Nonparametric Deep Fine-grained Clustering with Low-Rank Guided Vision-Language Model

Xulun Ye[1*] Benyu Wu[1*] Jie Hong[1] Kun Zhou[2†]
[1]Ningbo University, [2]Shenzhen University

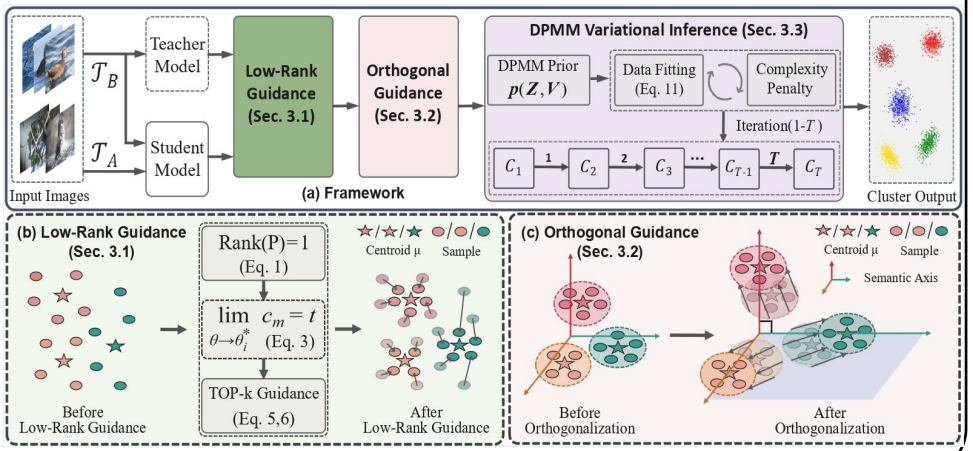
CVPR JUNE 3-7, 2026

Introduction

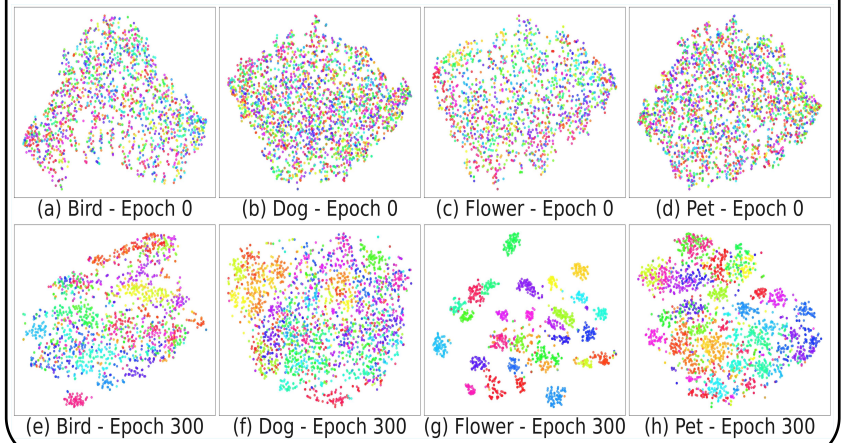
Deep clustering has achieved significant success, with recent large model-assisted approaches representing a paradigm shift by leveraging rich prior knowledge. However, these methods face major hurdles in fine-grained tasks. Generalist large models are pre-trained on coarse data with high inter-class and low intra-class variance, failing to capture subtle semantic differences under the opposite properties of fine-grained data. Additionally, existing fine-tuned models rely on unavailable supervised labels and require a pre-specified cluster number, which is impractical for real-world exploration. To address these dual challenges, this paper introduces a novel framework that dynamically infers categories through a unified Bayesian learning procedure. First, a low-rank guided Vision-Language Model transforms optimization into a top-k selection task to narrow intra-class variance. Second, perturbation-enhanced instance contrast and global orthogonality guidance amplify inter-class separation among cluster centers. Finally, integrating these with the Dirichlet process allows the framework to jointly infer cluster assignments and category numbers, significantly outperforming state-of-the-art approaches.

Methods and Materials

- Low-Rank Guided VLM
- A top-k selection task guides student predictions toward sparse targets, achieving tight intra-cluster compactness. Orthogonal and Perturbation Guidance
- Noise perturbation and center orthogonality constraints amplify global inter-class separation. DPMM Variational Inference
- A Dirichlet process dynamically infers optimal category numbers while simultaneously estimating cluster assignments. y.



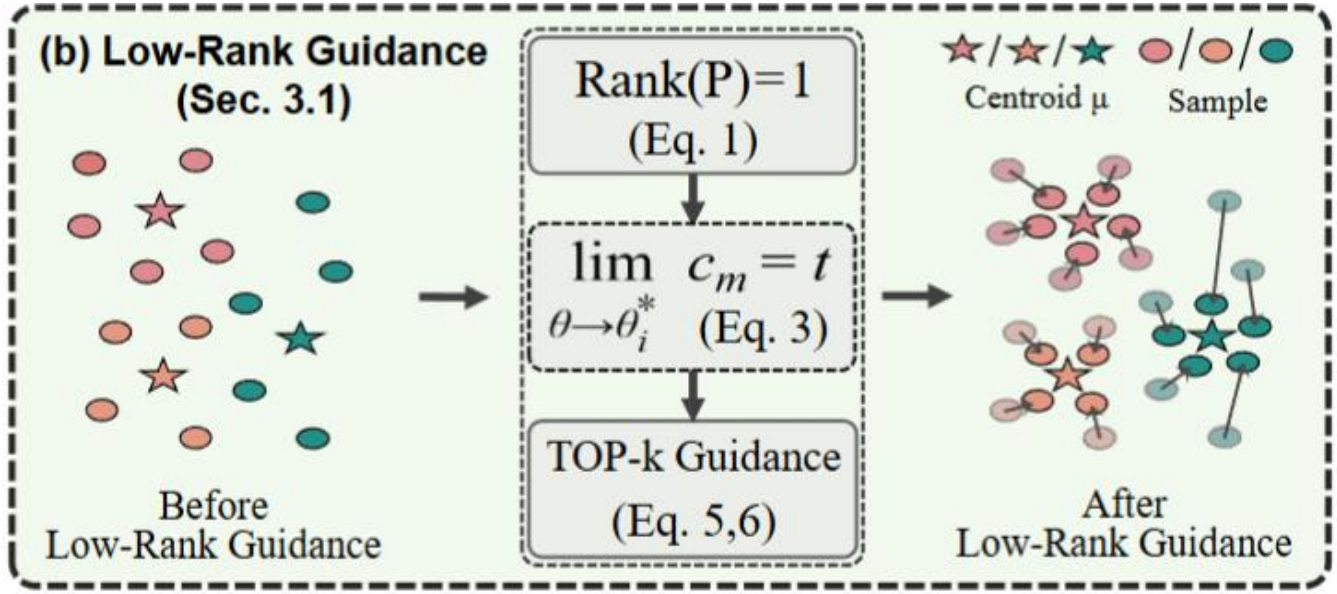
Experiments & Visualization Results

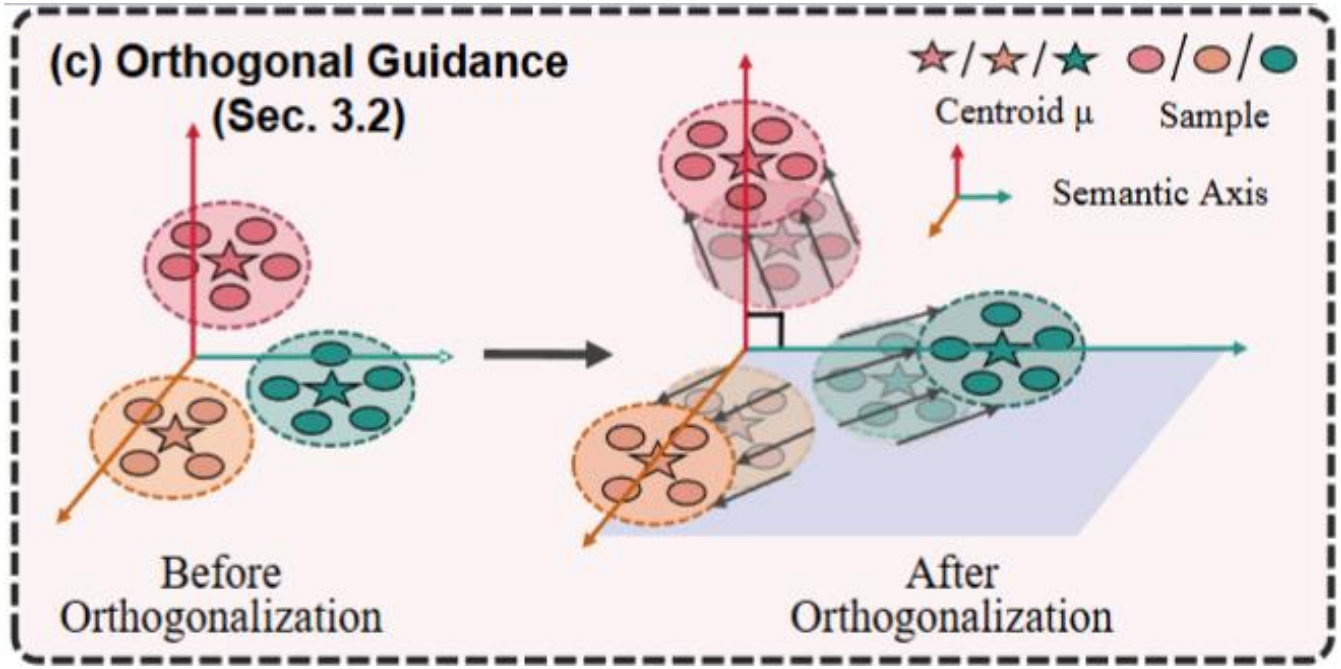


Contributions & Conclusions

A VLM-driven framework using low-rank guidance for unsupervised fine-grained clustering Top-k selection task implements a differentiable proxy for intra-cluster compactness Perturbation and global center orthogonality loss amplify inter-cluster feature separability Dirichlet Process Mixture Model dynamically infers optimal cluster numbers via variational inference State-of-the-art performance across multiple standard datasets in realistic unknown cluster scenarios

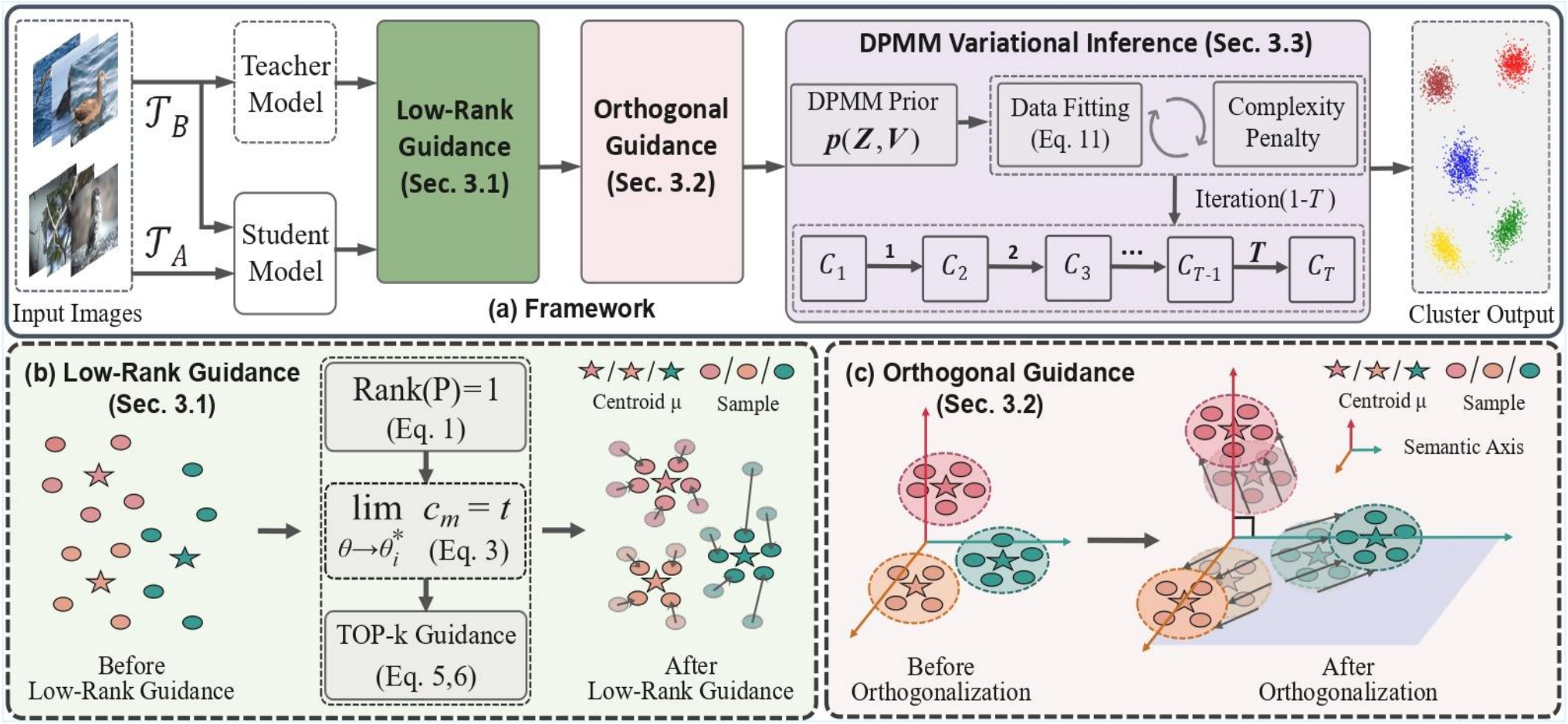
Introduction





- Deep clustering forms a cornerstone of unsupervised end-to-end representation learning
- Large model-assisted clustering leverages rich priors but struggles with fine-grained tasks
- Coarse-grained pre-training fails to capture subtle fine-grained intra-class variations
- Existing approaches rely on supervised labels and impractical pre-specified cluster numbers

Methodology



Three core modules

- Low-Rank Guidance
- Orthogonal Guidance
- DPMM Variational Inference

Experiments:Main Results

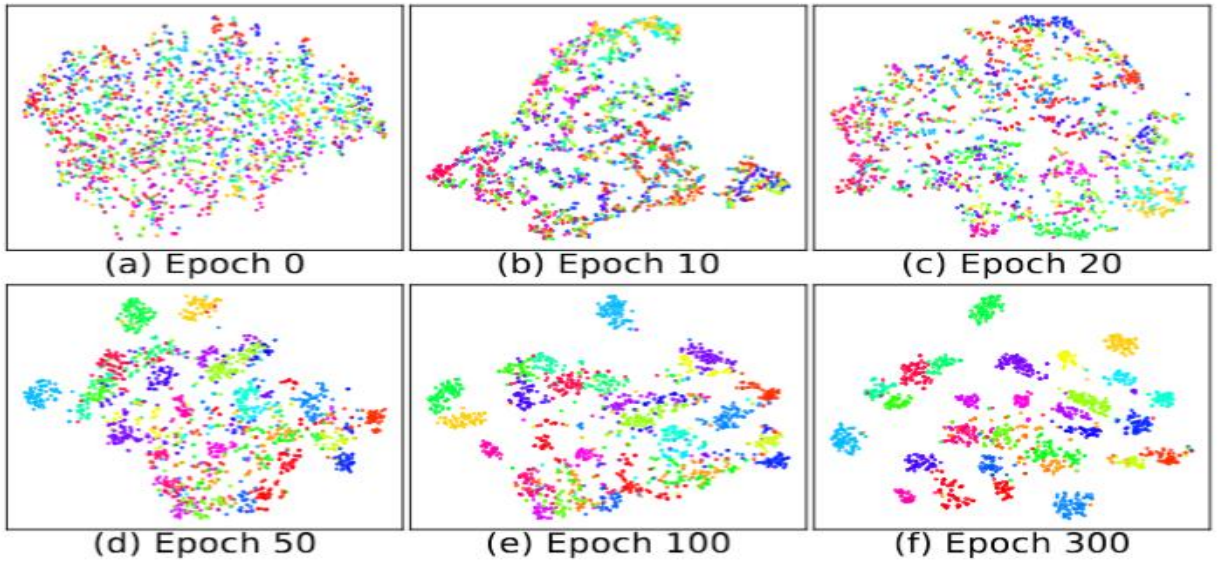
Table 1. The accuracy of NMI and ACC is compared on the CUB-200-2011, Stanford Dogs, Oxford Flower, and Oxford-IIIT Pet datasets. The best results are highlighted in **bold** and the second-best results are underlined. †Indicates guidance from the vision language model.

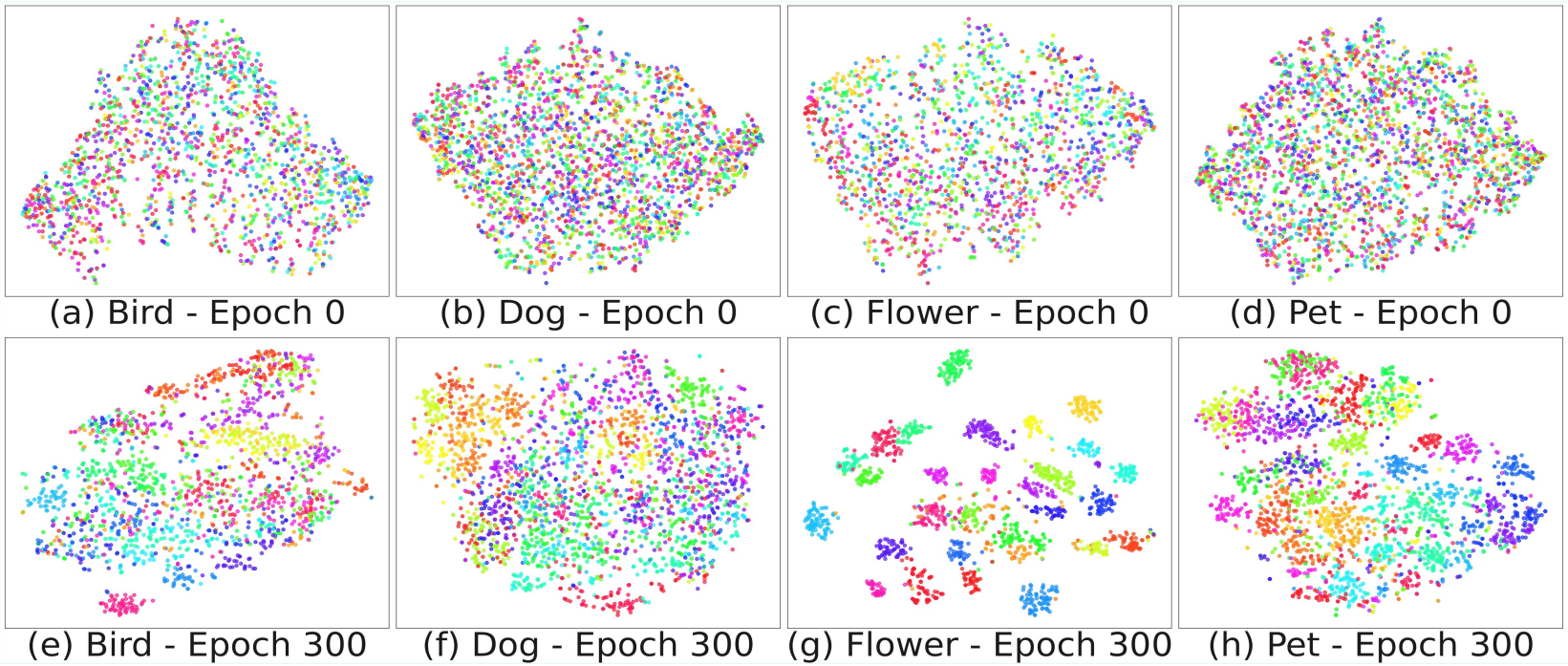
Methods	Backbone	CUB-200-2011		Stanford Dogs		Oxford Flower		Oxford-IIIT Pet	
		NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC
IIC (CVPR'19) [22]	Resnet-34	36.0	7.4	18.0	5.0	24.0	8.7	23.0	9.6
FineGAN (CVPR'19) [45]	-	37.0	6.9	22.0	6.0	24.0	8.1	25.0	8.9
MixNMatch (CVPR'20) [27]	-	41.0	10.2	30.0	10.3	57.0	39.0	56.0	42.3
SimCLR (ICML'20) [9]	Resnet-50	40.0	8.4	19.0	6.8	29.0	12.5	30.0	13.8
SCAN (ECCV'20) [47]	Resnet-50	45.0	11.9	35.0	12.3	77.0	56.5	76.0	60.4
CC (AAAI'21) [28]	Resnet-18	25.1	5.2	22.2	7.6	56.5	27.7	55.8	28.7
C3-GAN (ICLR'22) [25]	-	53.0	27.6	36.0	17.9	67.0	67.8	57.0	51.2
DivClust (CVPR'23) [33]	Resnet-50	39.7	10.8	27.6	9.5	59.0	30.3	58.6	33.5
IDC (NeurIPS'24) [31]	Resnet-50	29.3	17.8	30.5	18.1	43.3	27.5	46.2	28.7
TEMI (BMVC'23)† [1]	ViT	60.9	24.4	43.5	19.2	66.9	39.8	62.6	36.6
TAC (ICML'24)† [29]	ViT	64.6	<u>34.7</u>	<u>64.8</u>	<u>48.7</u>	73.2	53.2	76.5	61.3
CLUDI (ICML'25)† [46]	ViT	61.3	30.2	63.7	42.6	81.5	<u>69.7</u>	<u>87.3</u>	74.1
Ours+Resnet	Resnet-50	<u>65.1</u>	33.2	63.3	40.3	<u>84.7</u>	65.5	82.8	<u>77.7</u>
Ours+CLIP	ViT	**70.9**	**41.8**	**69.1**	**53.2**	**88.4**	**72.6**	**88.0**	**82.2**

Table 2. Comparison of the cluster number inferred by our framework versus the Ground Truth number of classes.

	CUB-200-2011	Stanford Dogs	Oxford Flower	Oxford-IIIT Pet	ImageNet50	ImageNet100	ImageNet200
Predicted cluster number	210.4	127.1	104.6	39.3	57.8	112.4	221.7
Ground Truth	200	120	102	37	50	100	200

Experiments:Main Results





Conclusion

1. *Proposed a novel framework for fine-grained clustering*
2. *Reframed low-rank optimization into top-k selection*
3. *Enhanced inter-class separability via feature perturbation*
4. *Integrated Dirichlet process for dynamic cluster inference*
5. *Achieved state-of-the-art performance on multiple benchmarks*