PEFTDiff: Diffusion-Guided Transferability Estimation for Parameter-Efficient Fine-Tuning

Supplementary Material

The supplementary material contains additional details and further results related to the main paper.

1. Visualization of diffusion-based technique on a certain dataset

To qualitatively assess the impact of the proposed diffusion process on feature representations, t-SNE visualizations are provided on the DTD dataset, comparing original features (left) and diffused features (right). Figure 1 illustrates the effect on two PEFT methods: Convpass Attention and LoRA.

For Convpass Attention, the diffusion process enhances inter-class separation and intra-class compactness, resulting in more distinct clusters, as reflected by an increase in Silhouette Score from 0.0189 to 0.0195. This indicates that diffusion refines the feature space to form more discriminative representations, which aligns with Convpass Attention's superior classification performance. In contrast, for LoRA, diffusion slightly disrupts the feature structure, reducing the Silhouette Score from 0.0188 to 0.0148. These observations suggest that the impact of diffusion depends on the initial geometry of the PEFT embeddings, benefiting methods with well-aligned feature spaces while offering limited gains for others.

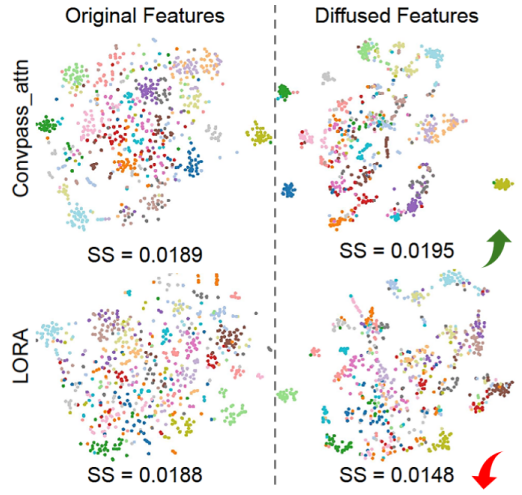

Figure 1. t-SNE visualization of original and diffused feature embeddings on the DTD dataset for Convpass Attention and LoRA.

2. More experimental results

This section presents experimental results that were not included in the main paper due to space limitations.

2.1. Comparison with Clustering Methods

The proposed method is further compared against DeepCluster and three traditional clustering-based approaches: DBSCAN, K-Means, and Agglomerative Clustering. As shown in Table 1, the proposed method consistently outperforms these baselines. For DeepCluster, its original evaluation protocol is followed to assess the quality of feature clustering, while for DBSCAN, K-Means, and Agglomerative Clustering, performance is evaluated based on clustering accuracy computed via assignment correctness.

Table 1. Comparison of clustering methods on VTAB-1k benchmark.

Metric	DeepCl.	DBSCAN	Kmeans	Agglomerative	Ours
Avg. (19)	0.050	-0.043	0.420	0.382	**0.517**

2.2. Effect of Dimensionality

The number of top c eigenvectors determines how much structural information is retained in the transformed diffusion space, directly impacting the accuracy of diffusion distance computation (Eq. 4). Selecting too few eigenvectors discards important manifold structure, whereas too many can introduce noise, reducing correlation performance. To isolate the effect of c, we fix the optimal values of k and t and vary c. We evaluate correlation performance for $c \in \{8, 16, 32, 64\}$, as shown in Figure 2. As c increases, correlation improves, peaking at $c = 32$ ($\tau_w = 0.517$). Beyond this point, adding more components does not enhance performance, indicating that an optimal balance must be maintained between preserving structural information and mitigating noise.

3. Normalization Formula

To ensure that the intra-class and inter-class diffusion scores are on a comparable scale, we apply normalization. Since a higher inter-class diffusion score S_t^{inter} is desirable, we use standard min-max normalization:

$$\hat{S}_t^{\text{inter}} = \frac{S_t^{\text{inter}} - \min(S_t^{\text{inter}})}{\max(S_t^{\text{inter}}) - \min(S_t^{\text{inter}})}. \quad (1)$$

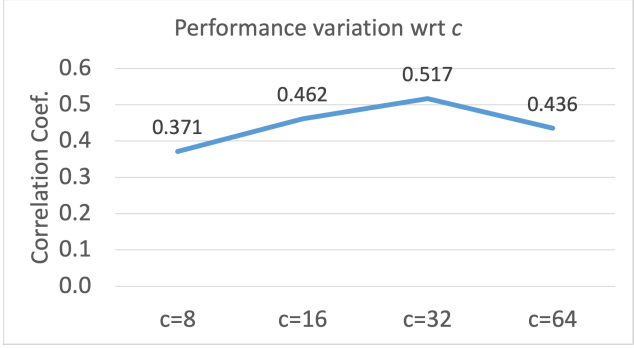

Figure 2. Effect of the number of components c on correlation performance. Correlation peaks at $c = 32$.

Conversely, for intra-class diffusion, a lower score is preferred, as it indicates tighter clustering within each class. To ensure that lower values contribute positively to the final ranking, we apply inverse normalization:

$$\hat{S}_t^{\text{intra}} = 1 - \frac{S_t^{\text{intra}} - \min(S_t^{\text{intra}})}{\max(S_t^{\text{intra}}) - \min(S_t^{\text{intra}})}. \tag{2}$$

This transformation ensures that both intra-class and inter-class diffusion scores align with our ranking objective. After normalization, higher values indicate better model performance, making the final PEFT selection score comparable across different models.

4. Relation between diffusion distance and euclidean distance [5]

If we choose the diffusion coordinates as:

$$\Psi_t(x) = (\lambda_1^t \phi_1(x), \lambda_2^t \phi_2(x), \ldots, \lambda_d^t \phi_d(x)),$$

then the diffusion distance between points in the original space is equal to the Euclidean distance in the diffusion space:

$$D_t^2(x_i, x_j) = \|\Psi_t(x_i) - \Psi_t(x_j)\|_2^2.$$

Proof:

The diffusion distance between two points x_i and x_j at time t is given by:

$$D_t^2(x_i, x_j) = \sum_{u \in X} |p_t(x_i, u) - p_t(x_j, u)|^2, \tag{3}$$

where $p_t(x, u)$ is the probability of transitioning from x to u in t steps.

Using the spectral decomposition of the transition matrix P, the probability distribution can be written as:

$$p_t(x, u) = \sum_{l \geq 0} \lambda_l^t \phi_l(x) \phi_l(u). \tag{4}$$

Substituting this into the diffusion distance:

$$D_t^2(x_i, x_j) = \sum_{u \in X} \left| \sum_{l \geq 0} \lambda_l^t \phi_l(x_i) \phi_l(u) - \sum_{l \geq 0} \lambda_l^t \phi_l(x_j) \phi_l(u) \right|^2. \tag{5}$$

Rearranging the terms:

$$D_t^2(x_i, x_j) = \sum_{u \in X} \left| \sum_{l \geq 0} \lambda_l^t \phi_l(u) \left(\phi_l(x_i) - \phi_l(x_j) \right) \right|^2. \tag{6}$$

Expanding the squared term:

$$D_t^2(x_i, x_j) = \sum_{u \in X} \left(\sum_{l \geq 0} \lambda_l^t \phi_l(u) \left(\phi_l(x_i) - \phi_l(x_j) \right) \right) \times \left(\sum_{m \geq 0} \lambda_m^t \phi_m(u) \left(\phi_m(x_i) - \phi_m(x_j) \right) \right). \tag{7}$$

$$D_t^2(x_i, x_j) = \sum_{u \in X} \sum_{l \geq 0} \sum_{m \geq 0} \lambda_l^t \phi_l(u) (\phi_l(x_i) - \phi_l(x_j)) \times \lambda_m^t \phi_m(u) (\phi_m(x_i) - \phi_m(x_j)). \tag{8}$$

$$D_t^2(x_i, x_j) = \sum_{u \in X} \sum_{l \geq 0} \sum_{m \geq 0} \lambda_l^t \phi_l(u) \phi_m(u) \lambda_m^t \times (\phi_l(x_i) - \phi_l(x_j))(\phi_m(x_i) - \phi_m(x_j)). \tag{9}$$

Since ϕ_l is an orthonormal basis, we use the property:

$$\sum_{u \in X} \phi_l(u) \phi_m(u) = \delta_{lm} = \begin{cases} 1, & l = m, \\ 0, & l \neq m. \end{cases} \tag{10}$$

This eliminates cross terms where $l \neq m$, simplifying our expression:

$$D_t^2(x_i, x_j) = \sum_{l \geq 0} \lambda_l^{2t} (\phi_l(x_i) - \phi_l(x_j))^2 \sum_{u \in X} \phi_l^2(u). \tag{11}$$

As provided in reference, the diffusion coordinates form an orthonormal basis. By the normalization property of eigenfunctions:

$$\sum_{u \in X} \phi_l^2(u) = 1. \tag{12}$$

Thus, we obtain:

$$D_t^2(x_i, x_j) = \sum_{l \geq 0} \lambda_l^{2t} (\phi_l(x_i) - \phi_l(x_j))^2. \quad (13)$$

Thus, we have proved that the diffusion distance in the original space is equal to the Euclidean distance in the diffusion space when the diffusion coordinates are chosen as:

$$\Psi_t(x) = (\lambda_1^t \phi_1(x), \lambda_2^t \phi_2(x), \ldots, \lambda_d^t \phi_d(x)). \quad (14)$$

This confirms that diffusion maps provide a transformation where the diffusion distance in the original space becomes equivalent to the Euclidean distance in the new diffusion space, ensuring that the geometry of the data is well preserved.

5. Algorithm and Time Complexity

Although the class count (C) introduces an $O(C^2)$ complexity term, the overall runtime is predominantly influenced by dataset size (N). Specifically, operations such as similarity matrix computation ($O(N \cdot k \cdot \log N)$) and spectral decomposition ($O(N^2)$) dominate the runtime. Therefore, we capped our experiments at 10,000 samples. Below, we provide a pseudocode representation (Alg. 1) of our approach to better illustrate the step-by-step process involved in selecting the most suitable PEFT technique.

6. Ground Truth

To establish an accurate ground-truth ranking of PEFT techniques, we fine-tune each model on the target dataset, following prior methodologies [33,37]. Earlier approaches [14,16,18] initialized PEFT techniques randomly; however, recent findings [34] suggest that initializing with ImageNet weights significantly improves fine-tuning performance. Ranking PEFT techniques using randomly initialized models is infeasible, as the extracted target embeddings would be highly noisy and lack meaningful structure. Since embeddings are obtained by performing inference with different PEFT methods, a randomly initialized PEFT produces feature representations that contain no task-specific information. Consequently, rankings based on these noisy embeddings would be unreliable, as variations would arise from random initialization rather than the actual fine-tuning effectiveness of each method.

To ensure meaningful feature extraction and stable rankings, we initialize all PEFT techniques with pre-trained ImageNet weights. Specifically, each PEFT method is pre-trained on a subset of ImageNet containing 30 samples per class before fine-tuning on the target dataset. Fine-tuning is performed using a grid search over key hyperparameters, as learning rate and weight decay significantly influence downstream performance. We explore learning rates of $\{10^{-2}, 10^{-3}, 10^{-4}\}$ and weight decay values of $\{10^{-3}, 10^{-4}, 10^{-5}\}$. The optimal hyperparameter configuration is selected based on validation performance. All models are fine-tuned on 8 NVIDIA A100 GPUs with a batch size of 128, and input images are resized to 224×224 pixels.

Algorithm 1: Diffusion-based PEFT Selection method

Input: Target dataset $D = \{Z, Y\}$ with samples Z and labels Y; Set of L PEFT models $\{f_l\}_{l=1}^{L}$; Diffusion step parameter t; Number of nearest neighbors k.

Output: PEFT selection scores S_{PEFT} for each PEFT.

for *each model* f_l *in* $\{f_l\}_{l=1}^{L}$ **do**
- Extract feature representations: $X_l = f_l(Z)$;
- Normalize features: $X_l \leftarrow X_l / \|X_l\|_2$;
- `// Compute Pairwise Similarities`
- Initialize affinity matrix $K \in \mathbb{R}^{N \times N}$
- **for** *each pair* (x_i, x_j) *in* X_l **do**
  - $K_{ij} = \exp(-\|x_i - x_j\|^2 / \sigma^2)$ `// Gaussian RBF kernel`
- `// Construct k-NN Graph`
- Apply k-NN filtering to retain only k-nearest neighbors; Compute degree matrix D where $D_{ii} = \sum_j K_{ij}$; Compute transition matrix $P = D^{-1}K$, ensuring P is row-stochastic;
- `// Compute Multi-Step Transition Probabilities`
- Compute $P^t = P \cdot P \cdots P$ (for t steps);
- `// Compute Diffusion Distances`
- Initialize diffusion distance matrix $D_t \in \mathbb{R}^{N \times N}$
- **for** *each pair* (x_i, x_j) *in* X_l **do**
  - $D_t^2(x_i, x_j) = \sum_{u=1}^{N} (P_t[i,u] - P_t[j,u])^2$;
- `// Compute Intra-Class and Inter-Class Scores`
- Compute intra-class diffusion score: S_t^{intra}
- Compute inter-class diffusion score: S_t^{inter}
- `// Compute PEFT Selection Score`
- Compute $\Delta S_t^{\text{intra}} = S_t^{\text{intra}}(P_{\text{PEFT}_i}) - S_t^{\text{intra}}(P_{\text{backbone}})$
- Compute $\Delta S_t^{\text{inter}} = S_t^{\text{inter}}(P_{\text{PEFT}_i}) - S_t^{\text{inter}}(P_{\text{backbone}})$;
- `// Compute final selection score after normalization`
- $$S_{\text{PEFT}_i} = \hat{\Delta S_t^{\text{inter}}} + \hat{\Delta S_t^{\text{intra}}}$$

return S_{PEFT};

Table 2. VTAB-1k Ground Truth Ranking: Top-1 classification accuracy (%) of different PEFT techniques across VTAB-1k datasets. This ranking is used to compute Kendall's correlation coefficient with the predicted ranking.

Dataset	ADAPTER	LORA	Convpass	ConvpassAttn	FactTT	FactTK	VPT	BitFit	NOAH
caltech101	88.78	88.82	90.91	91.77	91.30	91.39	92.24	86.88	91.01
cifar100	72.39	69.94	72.65	68.58	72.08	72.21	73.92	72.83	72.49
dtd	69.82	69.16	73.65	73.15	73.11	72.00	73.78	67.48	73.42
oxford_flowers102	97.62	97.54	98.17	98.79	99.55	98.42	99.42	97.81	97.97
oxford_iiit_pet	91.11	90.34	90.94	91.18	91.34	91.45	91.97	89.12	90.80
svhn	91.01	90.67	90.42	92.83	89.47	89.91	91.25	92.50	89.32
sun397	53.20	53.13	52.94	53.43	53.28	53.20	54.12	51.27	52.06
patch_camelyon	85.77	88.98	88.39	88.13	87.78	88.59	88.23	85.87	87.23
eurosat	96.93	97.91	97.22	98.29	97.21	97.43	97.13	97.82	97.05
resisc45	84.03	84.67	86.61	85.82	85.42	85.63	85.89	82.43	85.37
diabetic_retinopathy	74.60	73.78	75.49	74.44	73.25	73.86	74.00	74.19	73.14
clevr_count	82.48	82.17	83.17	82.09	82.96	82.20	82.91	80.37	82.83
clevr_dist	63.90	64.31	66.19	65.48	65.61	65.94	66.08	63.73	63.22
dmlab	50.97	50.62	51.93	51.88	51.07	52.39	49.14	49.91	50.26
kitti	78.10	78.93	81.86	78.90	77.19	76.09	79.80	75.11	79.06
dsprites_loc	82.68	82.01	86.72	84.29	86.35	86.19	82.41	81.46	86.69
dsprites_ori	54.48	54.30	54.13	53.73	53.25	53.18	54.63	51.58	53.84
smallnorb_azi	35.16	37.18	36.49	36.48	36.84	38.30	34.33	37.11	34.91
smallnorb_ele	43.16	43.08	45.79	43.18	42.55	43.08	42.88	37.47	42.81

Table 3. FGVC Ground Truth Ranking: Top-1 classification accuracy (%) of different PEFT techniques across FGVC datasets. This ranking is used to compute Kendall's correlation coefficient with the predicted ranking.

Dataset	ADAPTER	LORA	Convpass	ConvpassAttn	FactTT	FactTK	VPT	BitFit	NOAH
CUB_200_2011	88.81	89.65	88.99	88.76	87.37	87.12	89.57	88.51	89.59
NABirds	86.11	86.95	84.21	86.94	84.59	83.47	85.16	84.19	86.82
OxfordFlower	98.42	99.19	98.13	99.21	99.05	98.73	99.11	98.91	99.06
StanfordCars	86.21	84.39	85.61	85.19	85.48	85.92	84.77	84.04	85.14
StanfordDogs	91.92	91.04	92.82	93.12	91.01	92.49	93.48	92.81	93.22

Table 2 and Table 3 present the Top-1 classification accuracy of 9 PEFT methods across 19 VTAB-1k datasets and 5 FGVC datasets, respectively. These accuracy scores serve as ground-truth rankings for evaluating the correlation with our predicted rankings using Kendall's τ_w coefficient.