Balanced Dataset Distillation via Modeling Multiple Visual Pattern Distribution

Supplementary Material

6. Theoretical Justification

This section provides detailed derivations for **Proposition 2** and **Proposition 3** presented in the paper.

6.1. Proof of Proposition 2

Proposition 2: The coreset selected by BPS effectively covers the information of the original data manifold.

As discussed in Sec.3.4, BPS constructs a pattern-balanced coreset $\mathcal{V}_\mathcal{I} = \mathcal{V}_{\hat{c}} \cup \mathcal{V}_{\hat{b}}$ by evenly selecting samples from pattern centers and margins based on the modeled multiple visual pattern distribution. To verify that $\mathcal{V}_\mathcal{I}$ effectively covers the information of the manifold $\mathcal{M}$ (Eq. (4) in the paper), the fill distance $d_{fill}(\mathcal{V}_\mathcal{I}, \mathcal{M})$ is employed as a metric, defined as:

$$d_{fill}(\mathcal{V}_\mathcal{I}, \mathcal{M}) = \sup_{\mathbf{v}\in\mathcal{M}} \min_{\hat{g}\in\mathcal{V}_\mathcal{I}} \|\mathbf{v} - \hat{g}\|, \tag{18}$$

where the sample representation $\mathbf{v}$ is an arbitrary point on $\mathcal{M}$, and $\hat{g}$ denotes the representation of a sample in $\mathcal{V}_\mathcal{I}$.

We first analyze the information coverage of $\mathcal{V}_{\hat{c}}$ with respect to $\mathcal{M}$, i.e. $d_{fill}(\mathcal{V}_{\hat{c}}, \mathcal{M})$. For any point $\mathbf{v}$ on $\mathcal{M}$, assume it belongs to the m-th pattern, i.e., $\mathbf{v} \in \mathcal{M}_m$. Since $\hat{c}_m \in \mathcal{V}_{\hat{c}}$ is the center of this pattern, the minimum distance from $\mathbf{v}$ to $\mathcal{V}_{\hat{c}}$, $\min_{\hat{g}\in\mathcal{V}_{\hat{c}}} \|\mathbf{v} - \hat{g}\|$, must be less than or equal to the distance from $\mathbf{v}$ to $\hat{c}_m$, $\|\mathbf{v} - \hat{c}_m\|$, that is

$$\min_{\hat{g}\in\mathcal{V}_{\hat{c}}} \|\mathbf{v} - \hat{g}\| \le \|\mathbf{v} - \hat{c}_m\|. \tag{19}$$

By applying the triangle inequality, Eq. (19) can be further decomposed as

$$\begin{aligned} \min_{\hat{g}\in\mathcal{V}_{\hat{c}}} \|\mathbf{v} - \hat{g}\| &\le \|\mathbf{v} - \hat{c}_m\| \\ &= \|\mathbf{v} - \mu_m + \mu_m - \hat{c}_m\| \\ &\le \|\mathbf{v} - \mu_m\| + \|\mu_m - \hat{c}_m\|, \end{aligned} \tag{20}$$

where μ_m is the center of $\mathcal{M}_m$. For $\|\mathbf{v} - \mu_m\|$, according to the definition of manifold in the paper ($\mathcal{M}_m = \{\mathbf{v} \in \mathcal{V} \mid \|\mathbf{v} - \mu_m\| \le R_m\}$ in Eq. (4)), we have $\|\mathbf{v} - \mu_m\| \le R_m$. For $\|\mu_m - \hat{c}_m\|$, we define

$$\epsilon_{\text{clu}} = \max_{m\in\{1,..,M\}} \|\hat{c}_m - \mu_m\|. \tag{21}$$

Consequently, for $\forall \mathbf{v} \in \mathcal{M}$, the upper bound of $d_{fill}(\mathcal{V}_{\hat{c}}, \mathcal{M})$ becomes

$$\begin{aligned} d_{fill}(\mathcal{V}_{\hat{c}}, \mathcal{M}) &= \sup_{\mathbf{v}\in\mathcal{M}} \min_{\hat{g}\in\mathcal{V}_{\hat{c}}} \|\mathbf{v} - \hat{g}\| \\ &\le \max_m R_m + \epsilon_{\text{clu}}. \end{aligned} \tag{22}$$

Similarly, for the representation set $\mathcal{V}_{\hat{b}}$ of selected marginal samples, we define

$$\epsilon_{\text{eps}} = \max_{m\in\{1,..,M\}} \|\hat{b}_m - \mu_m\|. \tag{23}$$

Analogous to the decomposition in Eq. (20), the upper bound of $d_{fill}(\mathcal{V}_{\hat{b}}, \mathcal{M})$ between $\mathcal{V}_{\hat{b}}$ and $\mathcal{M}$ becomes

$$\begin{aligned} d_{fill}(\mathcal{V}_{\hat{b}}, \mathcal{M}) &= \sup_{\mathbf{v}\in\mathcal{M}} \min_{\hat{g}\in\mathcal{V}_{\hat{b}}} \|\mathbf{v} - \hat{g}\| \\ &\le \max_m R_m + \epsilon_{\text{eps}}. \end{aligned} \tag{24}$$

Since $\mathcal{V}_\mathcal{I} = \mathcal{V}_{\hat{c}} \cup \mathcal{V}_{\hat{b}}$, the fill distance between $\mathcal{V}_\mathcal{I}$ and $\mathcal{M}$ is determined by the minimum of the two individual fill distances, denoted as

$$\begin{aligned} d_{fill}(\mathcal{V}_\mathcal{I}, \mathcal{M}) &= \min\left(d_{fill}(\mathcal{V}_{\hat{c}}, \mathcal{M}), d_{fill}(\mathcal{V}_{\hat{b}}, \mathcal{M})\right) \\ &\le \max_m R_m + \min(\epsilon_{\text{clu}}, \epsilon_{\text{eps}}). \end{aligned} \tag{25}$$

For R_m, it is known that the first term of $\mathcal{L}_{\text{clu}}$ (Eq. (8)) in the paper is

$$\mathcal{L}_{\text{clu, intra}} = \frac{1}{|\mathcal{T}|} \sum_{i=1}^{|\mathcal{T}|} \frac{1}{|\mathcal{O}^*_{o_i}|} \sum_{j\in\mathcal{O}^*_{o_i}} \|\mathbf{v}_i - \mathbf{v}_j\|_1, \tag{26}$$

which aims to minimize the average pairwise distance within each pattern. Based on the GMM assumption in Axiom 1, the pattern radius R_m is proportional to the standard deviation of the pattern distribution σ_m, i.e., $R_m \propto \sigma_m$. Since $\mathcal{L}_{\text{clu, intra}}$ quantifies the scattering of sample representations (i.e., the standard deviation σ_m), R_m satisfies the following inequality,

$$R_m \propto \sigma_m \le \lambda \cdot \mathcal{L}^m_{\text{clu, intra}}, \tag{27}$$

where λ is a constant, and $\mathcal{L}^m_{\text{clu, intra}}$ is the computed loss of the m-th pattern. Consequently, Eq. (25) becomes

$$d_{fill}(\mathcal{V}_\mathcal{I}, \mathcal{M}) \le \lambda \cdot \mathcal{L}_{\text{clu, intra}} + \min(\epsilon_{\text{clu}}, \epsilon_{\text{eps}}). \tag{28}$$

Therefore, during training, $\mathcal{L}_{\text{clu, intra}}$ will decrease to a small value. Meanwhile, $\arg\min_\mathcal{O} \mathrm{H}(P_\mathcal{T}|\mathcal{O})$ in Eq. (7) ensures that $\hat{c}_m$ approaches μ_m, driving ϵ_{clu} to approximate zero by the end of training. Furthermore, benefitting from $\mathcal{L}_{\text{clu}}$, ϵ_{eps} is maintained at a small value during training. Thus, when $d_{fill}(\mathcal{V}_\mathcal{I}, \mathcal{M})$ is decreasing to a negligible value, it indicates that the pattern-balanced coreset selected by BPS can cover the information of the original data manifold.

6.2. Proof of Proposition 3

Proposition 3. The risk of $h_{\mathcal{T}_\mathcal{I}}$ on the original dataset can approximate the risk of $h_\mathcal{T}$, i.e., $\mathcal{R}_\mathcal{T}(h_\mathcal{T}) \simeq \mathcal{R}_\mathcal{T}(h_{\mathcal{T}_\mathcal{I}})$.

Following Eq. (15) in the paper, the empirical risk gap $\mathcal{R}_{\text{diff}}$ between $h_{\mathcal{T}_\mathcal{I}}$ and $h_\mathcal{T}$ is defined as

$$\mathcal{R}_{\text{diff}} = \mathop{\mathbb{E}}_{\mathbf{v}\sim\mathcal{V}_\mathcal{T}} [\, \|h_{\mathcal{T}_\mathcal{I}}(\mathbf{v}) - h_\mathcal{T}(\mathbf{v})\| \,]. \tag{29}$$

For any point $\mathbf{v}$ on the manifold $\mathcal{M}$, let $\hat{g}^* \in \mathcal{V}_\mathcal{I}$ denote the nearest sample in the coreset to $\mathbf{v}$, i.e., $\hat{g}^* = \arg\min_{\hat{g}\in\mathcal{V}_\mathcal{I}} \|\mathbf{v} - \hat{g}\|$. By applying the triangle inequality, the risk gap at $\mathbf{v}$, denoted as $\|h_{\mathcal{T}_\mathcal{I}}(\mathbf{v}) - h_\mathcal{T}(\mathbf{v})\|$, can be decomposed as

$$\begin{aligned}\|h_{\mathcal{T}_\mathcal{I}}(\mathbf{v}) - h_\mathcal{T}(\mathbf{v})\| \le \underbrace{\|h_{\mathcal{T}_\mathcal{I}}(\hat{g}^*) - h_\mathcal{T}(\hat{g}^*)\|}_{\textbf{Part 1}} + \\ \underbrace{\|(h_{\mathcal{T}_\mathcal{I}}(\mathbf{v}) - h_\mathcal{T}(\mathbf{v})) - (h_{\mathcal{T}_\mathcal{I}}(\hat{g}^*) - h_\mathcal{T}(\hat{g}^*))\|}_{\textbf{Part 2}}.\end{aligned} \tag{30}$$

For **Part 1**, we define

$$\epsilon_{appr} = \sup_{\hat{g}\in\mathcal{V}_\mathcal{I}} \|h_{\mathcal{T}_\mathcal{I}}(\hat{g}) - h_\mathcal{T}(\hat{g})\|, \tag{31}$$

which represents the difference between the outputs of $h_{\mathcal{T}_\mathcal{I}}$ and $h_\mathcal{T}$ on the coreset $\mathcal{V}_\mathcal{I}$. It is the optimization objective of $\mathcal{L}_{\text{KD}}$ (Eq. (17)) in Stage 3. Therefore, as $h_{\mathcal{T}_\mathcal{I}}$ converges during training, ϵ_{appr} will be minimized.

For **Part 2**, let's define an interpolation function $\Delta(\mathbf{v}) = h_{\mathcal{T}_\mathcal{I}}(\mathbf{v}) - h_\mathcal{T}(\mathbf{v})$. Then, **Part 2** can be formulated as

$$\textbf{Part 2} = \|\Delta(\mathbf{v}) - \Delta(\hat{g}^*)\|. \tag{32}$$

According to the Mean Value Theorem, there exists a point $\xi \in \mathcal{M}$ on the line connecting $\mathbf{v}$ and $\hat{g}^*$, yielding the following upper bound for **Part 2**:

$$\begin{aligned}\textbf{Part 2} &= \|\Delta(\mathbf{v}) - \Delta(\hat{g}^*)\| = \|\nabla\Delta(\xi)(\mathbf{v} - \hat{g}^*)\| \\ &\le \sup_{\mathbf{v}\in\mathcal{M}} \|\nabla\Delta(\mathbf{v})\| \cdot \|\mathbf{v} - \hat{g}^*\|,\end{aligned} \tag{33}$$

where the gradient norm $\|\nabla\Delta(\mathbf{v})\|$ of the interpolation function $\Delta(\mathbf{v})$ is

$$\begin{aligned}\|\nabla\Delta(\mathbf{v})\| &= \|\nabla h_{\mathcal{T}_\mathcal{I}}(\mathbf{v}) - \nabla h_\mathcal{T}(\mathbf{v})\| \\ &\le \|\nabla h_{\mathcal{T}_\mathcal{I}}(\mathbf{v})\| + \|\nabla h_\mathcal{T}(\mathbf{v})\|.\end{aligned} \tag{34}$$

Since $h_\mathcal{T}$ is trained on the original dataset $\mathcal{T}$, its learned manifold exhibits good smoothness, implying that $\|\nabla h_\mathcal{T}(\mathbf{v})\|$ approximates zero. Furthermore, existing studies [34, 35, 41] report that employing regularization strategies, such as data augmentation, can enhance the smoothness of the model $h_{\mathcal{T}_\mathcal{I}}$ in the sample interpolation region, thereby maintaining $\|\nabla h_{\mathcal{T}_\mathcal{I}}(\mathbf{v})\|$ at a small value. Following these studies, it can be assumed that the interpolation function $\Delta(\cdot)$ satisfies the ϵ_{inter}-Lipschitz continuity condition, indicating the upper bound of Eq. (34) is

$$\sup_{\mathbf{v}\in\mathcal{M}} \|\nabla\Delta(\mathbf{v})\| \le \epsilon_{inter}, \tag{35}$$

where ϵ_{inter} is a small constant. Based on Eqs. (18) and (35), the upper bound of **Part 2** in Eq. (33) becomes

$$\begin{aligned}\textbf{Part 2} &\le \sup_{\mathbf{v}\in\mathcal{M}} \|\nabla\Delta(\mathbf{v})\| \cdot \|\mathbf{v} - \hat{g}^*\| \\ &\le \epsilon_{inter} \cdot d_{fill}(\mathcal{V}_\mathcal{I}, \mathcal{M}).\end{aligned} \tag{36}$$

In summary, by substituting Eqs. (31) and (36) into (30), the upper bound of $\mathcal{R}_{\text{diff}}$ (Eq. (16) in the paper) becomes

$$\mathcal{R}_{\text{diff}} \le \epsilon_{appr} + \epsilon_{inter} \cdot d_{fill}(\mathcal{V}_\mathcal{I}, \mathcal{M}). \tag{37}$$

Since Stage 1 and Stage 2 have ensured that d_{fill} in Eq. (11) decreases to a small value, $\mathcal{R}_{\text{diff}}$ can be minimized. In other words, the performance of $h_{\mathcal{T}_\mathcal{I}}$ can approach that of $h_\mathcal{T}$.

7. Implementation Details of Pattern Marginal Sample Selection

As described in Sec. 3.4 of the paper, at the initial phase of Stage 2, a candidate set $\mathcal{I}^m_{hard}$ containing K samples (Eq. (13)) is constructed by selecting low-confidence samples from pattern margins. However, not all low-confidence samples are beneficial for learning [7]. The candidate set $\mathcal{I}^m_{hard}$ may contain samples with two distinct properties: 1) Valuable hard samples, which possess certain valuable semantics but are hard to identify, corresponding to the marginal patterns; and 2) Harmful noisy samples, which contain misleading information, such as incorrect annotations. The former are crucial for enhancing model generalization, whereas the latter impose a negative impact on model training. To filter valuable samples containing margin patterns from $\mathcal{I}^m_{hard}$, we employ a noisy label learning method. Specifically, for each visual pattern $\mathcal{O}_m$ (Eq. (7)), we assume that the confidence distribution $P_{\mathcal{I}^m_{hard}}$ of its candidate set $\mathcal{I}^m_{hard}$ follows a two-component Gaussian Mixture Model (GMM), denoted as:

$$\begin{aligned}P_{\mathcal{I}^m_{hard}} = w_{val,m}\mathcal{N}(\mu_{val,m}, \sigma^2_{val,m}) + \\ w_{noise,m}\mathcal{N}(\mu_{noise,m}, \sigma^2_{noise,m}),\end{aligned} \tag{38}$$

where w, μ, and σ denote the weight, mean, and standard deviation of the Gaussian components, respectively. The subscripts val and $noise$ correspond to the Gaussian components for valuable samples and noisy samples, respectively. These parameters can be estimated via the Expectation-Maximization (EM) algorithm. According to

[7], noisy samples typically exhibit lower confidence compared to valuable samples, i.e., $\mu_{noise,m} \leq \mu_{val,m}$. Then, based on Eq. (38), the posterior probability $P(\mathcal{N}_{val,m}|\mathbf{x}_i)$ that each sample $\mathbf{x}_i$ in $\mathcal{I}^m_{hard}$ belongs to $\mathcal{N}_{val,m}$ is computed, denoted as

$$P(\mathcal{N}_{val,m}|\mathbf{x}_i) = \frac{w_{val,m}\mathcal{N}(\mathbf{x}_i; \mu_{val,m}, \sigma^2_{val,m})}{P_{\mathcal{I}^m_{hard}}(\mathbf{x}_i)}. \quad (39)$$

Finally, the candidate samples in $\mathcal{I}^m_{hard}$ are sorted according to their posterior probability $P(\mathcal{N}_{val,m}|\mathbf{x}_i)$, and the top N_m samples with the highest probabilities are selected to form the final marginal sample set $\mathcal{I}^{m,filter}_{hard}$. Note that N_m is set equal to the number of samples selected from the pattern center $\mathcal{I}^m_{cen}$ (Eq. (12)) to ensure pattern balance.

8. Visualization Analysis

To qualitatively validate the effectiveness of BPS, we selecte three classes (*Stone Wall*, *Bald Eagle*, and *Amphibious Vehicle*) from ImageNet-1K, and first visualize their modeled multi-pattern distributions alongside some samples selected by BPS that capture class-general and marginal patterns, respectively. Second, we projecte the data distilled by BPS and other SOTAs into a unified reference space $\mathcal{V}_{ref}$, to further visualize their respective degrees of pattern balance.

8.1. Visualization of Multiple Visual Pattern Distribution & Class-general and Marginal Pattern Instances

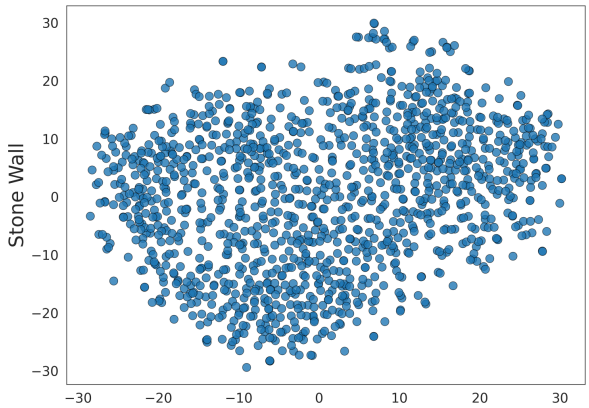
Figure 7 illustrates the distributions of the original dataset modeled by BPS and other SOTAs. Since other methods commonly impose constraints to minimize intra-class distance and maximize inter-class distance, they implicitly enforce a unimodal cluster structure per class, thereby neglecting the intricate diversity within classes. In contrast, by explicitly modeling the multiple visual pattern distribution, BPS captures a multi-pattern structure that better aligns with the nature of real-world data.

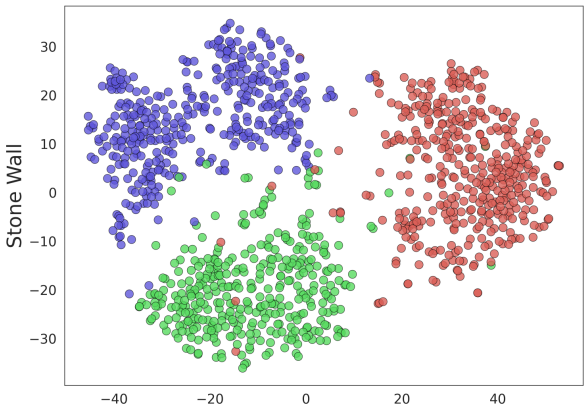
To further intuitively illustrate the visual semantics encoded in these multiple visual patterns, Fig. 8 shows some instances selected from these pattern centers and margins. Our observations reveal two key insights: 1) Even within a single class, the distinct patterns discovered by BPS exhibit notable visual variations, such as different viewpoints or backgrounds. This finding aligns with Axiom 1 in the paper: the distribution of each class in real-world data often follows a multi-pattern structure rather than a single cluster. 2) Samples from pattern centers represent class-general patterns, capturing the most typical features of the class; conversely, samples from pattern margins represent marginal patterns, carrying more complex visual features that are inherently harder to identify. Although marginal patterns are visually more challenging and less frequent, they are critical for enhancing model generalization.

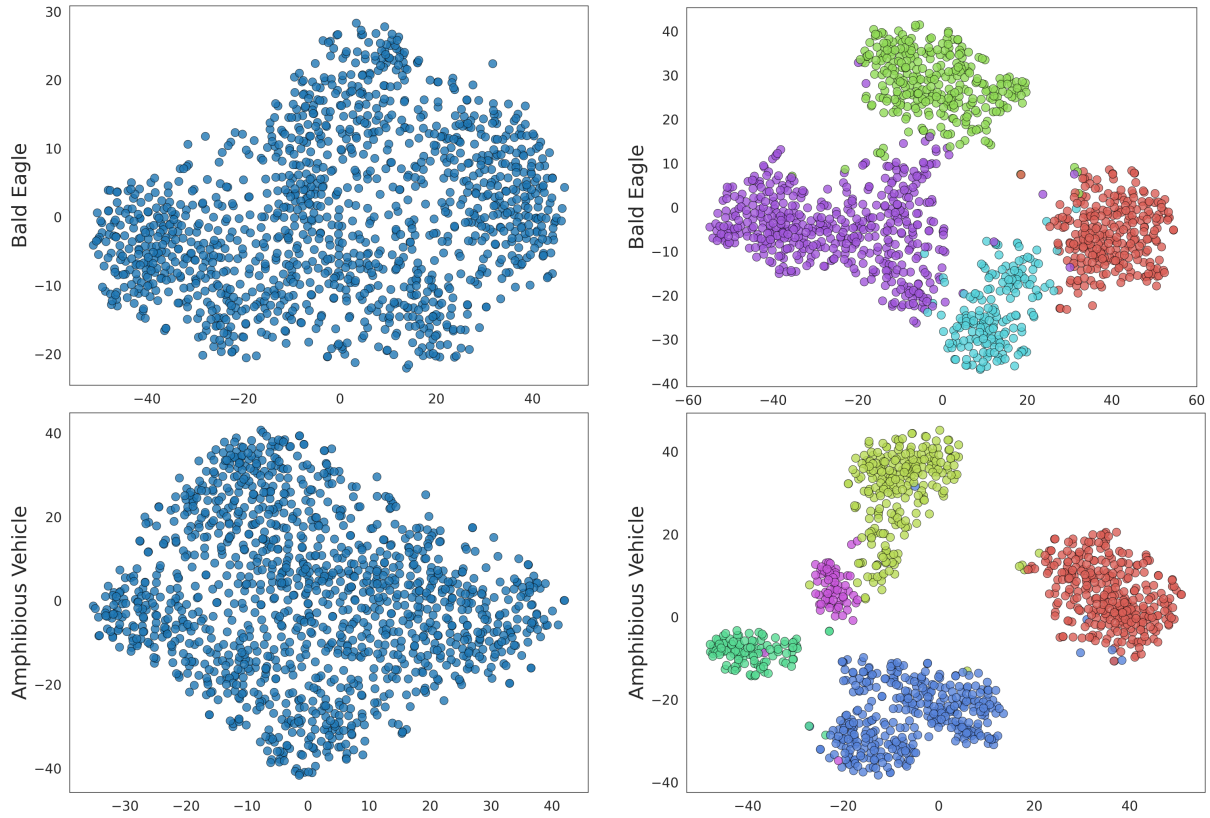
These visualizations underscore the importance of BPS, which explicitly balances patterns based on the modeled multiple visual pattern distribution, thereby offering a novel perspective for understanding dataset natures in the field of dataset distillation.

8.2. Visualization of Pattern Balance

To qualitatively assess the pattern balance of datasets distilled by all methods, we present t-SNE visualization in Fig. 9, comparing BPS with other SOTAs (excluding SRe2L) on ImageNet-1K with IPC $= 50$. For a fair comparison, we adopt the same experimental setup in Sec. 4.2.2 of the paper, projecting all distilled data into a unified, pattern-balanced reference space $\mathcal{V}_{ref}$ derived from the original dataset, rather than the space with multi-pattern structure learned by BPS. It can be observed that the samples selected by BPS are distributed uniformly and sparsely in $\mathcal{V}_{ref}$. This observation empirically validates Proposition 2 in the paper, indicating that BPS effectively covers the information of the original data manifold and suggesting that these samples contain balanced class-general and marginal patterns. In contrast, the distilled data from some SOTAs exhibit dense and localized clustering, reflecting an underlying pattern imbalance.







(a) The feature distribution modeled by other methods.

(b) The multi-pattern distribution modeled by BPS.

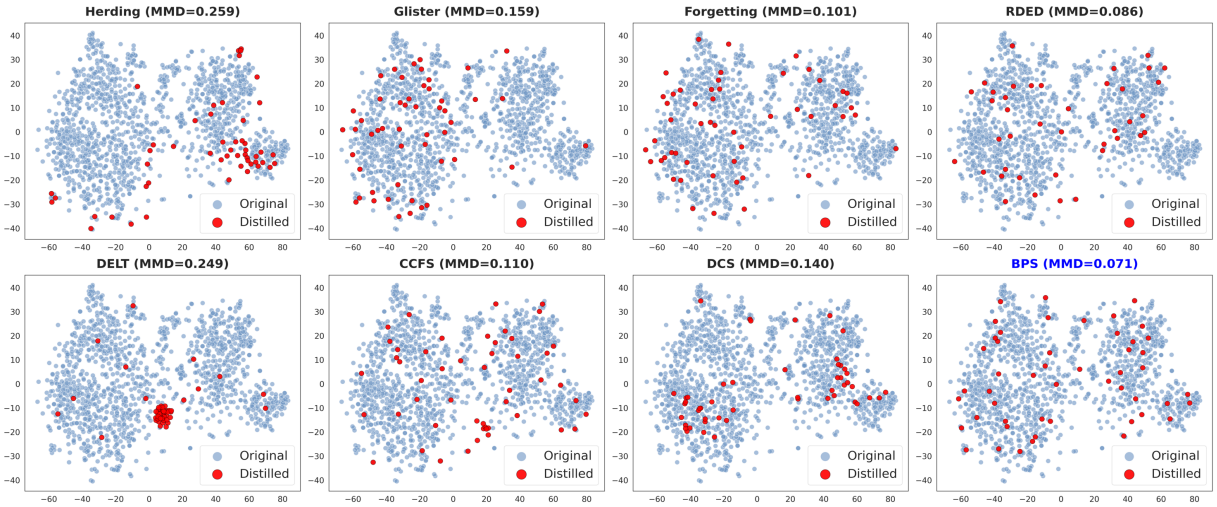
Figure 7. T-SNE visualization of feature distributions of three classes (*Stone Wall*, *Bald Eagle* and *Amphibious Vehicle*) in ImageNet-1K. (a) The single cluster distribution of a class modeled by other SOTA methods. (b) The multi-pattern structure modeled by BPS, where different colors denote distinct patterns within one class.

(a) The class of *Stone Wall*

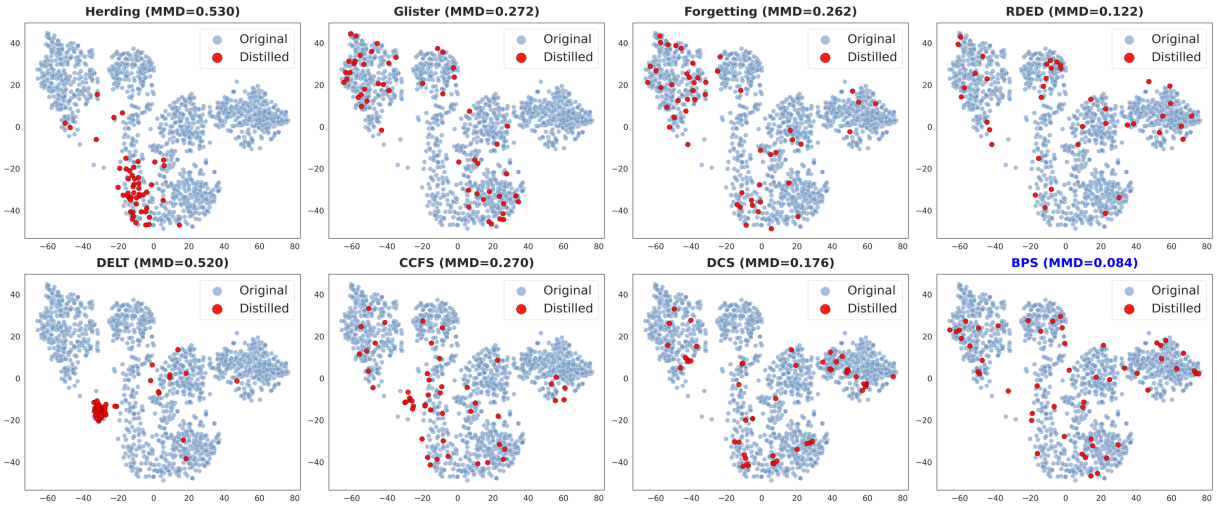
(b) The class of *Bald Eagle*

(c) The class of *Amphibious Vehicle*

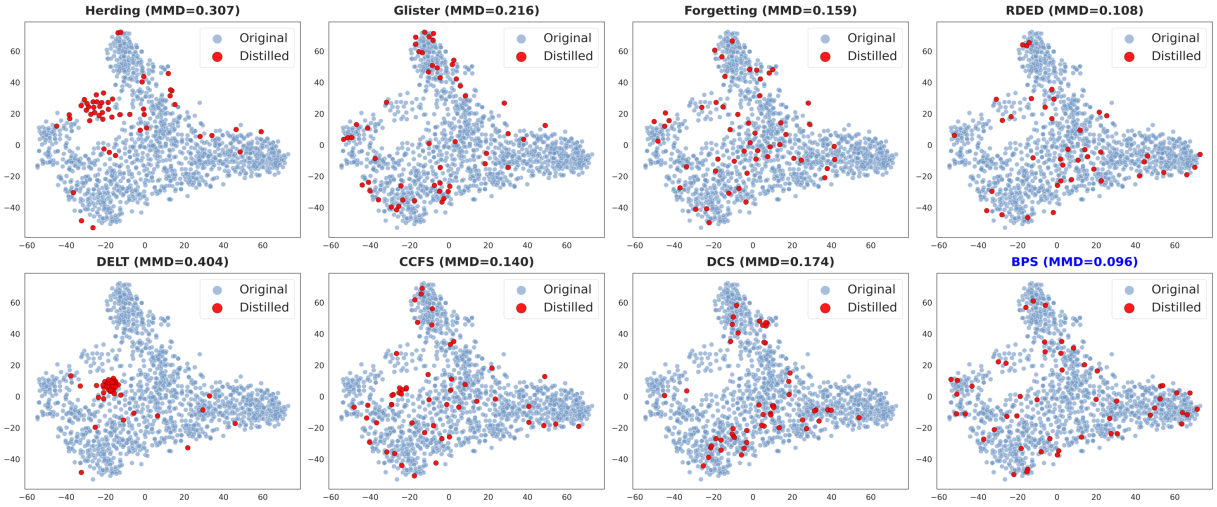
Figure 8. The instances capturing the class-general (yellow boxes) and marginal (blue boxes) patterns discovered by BPS in ImageNet-1K. Each row represents a visual pattern.



(a) The class of *Stone Wall*



(b) The class of *Bald Eagle*



(c) The class of *Amphibious Vehicle*

Figure 9. T-SNE visualization of distilled data (marked by red) and original data (marked by blue) on ImageNet-1K with IPC $= 50$. All data are projected onto a unified and pattern-balanced reference space $\mathcal{V}_{ref}$.