AdjointDEIS: Efficient Gradients for Diffusion Models

Zander W. Blasingame   Chen Liu

Clarkson University

Problem

Diffusion models use a learned score function, $\boldsymbol{s}_\theta(\mathbf{x}_t, t) \approx \nabla_{\mathbf{x}} \log p_t(\mathbf{x}_t)$, or closely related function like a noise-prediction network $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) \approx -\sigma_t \nabla_{\mathbf{x}} \log p_t(\mathbf{x}_t)$ to turn noise from p_T into samples from the data distribution p_0.

Sampling is done via the reverse-time SDE

$$\mathrm{d}\mathbf{x}_t = [f(t)\mathbf{x}_t + \frac{g^2(t)}{\sigma_t}\boldsymbol{\epsilon}(\mathbf{x}_t, \mathbf{z}, t)]\,\mathrm{d}t + g(t)\,\mathrm{d}\bar{\mathbf{w}}_t, \tag{1}$$

where '$\mathrm{d}t$' is a *negative* timestep and $\{\bar{\mathbf{w}}_t\}_{t\in[0,T]}$ is the Wiener process as time flows backwards.

The Probability Flow ODE [1] has the same marginal distributions as Eq. (1):

$$\frac{\mathrm{d}\mathbf{x}_t}{\mathrm{d}t} = \boldsymbol{f}_\theta(\mathbf{x}_t, \mathbf{z}, t) \coloneqq f(t)\mathbf{x}_t + \frac{g^2(t)}{2\sigma_t}\boldsymbol{\epsilon}(\mathbf{x}_t, \mathbf{z}, t). \tag{2}$$

Problem Statement

We wish to solve the following optimization problems for the ODE or SDE variants:

$$\underset{\mathbf{x}_T, \mathbf{z}, \theta}{\arg\min}\ \mathcal{L}\left(\mathbf{x}_T + \int_T^0 f(t)\mathbf{x}_t + \frac{g^2(t)}{\sigma_t}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \mathbf{z}, t)\,\mathrm{d}t + \int_T^0 g(t)\,\mathrm{d}\bar{\mathbf{w}}_t\right), \tag{3}$$

$$\underset{\mathbf{x}_T, \mathbf{z}, \theta}{\arg\min}\ \mathcal{L}\left(\mathbf{x}_T + \int_T^0 f(t)\mathbf{x}_t + \frac{g^2(t)}{2\sigma_t}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \mathbf{z}, t)\,\mathrm{d}t\right). \tag{4}$$

Where $\mathcal{L}: \mathbb{R}^d \to \mathbb{R}$ is some scalar-valued differentiable guidance function defined on the output of the diffusion model.

To find the optimal $\mathbf{x}_T$, $\mathbf{z}$, and θ we perform gradient descent using the **continuous adjoint equations** [2], *i.e.*, we need to find

$$\mathbf{a}_{\mathbf{x}}(t) \coloneqq \frac{\partial \mathcal{L}}{\partial \mathbf{x}_t}, \qquad \mathbf{a}_{\mathbf{z}}(T) \coloneqq \frac{\partial \mathcal{L}}{\partial \mathbf{z}}, \qquad \mathbf{a}_\theta(T) \coloneqq \frac{\partial \mathcal{L}}{\partial \theta}. \tag{5}$$

Overview

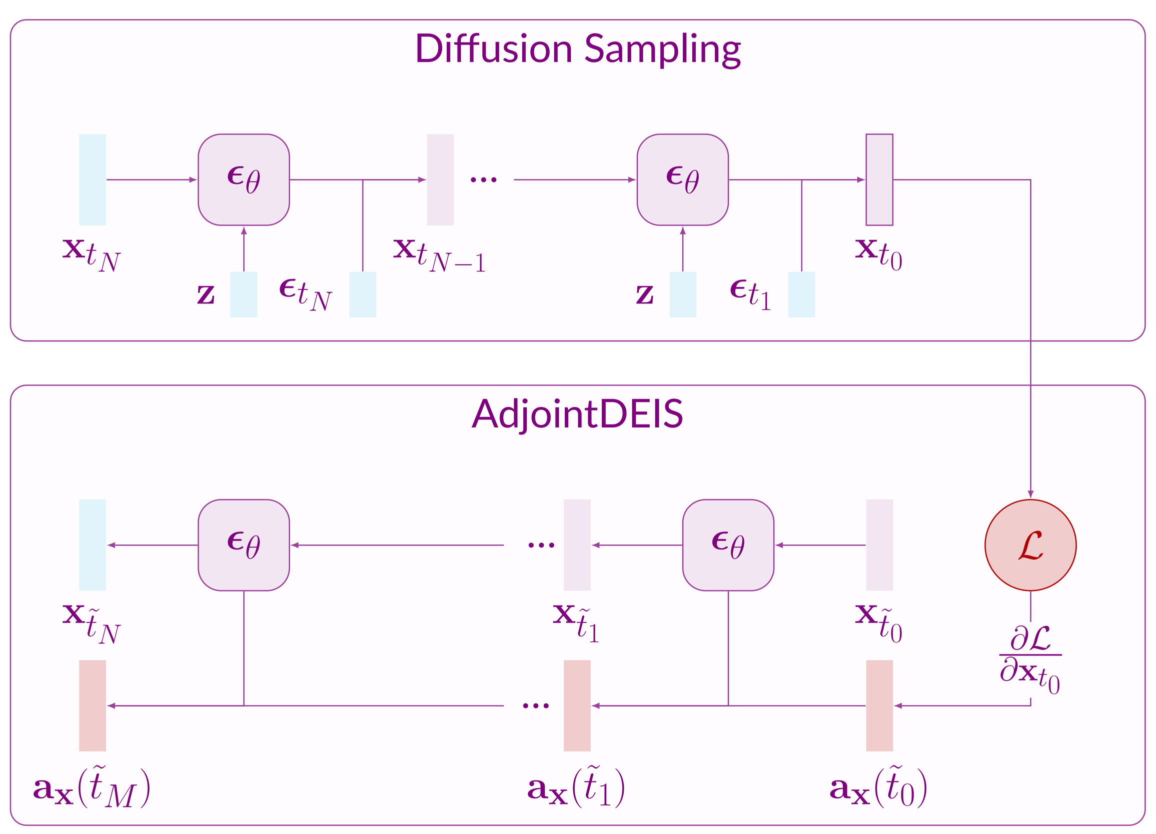


Figure 1. A high-level overview of the AdjointDEIS solver to the continuous adjoint equations for diffusion models. The sampling schedule consists of $\{t_n\}_{n=0}^N$ timesteps for the diffusion model and $\{\tilde{t}_n\}_{n=0}^M$ timesteps for AdjointDEIS. The gradients $\mathbf{a}_{\mathbf{x}}(T)$ can be used to optimize $\mathbf{x}_T$ to find some optimal $\mathbf{x}_T^*$.

The continuous adjoint equations for Eq. (4) are given by the IVP:

$$\begin{aligned}
\mathbf{a}_{\mathbf{x}}(0) &= \frac{\partial \mathcal{L}}{\partial \mathbf{x}_0}, & \frac{\mathrm{d}\mathbf{a}_{\mathbf{x}}}{\mathrm{d}t}(t) &= -\mathbf{a}_{\mathbf{x}}(t)^\top \frac{\partial \boldsymbol{f}_\theta(\mathbf{x}_t, \mathbf{z}, t)}{\partial \mathbf{x}_t}, \\
\mathbf{a}_{\mathbf{z}}(0) &= \mathbf{0}, & \frac{\mathrm{d}\mathbf{a}_{\mathbf{z}}}{\mathrm{d}t}(t) &= -\mathbf{a}_{\mathbf{x}}(t)^\top \frac{\partial \boldsymbol{f}_\theta(\mathbf{x}_t, \mathbf{z}, t)}{\partial \mathbf{z}}, \\
\mathbf{a}_\theta(0) &= \mathbf{0}, & \frac{\mathrm{d}\mathbf{a}_\theta}{\mathrm{d}t}(t) &= -\mathbf{a}_{\mathbf{x}}(t)^\top \frac{\partial \boldsymbol{f}_\theta(\mathbf{x}_t, \mathbf{z}, t)}{\partial \theta}.
\end{aligned} \tag{6}$$

This treats $\boldsymbol{f}_\theta(\mathbf{x}_t, \mathbf{z}, t)$ as a black-box; however, we can exploit the *special* formulation of diffusion ODEs to greatly simplify these equations.

Simplified Formulation of the Continuous Adjoint Equations

Proposition 1 (Exact solution of adjoint diffusion ODEs). *Given initial values* $[\mathbf{a}_{\mathbf{x}}(t), \mathbf{a}_{\mathbf{z}}(t), \mathbf{a}_\theta(t)]$ *at time* $t \in (0, T)$, *the solution* $[\mathbf{a}_{\mathbf{x}}(s), \mathbf{a}_{\mathbf{z}}(s), \mathbf{a}_\theta(s)]$ *at time* $s \in (t, T]$ *of adjoint diffusion ODEs in Eq. (6) is*

$$\mathbf{a}_{\mathbf{x}}(s) = \frac{\alpha_t}{\alpha_s}\mathbf{a}_{\mathbf{x}}(t) + \frac{1}{\alpha_s}\int_{\lambda_t}^{\lambda_s} \alpha_\lambda^2 e^{-\lambda}\mathbf{a}_{\mathbf{x}}(\lambda)^\top \frac{\partial \boldsymbol{\epsilon}_\theta(\mathbf{x}_\lambda, \mathbf{z}, \lambda)}{\partial \mathbf{x}_\lambda}\,\mathrm{d}\lambda, \tag{7}$$

$$\mathbf{a}_{\mathbf{z}}(s) = \mathbf{a}_{\mathbf{z}}(t) + \int_{\lambda_t}^{\lambda_s} \alpha_\lambda e^{-\lambda}\mathbf{a}_{\mathbf{x}}(\lambda)^\top \frac{\partial \boldsymbol{\epsilon}_\theta(\mathbf{x}_\lambda, \mathbf{z}, \lambda)}{\partial \mathbf{z}}\,\mathrm{d}\lambda, \tag{8}$$

$$\mathbf{a}_\theta(s) = \mathbf{a}_\theta(t) + \int_{\lambda_t}^{\lambda_s} \alpha_\lambda e^{-\lambda}\mathbf{a}_{\mathbf{x}}(\lambda)^\top \frac{\partial \boldsymbol{\epsilon}_\theta(\mathbf{x}_\lambda, \mathbf{z}, \lambda)}{\partial \theta}\,\mathrm{d}\lambda. \tag{9}$$

We can construct k-th order solvers using multi-step methods to estimate the n-th derivative of the *scaled* vector-Jacobian product, which we denote by

$$\mathbf{V}^{(n)}(\mathbf{x}; \lambda_t) = \frac{\mathrm{d}^n}{\mathrm{d}\lambda^n}\left[\alpha_\lambda^2 \mathbf{a}_{\mathbf{x}}(\lambda)^\top \frac{\partial \boldsymbol{\epsilon}_\theta(\mathbf{x}_\lambda, \mathbf{z}, \lambda)}{\partial \mathbf{x}_\lambda}\right]_{\lambda=\lambda_t}. \tag{10}$$

Using a Taylor Expansion about λ_t and letting $h = \lambda_s - \lambda_t$ we can rewrite Eq. (7) as

$$\mathbf{a}_{\mathbf{x}}(s) = \underbrace{\frac{\alpha_t}{\alpha_s}\mathbf{a}_{\mathbf{x}}(t)}_{\substack{\text{Linear term} \\ \textbf{Exactly computed}}} + \frac{1}{\alpha_s}\sum_{n=0}^{k-1} \underbrace{\mathbf{V}^{(n)}(\mathbf{x}; \lambda_t)}_{\substack{\text{Derivatives} \\ \textbf{Approximated}}} \underbrace{\int_{\lambda_t}^{\lambda_s} \frac{(\lambda - \lambda_t)^n}{n!} e^{-\lambda}\,\mathrm{d}\lambda}_{\substack{\text{Coefficients} \\ \textbf{Analytically computed}}} + \underbrace{\mathcal{O}(h^{k+1})}_{\substack{\text{Higher-order errors} \\ \textbf{Omitted}}}. \tag{11}$$

AdjointDEIS-1. Given an initial augmented adjoint state $[\mathbf{a}_{\mathbf{x}}(t), \mathbf{a}_{\mathbf{z}}(t), \mathbf{a}_\theta(t)]$ at time $t \in (0, T)$, the solution $[\mathbf{a}_{\mathbf{x}}(s), \mathbf{a}_{\mathbf{z}}(s), \mathbf{a}_\theta(s)]$ at time $s \in (t, T]$ is approximated by

$$\begin{aligned}
\mathbf{a}_{\mathbf{x}}(s) &= \frac{\alpha_t}{\alpha_s}\mathbf{a}_{\mathbf{x}}(t) + \sigma_s(e^h - 1)\frac{\alpha_t^2}{\alpha_s^2}\mathbf{a}_{\mathbf{x}}(t)^\top \frac{\partial \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \mathbf{z}, t)}{\partial \mathbf{x}_t}, \\
\mathbf{a}_{\mathbf{z}}(s) &= \mathbf{a}_{\mathbf{z}}(t) + \sigma_s(e^h - 1)\frac{\alpha_t}{\alpha_s}\mathbf{a}_{\mathbf{x}}(t)^\top \frac{\partial \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \mathbf{z}, t)}{\partial \mathbf{z}}, \\
\mathbf{a}_\theta(s) &= \mathbf{a}_\theta(t) + \sigma_s(e^h - 1)\frac{\alpha_t}{\alpha_s}\mathbf{a}_{\mathbf{x}}(t)^\top \frac{\partial \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \mathbf{z}, t)}{\partial \theta}.
\end{aligned} \tag{12}$$

Theoretical Results

For the AdjointDEIS solvers, we make similar assumptions to Lu *et al.* [3].

Assumption 1. *The total derivatives of the vector-Jacobian products* $\mathbf{V}^{(n)}(\{\mathbf{x}_\lambda, \mathbf{z}, \theta\}, \lambda)$ *as a function of* λ *exist and are continuous for* $0 \le n \le k-1$ *(and hence bounded).*

Assumption 2. *The function* $\boldsymbol{\epsilon}_\theta(\mathbf{x}, \mathbf{z}, t)$ *is continuous in* t *and uniformly Lipschitz and continuously differentiable w.r.t. its first parameter* $\mathbf{x}$.

Assumption 3. $h_{max} \coloneqq \max_{1\le j\le M} h_j = \mathcal{O}(1/M)$.

Assumption 4. $\rho_i > c > 0$ *for all* $i = 1, \dots, M$ *and some constant* c.

Theorem 1 (AdjointDEIS-k as a k-th order solver). *Assume the function* $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \mathbf{z}, t)$ *and its associated vector-Jacobian products follow the regularity conditions detailed above, then for* $k = 1, 2$, *AdjointDEIS-*k *is a* k*-th order solver for adjoint diffusion ODEs, i.e., for the sequence* $\{\tilde{\mathbf{a}}_{\mathbf{x}}(t_i)\}_{i=1}^M$ *computed by AdjointDEIS-*k, *the global truncation error at time* T *satisfies* $\tilde{\mathbf{a}}_{\mathbf{x}}(t_M) - \mathbf{a}_{\mathbf{x}}(T) = \mathcal{O}(h_{max}^2)$, *where* $h_{max} = \max_{1\le j\le M}(\lambda_{t_i} - \lambda_{t_{i-1}})$. *Likewise, AdjointDEIS-*k *is a* k*-th order solver for the estimated gradients w.r.t.* $\mathbf{z}$ *and* θ.

Theorem 2. *Let* $\boldsymbol{f}: \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d$ *be in* $\mathcal{C}_b^{\infty,1}$ *and* $\boldsymbol{g}: \mathbb{R} \to \mathbb{R}^{d\times w}$ *be in* $\mathcal{C}_b^1$. *Let* $\mathcal{L}: \mathbb{R}^d \to \mathbb{R}$ *be a scalar-valued differentiable function. Let* $\mathbf{w}_t: [0, T] \to \mathbb{R}^w$ *be a* w*-dimensional Wiener process. Let* $\mathbf{x}: [0, T] \to \mathbb{R}^d$ *solve the Stratonovich SDE*

$$\mathrm{d}\mathbf{x}_t = \boldsymbol{f}(\mathbf{x}_t, t)\,\mathrm{d}t + \boldsymbol{g}(t) \circ \mathrm{d}\mathbf{w}_t,$$

with initial condition $\mathbf{x}_0$. *Then the adjoint process* $\mathbf{a}_{\mathbf{x}}(t) \coloneqq \partial\mathcal{L}(\mathbf{x}_T)/\partial\mathbf{x}_t$ *is a strong solution to the backwards-in-time ODE*

$$\mathrm{d}\mathbf{a}_{\mathbf{x}}(t) = -\mathbf{a}_{\mathbf{x}}(t)^\top \frac{\partial \boldsymbol{f}}{\partial \mathbf{x}_t}(\mathbf{x}_t, t)\,\mathrm{d}t. \tag{13}$$

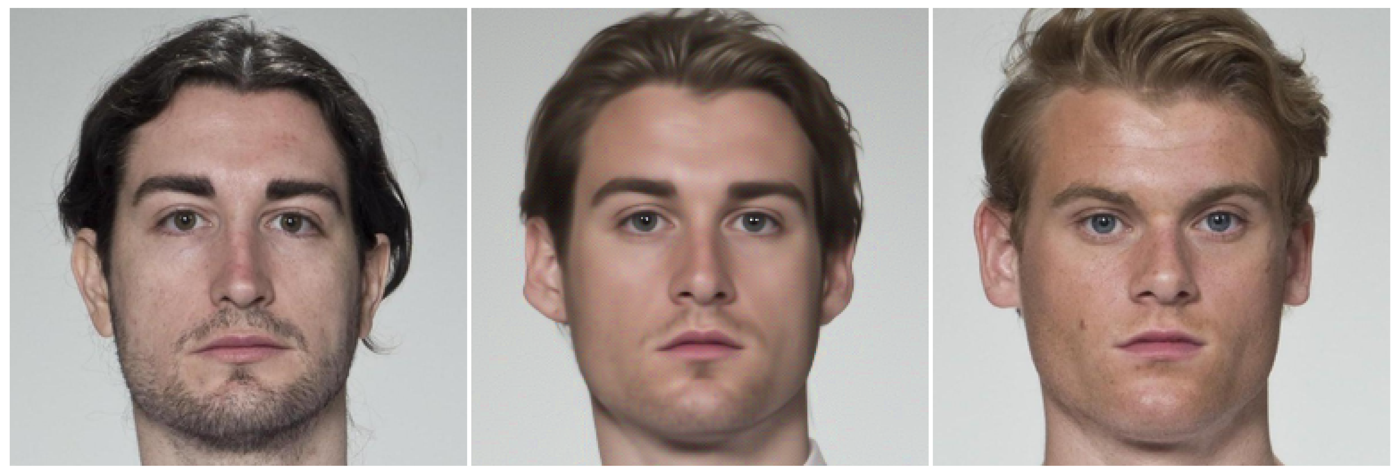
Proposition 2 (Exact solution of adjoint diffusion SDEs). *Given initial values* $[\mathbf{a}_{\mathbf{x}}(t), \mathbf{a}_{\mathbf{z}}(t), \mathbf{a}_\theta(t)]$ *at time* $t \in (0, T)$, *the solution* $[\mathbf{a}_{\mathbf{x}}(s), \mathbf{a}_{\mathbf{z}}(s), \mathbf{a}_\theta(s)]$ *at time* $s \in (t, T]$ *of adjoint diffusion SDEs is*

$$\mathbf{a}_{\mathbf{x}}(s) = \frac{\alpha_t}{\alpha_s}\mathbf{a}_{\mathbf{x}}(t) + \frac{2}{\alpha_s}\int_{\lambda_t}^{\lambda_s} \alpha_\lambda^2 e^{-\lambda}\mathbf{a}_{\mathbf{x}}(\lambda)^\top \frac{\boldsymbol{\epsilon}_\theta(\mathbf{x}_\lambda, \mathbf{z}, \lambda)}{\partial \mathbf{x}_\lambda}\,\mathrm{d}\lambda, \tag{14}$$

$$\mathbf{a}_{\mathbf{z}}(s) = \mathbf{a}_{\mathbf{z}}(t) + 2\int_{\lambda_t}^{\lambda_s} \alpha_\lambda e^{-\lambda}\mathbf{a}_{\mathbf{x}}(\lambda)^\top \frac{\partial \boldsymbol{\epsilon}_\theta(\mathbf{x}_\lambda, \mathbf{z}, \lambda)}{\partial \mathbf{z}}\,\mathrm{d}\lambda, \tag{15}$$

$$\mathbf{a}_\theta(s) = \mathbf{a}_\theta(t) + 2\int_{\lambda_t}^{\lambda_s} \alpha_\lambda e^{-\lambda}\mathbf{a}_{\mathbf{x}}(\lambda)^\top \frac{\partial \boldsymbol{\epsilon}_\theta(\mathbf{x}_\lambda, \mathbf{z}, \lambda)}{\partial \theta}\,\mathrm{d}\lambda. \tag{16}$$

Experimental Results



(a) Identity a (b) Guided face morph   (c) Identity b

Figure 2. Example of guided morphed face generation with AdjointDEIS on the FRLL dataset.

We demonstrate the use of AdjointDEIS in a guided generation problem of face morphing. The face morphing attack seeks to create an image which triggers a false accept with **both** identities in the targeted Face Recognition (FR) system.

We use AdjointDEIS to find the optimal $(\mathbf{x}_T, \mathbf{z})$ in the Diffusion Morph (DiM) pipeline [4]. We evaluate against three SOTA FR systems: ArcFace, AdaFace, and ElasticFace. We measure the performance with the Mated Morph Presentation Match Rate (MMPMR) metric which measures how many morphs are successful in fooling the FR system.

Table 1. Vulnerability of different FR systems across different morphing attacks on the SYN-MAD 2022 dataset. FMR = 0.1%.

		MMPMR(↑)		
Morphing Attack	NFE(↓)	AdaFace	ArcFace	ElasticFace
Webmorph	-	97.96	96.93	98.36
MIPGAN-I	-	72.19	77.51	66.46
MIPGAN-II	-	70.55	72.19	65.24
DiM-A	350	92.23	90.18	93.05
Fast-DiM	300	92.02	90.18	93.05
Morph-PIPE	2350	95.91	92.84	95.5
DiM + AdjointDEIS-1 (ODE)	2250	**99.8**	**98.77**	**99.39**
DiM + AdjointDEIS-1 (SDE)	2250	98.57	97.96	97.75

Key Insights

Exponential Integrators. The continuous adjoint equations preserve the same semi-linear structure of diffusion ODEs/SDEs allowing the use of numerical methods based on exponential integrators.

Adjoint diffusion SDEs are actually ODEs. The continuous adjoint equations for diffusion SDEs are greatly simplified allowing us to express them as essentially an ODE.

Efficient gradients. We can calculate the gradients using k-th order solvers with discretization steps **independent** of the sampling process.

References

[1] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *International Conference on Learning Representations*, 2021.

[2] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud, "Neural ordinary differential equations," in *Advances in Neural Information Processing Systems* (S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds.), vol. 31, Curran Associates, Inc., 2018.

[3] C. Lu, Y. Zhou, F. Bao, J. Chen, C. LI, and J. Zhu, "Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps," in *Advances in Neural Information Processing Systems* (S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, eds.), vol. 35, pp. 5775–5787, Curran Associates, Inc., 2022.

[4] Z. W. Blasingame and C. Liu, "Leveraging diffusion for strong and high quality face morphing attacks," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 6, no. 1, pp. 118–131, 2024.