Y})) := \mathbb{E}_t \left\{ \lambda(t)\, \mathbb{E}_{\mathbf{x}, \tilde{y} \sim \tilde{p}_{\mathrm{data}}}\, \mathbb{E}_{\mathbf{x}_t \sim p_{t|0}} \left[\left|\left| \sum_{y=1}^{c} \underbrace{w(\mathbf{x}_t, \tilde{y}, y, t)}_{=\, p_t(Y = y | \tilde{Y} = y, \mathbf{x}_t)} \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, y, t) - \nabla_{\mathbf{x}_t} \log p_{t|0}(\mathbf{x}_t | \mathbf{x}, \tilde{Y} = \tilde{y}) \right|\right|_2^2 \right] \right\}$$



(a) Dataset with noisy labels (b) Baseline (c) Ours

Problem Formulation Label Noise

- Setup
  - Data space $\mathcal{X} \in \mathbb{R}^d$, label space $\mathcal{Y} = \{1, ..., c\}$
  - Data instance $\mathbf{x} \in \mathcal{X}$, clean label $y \in \mathcal{Y}$, noisy label $\tilde{y} \in \tilde{\mathcal{Y}}$
  - Only have a noisy labeled training dataset $\tilde{D} = \{(\mathbf{x}^{(i)}, \tilde{y}^{(i)})\}_{i=1}^{n}$ from noisy-label data distribution $\tilde{p}_{\text{data}}(\mathbf{X}, \tilde{Y})$



- Class-conditional label-noise setting
  - The noisy label $\tilde{Y}$ is assumed to be independent of the instance X given the clean label Y.
  - From a generative perspective, it can be expressed as follows:

$$p(\mathbf{x}|\tilde{Y} = \tilde{y}) = \sum_{y=1}^{c} p(Y = y|\tilde{Y} = \tilde{y}) p(\mathbf{x}|Y = y, \tilde{Y} = \tilde{y}) = \sum_{y=1}^{c} p(Y = y|\tilde{Y} = \tilde{y}) p(\mathbf{x}|Y = y)$$

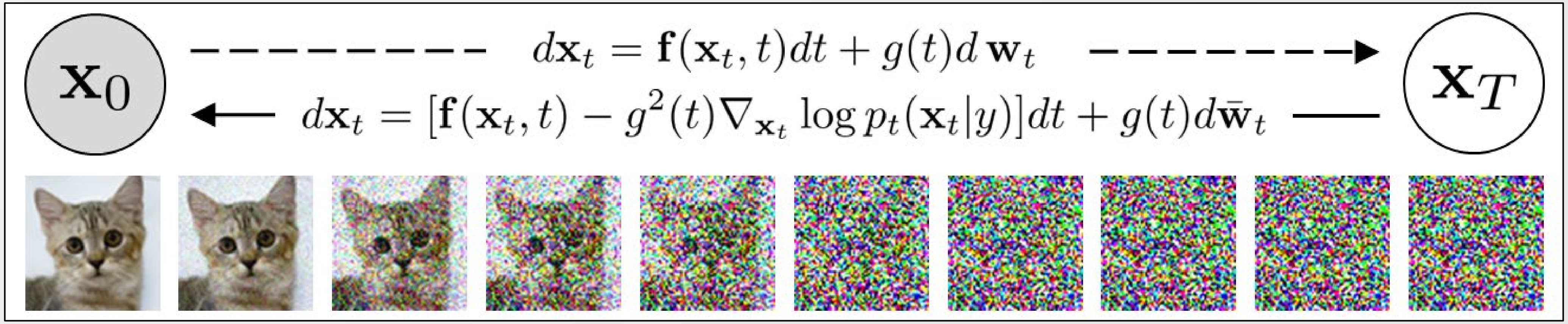
    - Each noisy-label conditional distribution is a mixture of clean-label conditional distribution.
  - We define a reverse transition matrix as $\boldsymbol{S} \in [0, 1]^{c \times c}$ where $S_{i,j} = p(Y = j|\tilde{Y} = i)$.

  - We will show that despite this instance-independent assumption, instance-dependent information is needed to overcome noisy labels in the diffusion model.

Problem Formulation Diffusion Models

- Diffusion models (or score-based generative models)
  - Sequentially corrupting training data with slowly increasing noise, and then learning to reverse this corruption in order to form a generative model of the data.
  - The key point is the score function, $\nabla_x \log p_t(x_t|y)$, which is the gradient of the log probability density with respect to data.
  - Therefore, the diffusion model aims to train the score network to approximate $\nabla_x \log p_t(x_t|y)$ through the score matching objective function, e.g., denoising score matching (DSM).

$$\mathcal{L}_{\mathrm{DSM}}(\boldsymbol{\theta}; p_{\mathrm{data}}(\mathbf{X}, Y)) := \mathbb{E}_t \left\{ \lambda(t)\, \mathbb{E}_{\mathbf{x}, y \sim p_{\mathrm{data}}} \mathbb{E}_{\mathbf{x}_t \sim p_{t|0}} \left[\left\| \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, y, t) - \nabla_{\mathbf{x}_t} \log p_{t|0}(\mathbf{x}_t | \mathbf{x}, \tilde{Y} = \tilde{y}) \right\|_2^2 \right] \right\}$$



- Learning diffusion models from noisy labels
  - If the score network is optimized by the original DSM objective with a noisy label dataset, then the score network converges on the noisy-label conditional score.

Remark. *Let* $\boldsymbol{\theta}^*_{DSM} := \arg\min_{\boldsymbol{\theta}} \mathcal{L}_{DSM}(\boldsymbol{\theta}; \tilde{p}_{data}(\mathbf{X}, \tilde{Y}))$ *be the optimal parameters obtained by minimizing the DSM objective. Then,* $\mathbf{s}_{\boldsymbol{\theta}^*_{DSM}}(\mathbf{x}_t, y, t) = \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|\tilde{Y} = y)$ *for all* $\mathbf{x}_t, y, t$.

  - To train the score network in the alignment of the clean-label conditional score, we modify the objective function to adjust the gradient signal from the score matching.
  - We start the adjustment by establishing the relationship between clean- and noisy-label conditional scores.

Methods Clean- and Noisy-Label Conditional Score

- Relationship between clean- and noisy-label conditional scores

$$\begin{bmatrix} \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|\tilde{Y}=1) \\ \vdots \\ \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|\tilde{Y}=c) \end{bmatrix} = \begin{bmatrix} w(\mathbf{x}_t, \tilde{Y}=1, Y=1, t) & \cdots & w(\mathbf{x}_t, \tilde{Y}=1, Y=c, t) \\ \vdots & \ddots & \vdots \\ w(\mathbf{x}_t, \tilde{Y}=c, Y=1, t) & \cdots & w(\mathbf{x}_t, \tilde{Y}=c, Y=c, t) \end{bmatrix} \begin{bmatrix} \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|Y=1) \\ \vdots \\ \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|Y=c) \end{bmatrix}$$

$$\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|\tilde{Y}=\tilde{y}) = \sum_{y=1}^{c} \quad w(\mathbf{x}_t, \tilde{y}, y, t) \quad \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|Y=y)$$

Noisy-label conditional scores | Transition-aware weight function | Clean-label conditional scores

$$\text{where } w(\mathbf{x}_t, \tilde{y}, y, t) := p(Y=y|\tilde{Y}=\tilde{y}) \frac{p_t(\mathbf{x}_t|Y=y)}{p_t(\mathbf{x}_t|\tilde{Y}=\tilde{y})} = p_t(Y=y|\tilde{Y}=\tilde{y}, \mathbf{x}_t)$$

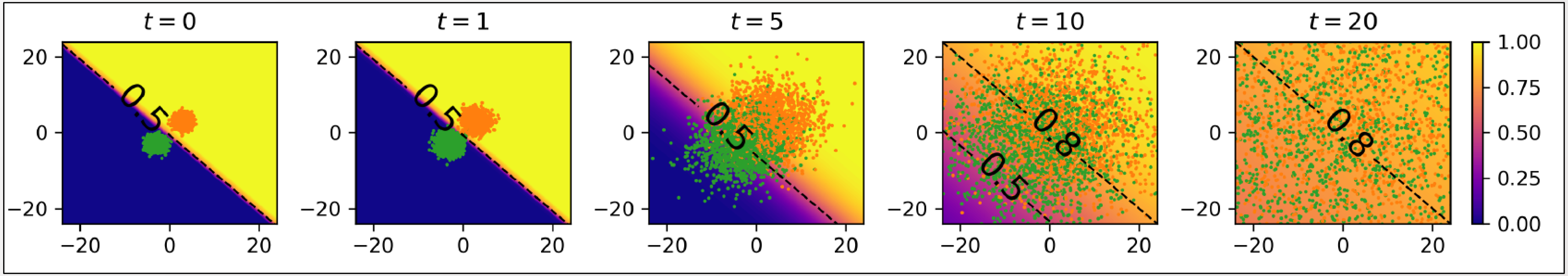
- The noisy-label conditional score can be expressed as a convex combination of the clean-label conditional scores with coefficient w.
  - $w(x_t, \tilde{y}, y, t) \geq 0$ & $\sum_{y=1}^{c} w(x_t, \tilde{y}, y, t) = 1$.

Methods Clean- and Noisy-Label Conditional Score

- Transition-aware weight function $w(x_t, \tilde{y}, y, t)$

$$w(\mathbf{x}_t, \tilde{y}, y, t) := p(Y = y|\tilde{Y} = \tilde{y}) \frac{p_t(\mathbf{x}_t|Y=y)}{p_t(\mathbf{x}_t|\tilde{Y}=\tilde{y})} = p_t(Y = y|\tilde{Y} = \tilde{y}, \mathbf{x}_t)$$

  - This function represents **instance-wise and time-dependent** (reverse) label transitions.
  - Training diffusion models with noisy labels poses a significant challenge because we need instance-dependent label noise information, even under the class-conditional label noise.



Contour maps of $w(x_t, \tilde{Y} = 1, Y = 1, t)$ in the 2-D Gaussian mixture model at different diffusion timesteps

- Transition-aware weighted Denoising Score Matching (TDSM)
  - Minimize the distance between the transition-aware weighted sum of conditional score network outputs and the perturbed data score.

$$\mathcal{L}_{\mathrm{TDSM}}(\boldsymbol{\theta}; \tilde{p}_{\mathrm{data}}(\mathbf{X}, \tilde{Y})) := \mathbb{E}_t \left\{ \lambda(t)\, \mathbb{E}_{\mathbf{x}, \tilde{y} \sim \tilde{p}_{\mathrm{data}}}\, \mathbb{E}_{\mathbf{x}_t \sim p_{t|0}} \left[\left|\left| \underbrace{\sum_{y=1}^{c} w(\mathbf{x}_t, \tilde{y}, y, t) \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, y, t)}_{\approx\, \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t | \tilde{Y} = \tilde{y})} - \nabla_{\mathbf{x}_t} \log p_{t|0}(\mathbf{x}_t | \mathbf{x}, \tilde{Y} = \tilde{y}) \right|\right|_2^2 \right] \right\}$$

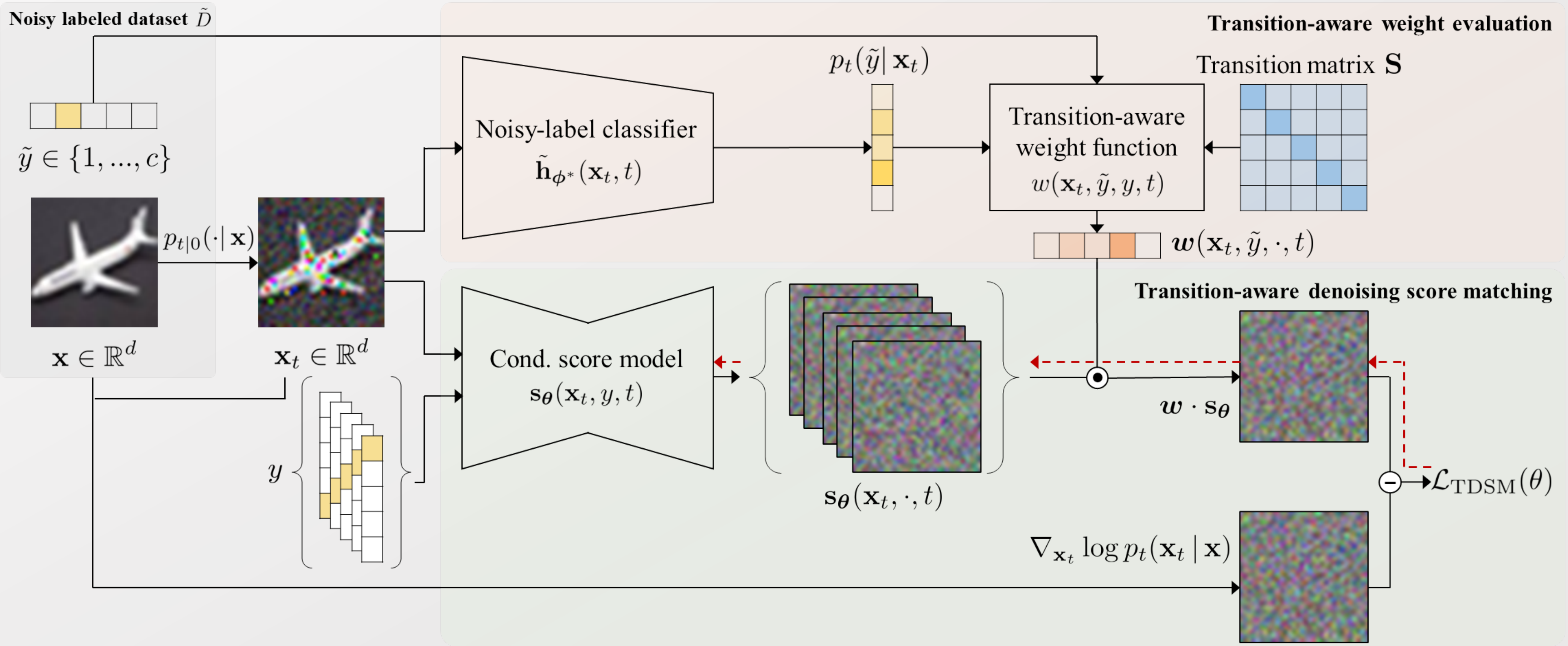
  - Theoretically, the score network trained by TDSM objective converges to the clean-label conditional score.

Theorem 3. *Let* $\boldsymbol{\theta}^*_{TDSM} := \arg\min_{\boldsymbol{\theta}} \mathcal{L}_{TDSM}(\boldsymbol{\theta}; \tilde{p}_{data}(\mathbf{X}, \tilde{Y}))$ *be the optimal parameters obtained by minimizing the TDSM objective. Then, under a class-conditional label noise setting with an invertible transition matrix,* $\mathbf{s}_{\boldsymbol{\theta}^*_{TDSM}}(\mathbf{x}_t, y, t) = \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t | Y = y)$ *for all* $\mathbf{x}_t, y, t$.

Methods Estimation of Transition-aware Weights

- We can estimate the transition-aware weight function $w(x_t, \tilde{y}, y, t)$ using the transition matrix S and the time-dependent noisy-label classifier $\tilde{h}_\phi(x_t, t)$.

$$w(\mathbf{x}_t, \tilde{y}, y, t) = p(Y = y|\tilde{Y} = \tilde{y})\frac{p_t(\mathbf{x}_t|Y=y)}{p_t(\mathbf{x}_t|\tilde{Y}=\tilde{y})} \quad \Rightarrow \quad \hat{w}(\mathbf{x}_t, \tilde{y}, y, t) = \frac{S_{\tilde{y},y} n_{\tilde{y}}}{\tilde{\mathrm{h}}_{\boldsymbol{\phi}}(\mathbf{x}_t,t)_{\tilde{y}}} \sum_{i=1}^{c} \frac{S^{-1}_{y,i} \tilde{\mathrm{h}}_{\boldsymbol{\phi}}(\mathbf{x}_t,t)_i}{n_i}$$

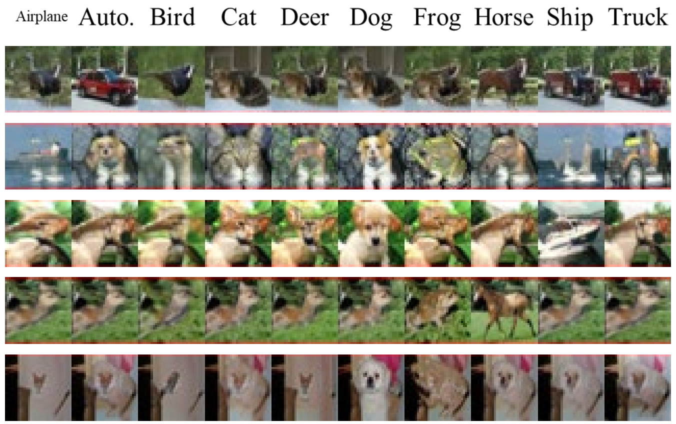


Experiment Results

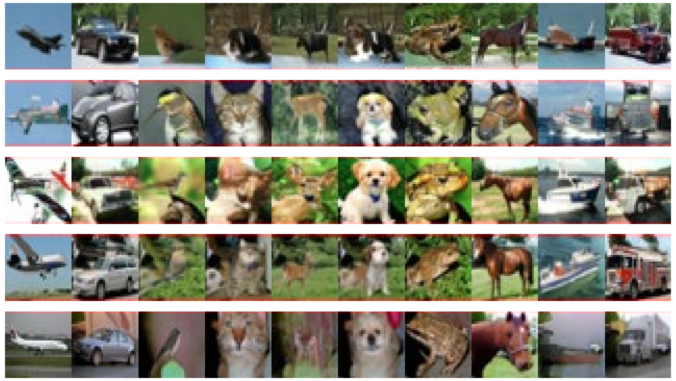
Analysis on benchmark dataset with synthetic label noise

		Metric		Symmetric				Asymmetric				Clean
				20%		40%		20%		40%		0%
				DSM	TDSM	DSM	TDSM	DSM	TDSM	DSM	TDSM	DSM
MNIST	un	Density	(↑)	81.11	**84.83**	81.93	**84.55**	84.23	**85.27**	84.47	**84.71**	86.20
		Coverage	(↑)	81.23	**82.16**	**81.65**	81.31	82.30	**82.45**	81.97	**82.27**	82.90
	cond	CAS	(↑)	94.31	**98.22**	72.52	**96.49**	95.25	**98.22**	89.29	**96.54**	98.55
		CW-Density	(↑)	69.78	**82.99**	55.70	**80.09**	78.58	**83.74**	73.54	**81.65**	85.79
		CW-Coverage	(↑)	76.77	**80.93**	70.45	**79.21**	79.97	**81.35**	77.50	**80.57**	82.09
CIFAR-10	un	FID	(↓)	**2.00**	2.06	**2.07**	2.43	2.02	**1.95**	2.23	**2.06**	1.92
		IS	(↑)	9.91	**9.97**	9.83	**9.96**	**10.06**	10.04	**10.09**	10.02	10.03
		Density	(↑)	100.03	**106.13**	100.94	**111.63**	100.66	**104.15**	101.25	**105.19**	103.08
		Coverage	(↑)	81.13	**81.89**	80.93	**82.03**	81.36	**81.81**	81.10	**81.90**	81.90
	cond	CW-FID	(↓)	16.21	**12.16**	30.45	**15.92**	11.97	**10.89**	15.18	**12.54**	10.23
		CAS	(↑)	66.80	**70.92**	47.21	**62.28**	72.66	**74.28**	68.98	**71.51**	77.74
		CW-Density	(↑)	88.45	**99.52**	73.02	**97.80**	96.10	**101.77**	92.13	**99.21**	102.63
		CW-Coverage	(↑)	77.80	**80.29**	71.63	**78.65**	79.95	**80.99**	78.12	**79.98**	81.57
CIFAR-100	un	FID	(↓)	**2.96**	4.26	**3.36**	6.85	2.76	**2.64**	**2.73**	2.81	2.51
		IS	(↑)	12.28	**12.29**	11.86	**12.07**	12.49	**12.79**	12.51	**12.57**	12.80
		Density	(↑)	83.01	**85.66**	81.70	**88.45**	87.36	**88.41**	**87.06**	87.01	87.98
		Coverage	(↑)	**75.02**	74.90	**73.92**	72.12	77.04	**77.46**	**76.56**	76.27	77.63
	cond	CW-FID	(↓)	79.91	**78.71**	100.04	**93.24**	75.39	**69.83**	89.13	**73.13**	66.97
		CAS	(↑)	25.49	**28.54**	15.41	**21.17**	33.31	**37.33**	23.50	**34.47**	39.50
		CW-Density	(↑)	66.47	**70.62**	49.77	**60.60**	72.14	**78.92**	60.27	**74.30**	82.58
		CW-Coverage	(↑)	70.11	**70.77**	60.64	**63.89**	71.08	**74.01**	64.19	**71.48**	75.78

Quantitative results with various noise settings



(a) DSM (base)


(b) TDSM (ours)

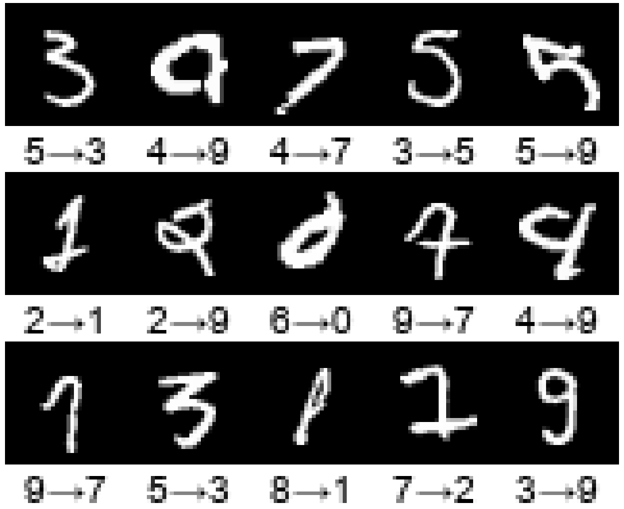
Generated images from baseline and our models

- Label noise in the diffusion model training degrades the sample quality and causes a class mismatch problem.
- The images generated by our model have better quality with an accurate class representation of the intended class than those generated by the baseline model.

Experiment Results
Analysis on benchmark dataset with annotated label

Metric		MNIST		CIFAR-10		CIFAR-100	
		DSM	TDSM	DSM	TDSM	DSM	TDSM
FID	(↓)	-	-	1.92	**1.91**	**2.51**	2.67
IS	(↑)	-	-	10.03	**10.10**	12.80	**12.85**
Density	(↑)	86.20	**88.08**	103.08	**104.35**	87.98	**90.04**
Coverage	(↑)	82.90	**83.69**	81.90	**82.07**	77.63	**78.28**

Quantitative results on the benchmark dataset with annotated label



Noisy labels of MNIST, captured by transition-aware weights.

- Our label-noise robust models consistently outperform the baseline models, indicating that existing benchmark datasets may suffer from noisy labels.
- Using the transition-aware weights, we find that the benchmark dataset also contains examples with noisy or ambiguous labels.

Experiment Results Combining with the existing noisy label corrector

	Metric		Symmetric		Asymmetric	
			DSM	TDSM	DSM	TDSM
un	FID	(↓)	**2.54**	2.84	4.00	**3.41**
	IS	(↑)	12.80	**12.94**	12.51	**12.83**
	Density	(↑)	87.28	**90.20**	83.65	**88.10**
	Coverage	(↑)	77.44	**77.63**	75.94	**77.57**
cond	CW-FID	(↓)	67.52	**67.33**	78.93	**76.62**
	CAS	(↑)	42.15	**42.39**	39.60	**39.72**
	CW-Density	(↑)	82.04	**85.44**	76.04	**81.69**
	CW-Coverage	(↑)	75.20	**75.61**	70.39	**71.62**

Quantitative results of combining with the noisy label corrector on the CIFAR-100 under 40% noise

- The existing classifiers to mitigate the noisy label can be considered as finding the true label after noise filtering.
- By pipelining this noisy label corrector and our TDSM approach, we can find a better noise filtering in terms of generation performance.
- Our approach tackles the noisy label problem from a diffusion model learning perspective, providing an orthogonal direction compared to conventional noisy label methods.

Thank you!

Paper

Code

Contact: byeonghu.na@kaist.ac.kr