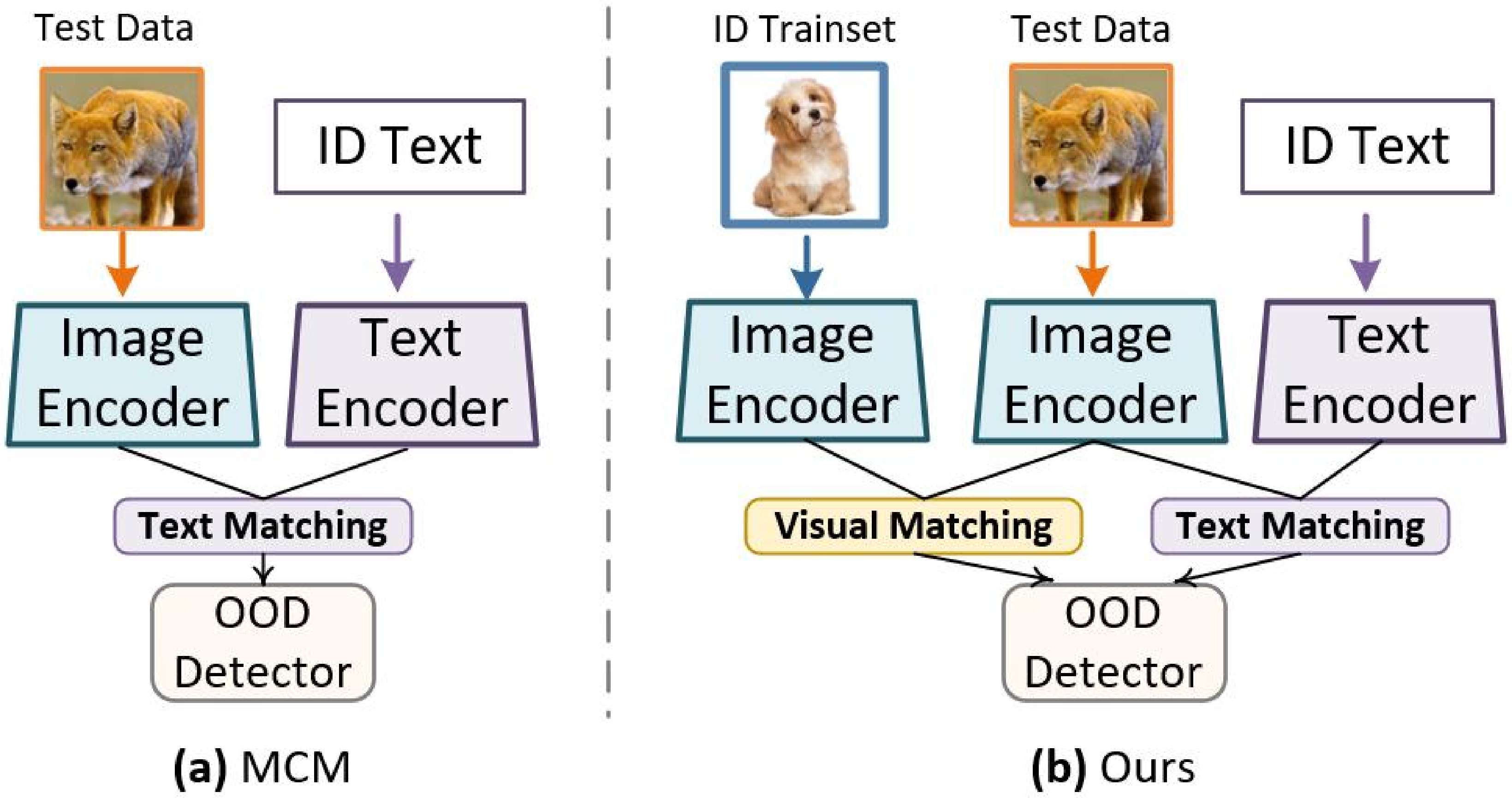
Vision-Language Dual-Pattern Matching for Out-of-Distribution Detection

Zihan Zhang*, Zhuo Xu*, and Xiang Xiang*(✉)

National Key Lab of Multi-Spectral Information Intelligent Processing Technology,
School of Artificial Intelligence and Automation,
Huazhong University of Science and Technology, Wuhan, China
xex@hust.edu.cn

Motivation



- Previous work has not fully exploited the information from the image modality, relying only on text matching.
- Efficient fine-tuning strategies for visual language models on out-of-distribution (OOD) detection have not been adequately explored in prior research.

Existing Methods

Methods without OOD data

Unimodal OOD Detection

- **Training-Free:** MSP,Maxlogit,ODIN,Energy
- **Training-Required:** MOS,VOS,LogitNorm,DML

Crossmodal OOD Detection

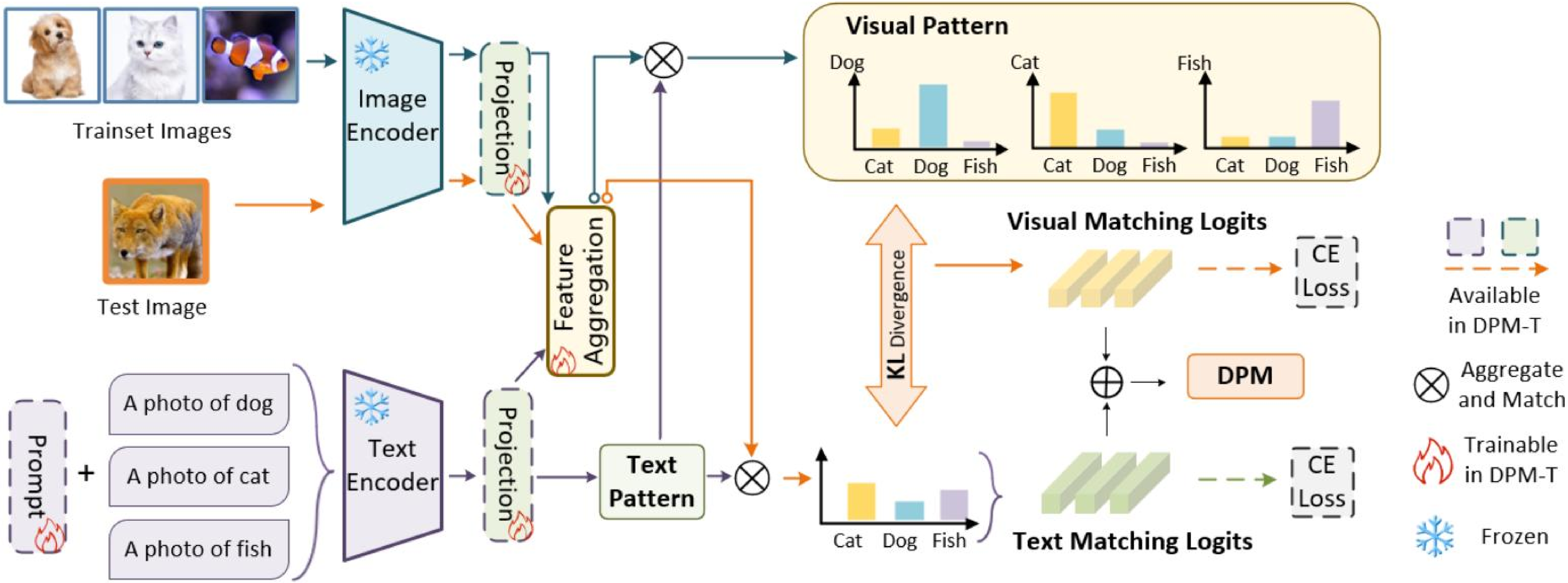
- **Training-Free:** MCM,GLMCM
- **Training-Required:** ZOC,CLIPN,LoCoOp

Adapting CLIP to downstream tasks

- **Training-free:** CALIP,Tip-Adapter
- **Prompt-based:** CoOp,Dual-CoOp,VPT
- **Adapter-based:** Clip-adapter,MaPLe,Tip-adapter(F)

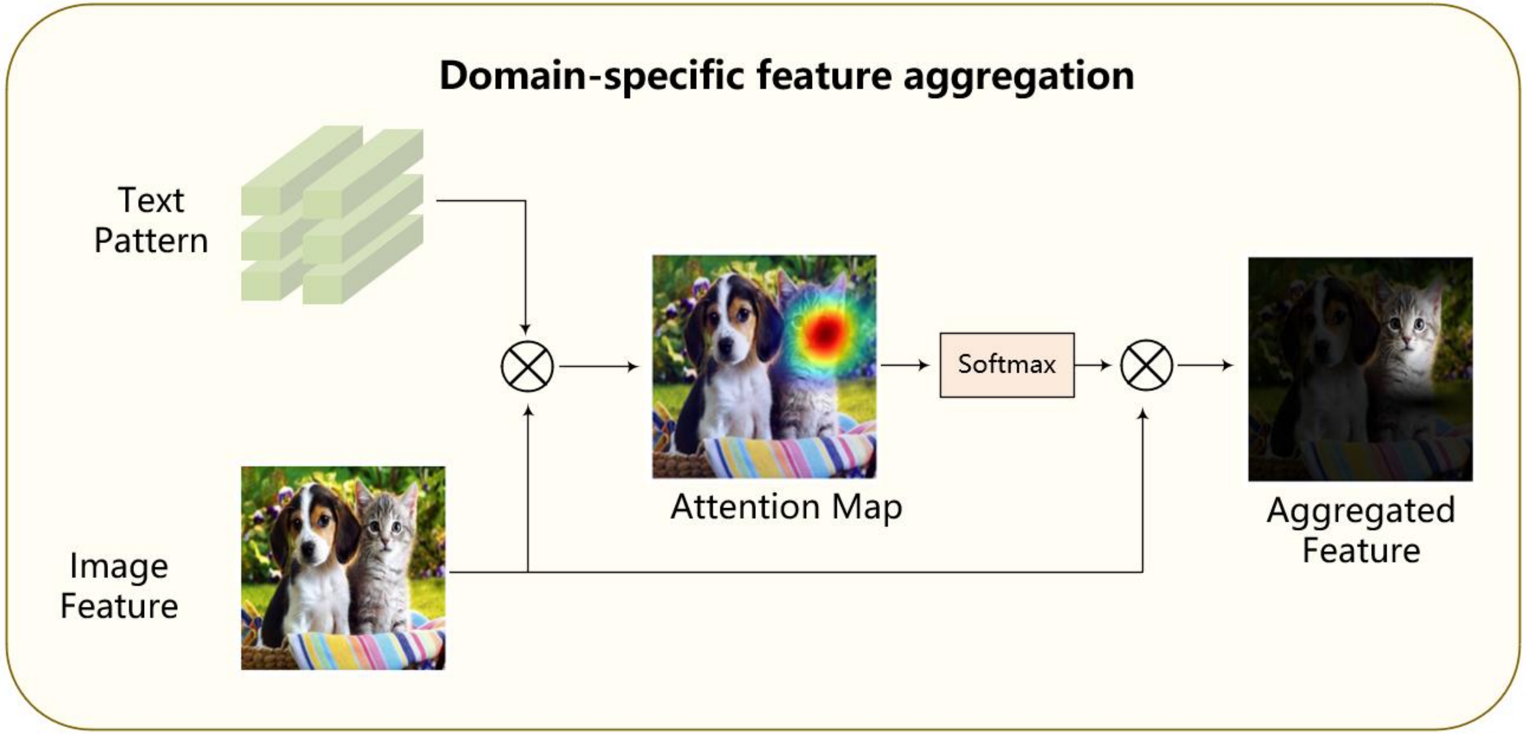
Our Solution

Compared to other methods that only textual prototypes, DPM introduces a visual prototype, combining the outputs from both modalities for a more comprehensive representation.



Our Solution

- CLIP contains a significant amount of redundant information in its features. Many channels are irrelevant.
- We select the rigion visual feature that are most aligned with the textual features.



org:

$$F_k^T = \Phi^T(prompt(y_k)), F^V = \text{Pooling}(f^V),$$

$$p(y_k \mid \mathbf{x}) = \frac{\exp\left(\cos\left(F^V, F_k^T\right)/\tau\right)}{\sum_{j=1}^{K} \exp\left(\cos\left(F^V, F_j^T\right)/\tau\right)}.$$

ours:

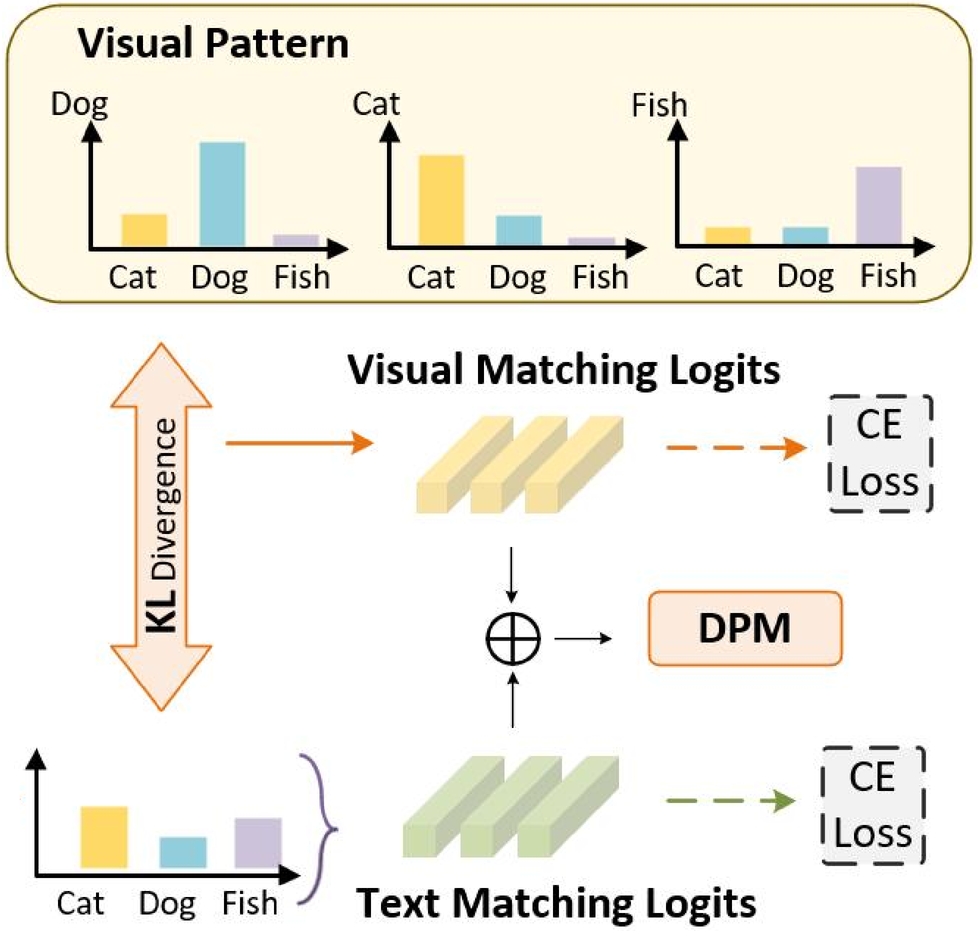
$$\bar{F^V} = \sum_{r=1}^{H\cdot W} \frac{\exp(\cos(F_r^V, F_k^T))}{\sum_{j=1}^{H\cdot W} \exp(\cos(F_j^V, F_k^T))} F_r^V \in \mathbb{R}^{1\times C}.$$

$$\tilde{F^V} = \gamma \bar{F^V} + F^V \in \mathbb{R}^{1\times C}.$$

$$\tilde{p}(y_k \mid \mathbf{x}) = \frac{\exp\left(\cos\left(\tilde{F^V}, F_k^T\right)/\tau\right)}{\sum_{j=1}^{K} \exp\left(\cos\left(\tilde{F^V}, F_j^T\right)/\tau\right)}.$$

Our Solution

Dual-pattern Matching (DPM)



Training-free DPM-F

- **Visual-pattern:** accumulating the similarity of the image-text matching score of all ID data.

$$s_j = \mathtt{Concat}[\tilde{p}(y_1 \mid \mathbf{x_j}), \tilde{p}(y_2 \mid \mathbf{x_j}), ..., \tilde{p}(y_K \mid \mathbf{x_j})],$$

$$P_k^V = \frac{1}{n}\sum_{j=1}^{n} s_j \in \mathbb{R}^{1\times K} \quad P^V = \mathtt{Concat}([P_1^V, P_2^V, ..., P_K^V]) \in \mathbb{R}^{K\times K}$$

- **Textual-pattern:** text feature of all ID data.

$$P^T = \mathtt{Concat}([P_1^T, P_2^T, ..., P_K^T]) \in \mathbb{R}^{K\times C}$$

- **Matching the test feature with dual-pattern:**

$$s_k^V = \mathrm{KL}(\mathrm{Softmax}(s^T)||P_k^V) \in \mathbb{R}$$

$$s^T = \tilde{F}^V P^{T\top} \in \mathbb{R}^{1\times K} \qquad s^V = \mathtt{Concat}([s_1^V, s_2^V, ..., s_K^V]) \in \mathbb{R}^{1\times K}$$

$$DPM(\mathbf{x}) = \max(s^T) - \beta \min(s^V) \in \mathbb{R},$$

$$OOD = \begin{cases} 1, & \text{if } DPM(\mathbf{x}) > \lambda \\ 0, & \text{if } DPM(\mathbf{x}) < \lambda, \end{cases}$$

Our Solution

Training-required DPM-T

- **Text-guided domain-specific feature**

$$\tilde{F^{V}t} = L_v(F^V) + \gamma \sum_{r=1}^{H\cdot W} \frac{\exp(\cos(L_s(F_r^V), L_t(F_k^T)))}{\sum_{j=1}^{H\cdot W} \exp(\cos(L_s(F_j^V), L_t(F_k^T)))} L_s(F_r^V) \in \mathbb{R}^{1\times C}$$

- **Vision-guided domain-specific feature**

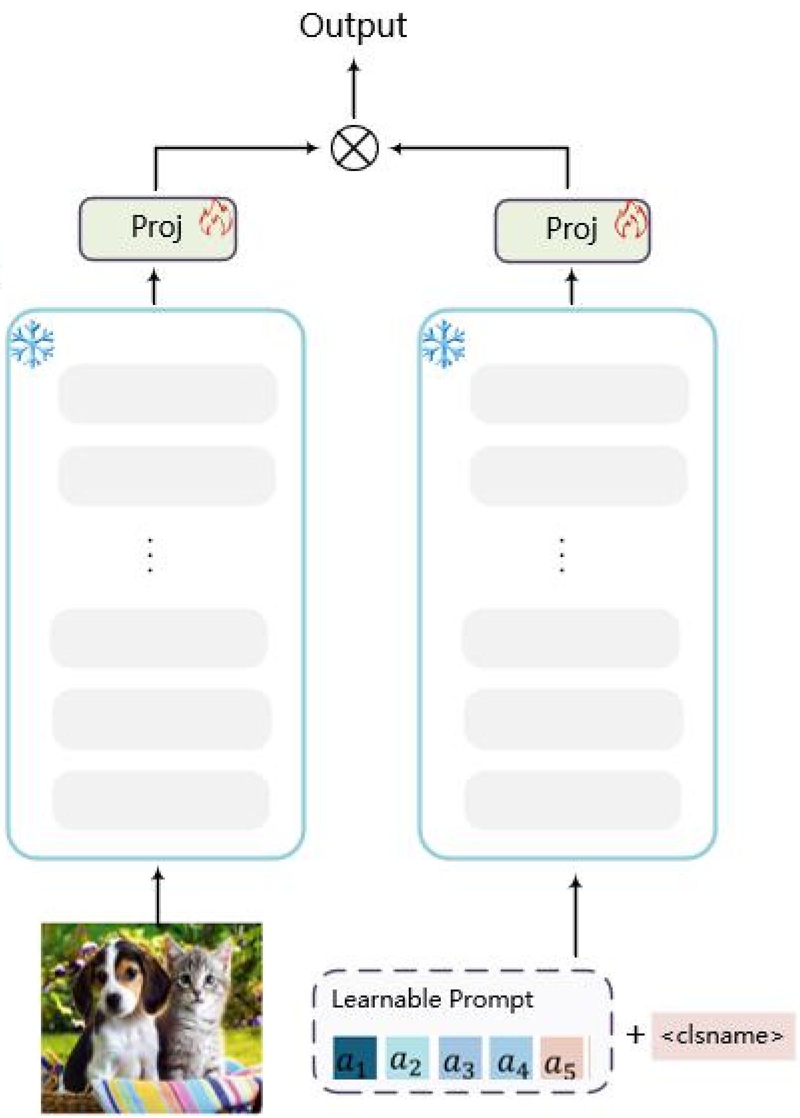
$$\tilde{F^{V}v} = L_v(F^V) + \gamma \sum_{r=1}^{H\cdot W} \frac{\exp(\cos(L_s(F_r^V), \mu_k))}{\sum_{j=1}^{H\cdot W} \exp(\cos(L_s(F_j^V), \mu_k))} L_s(F_r^V) \in \mathbb{R}^{1\times C}$$

- **Matching the visual and textual pattern**

$$p^V(y_k \mid x) = \underbrace{\frac{\exp\left(\cos\left(\tilde{F^{V}v}, \mu_k\right)/\tau\right)}{\sum_{j=1}^{K} \exp\left(\cos\left(\tilde{F^{V}v}, \mu_j\right)/\tau\right)}}_{\text{Visual Matching}} \quad p^T(y_k \mid x) = \underbrace{\frac{\exp\left(\cos\left(\tilde{F^{V}t}, L_t(F^T)\right)/\tau\right)}{\sum_{j=1}^{K} \exp\left(\cos\left(\tilde{F^{V}t}, L_t(F^T)\right)/\tau\right)}}_{\text{Textual Matching}}$$

- **Optimizing the learnable module**

$$\mathcal{L} = \mathcal{L}_{VM} + \mathcal{L}_{TM} = \ell_{ce}(p^V, y) + \ell_{ce}(p^T, y)$$



Main Results

Method	iNaturalist AUR↑	iNaturalist FPR↓	SUN AUR↑	SUN FPR↓	Places AUR↑	Places FPR↓	Texture AUR↑	Texture FPR↓	Avg. AUR↑	Avg. FPR↓
Image Encoder: ViT-B-16										
MSP	83.75	59.18	81.93	61.10	81.11	63.17	79.04	63.14	79.05	63.14
MaxLogit	88.03	60.88	81.32	61.10	80.11	63.17	79.05	64.14	81.06	61.90
Energy	87.54	64.36	87.21	67.26	87.80	56.55	74.89	96.33	84.36	71.13
MCM	90.89	46.51	91.12	42.14	88.68	46.49	87.19	52.38	89.47	46.88
ReAct	89.95	60.45	92.39	43.80	**92.01**	**38.84**	93.90	31.05	92.06	43.53
DPM-F	96.94	12.89	92.62	31.63	89.97	41.15	91.60	32.71	92.78	29.59
Requires training (or w/ fine-tuning)										
CLIPN-A	95.27	23.94	93.93	26.17	90.93	40.83	92.28	33.45	93.10	31.10
DPM-T	**99.03**	**5.08**	**97.07**	**16.77**	91.81	40.54	**94.96**	**21.98**	**95.72**	**21.09**
Image Encoder: ViT-B-32										
MSP	81.80	65.60	79.40	68.21	78.04	68.50	76.31	68.78	78.89	67.77
MaxLogit	86.68	65.12	87.62	58.30	88.65	50.91	76.74	79.72	84.92	63.59
Energy	83.73	75.04	86.00	68.11	87.87	56.52	72.66	87.66	82.57	71.83
MCM	88.80	57.44	90.09	48.52	88.15	50.40	85.83	57.04	88.20	53.10
ReAct	90.56	52.43	91.87	46.56	**92.67**	38.79	82.21	75.96	89.33	53.44
DPM-F	94.86	25.08	93.17	31.55	89.89	43.73	**87.84**	48.90	91.44	37.31
Requires training (or w/ fine-tuning)										
CLIPN-A	94.67	28.75	92.85	31.87	86.93	50.17	87.68	49.49	90.53	40.07
DPM-T	**97.66**	**12.79**	**94.94**	**26.89**	92.61	**31.42**	87.65	**45.91**	**93.22**	**29.25**

Main Results

Method	CIFAR-10		ImageNet-R		LSUN		Avg.	
	AUR ↑	FPR ↓	AUR ↑	FPR ↓	AUR ↑	FPR ↓	AUR ↑	FPR ↓
Image Encoder: ViT-B-16								
MSP [15]	74.53	84.43	58.46	95.59	86.43	62.01	73.14	80.67
MaxLogit [13]	74.36	84.78	84.22	53.33	83.28	85.01	80.62	74.37
Energy [25]	70.88	87.34	84.17	55.07	80.46	89.59	78.50	77.33
ReAct [35]	69.50	93.26	42.62	95.73	84.16	80.91	65.42	89.96
MCM [26]	76.08	85.55	56.81	97.43	88.21	59.02	73.71	80.66
DPM-F	84.81	61.48	84.69	50.95	86.47	59.62	85.32	57.35
Requires training (or w/ fine-tuning)								
CLIPN-A [43]	80.53	69.46	63.79	75.06	89.62	55.83	77.98	66.78
DPM-T	**90.55**	**45.60**	**90.41**	**40.28**	**93.92**	**30.73**	**91.62**	**38.87**
Image Encoder: ViT-B-32								
MSP [15]	70.95	87.29	58.14	97.23	88.43	60.63	72.51	81.72
MaxLogit [13]	74.94	87.38	72.51	71.62	90.98	59.71	79.48	78.91
Energy [25]	71.22	86.86	74.78	67.11	85.90	77.99	77.3	79.25
ReAct [35]	68.96	92.45	68.47	75.56	89.47	69.74	75.63	79.25
MCM [26]	72.37	87.86	56.68	98.02	90.57	53.81	73.21	79.89
DPM-F	78.72	74.53	78.30	64.44	95.41	23.81	84.14	54.26
Requires training (or w/ fine-tuning)								
CLIPN-A [43]	88.06	47.99	87.09	60.07	93.55	35.19	89.57	47.75
GROOD [40]	90.72	43.58	82.02	57.25	86.34	65.98	86.36	55.61
DPM-T	**94.11**	**30.80**	**93.87**	**28.54**	**96.39**	**19.89**	**94.79**	**26.41**

Ablation Study

Different VM calculations

Methods	CIFAR-10 AUROC ↑	CIFAR-10 FPR95 ↓	ImageNet-R AUROC ↑	ImageNet-R FPR95 ↓	LSUN AUROC ↑	LSUN FPR95 ↓
KL	63.91	86.32	61.11	85.69	88.98	42.74
ED	44.24	97.88	75.29	75.66	43.15	98.55
Cos	59.49	92.76	86.69	54.19	55.23	96.48
CosFea	60.40	88.22	92.93	36.87	49.15	97.23
TM	77.23	83.78	76.31	62.99	87.71	67.44
KL + TM	**78.72**	**74.53**	78.30	64.44	**95.41**	**23.81**
ED + TM	61.54	91.44	84.18	55.33	69.24	90.96
Cos + TM	76.04	79.01	88.49	39.31	84.81	73.05
CosFea + TM	69.36	84.56	**91.68**	**37.16**	68.53	93.59

Different loss

$\mathcal{L}_{org}$	$\mathcal{L}_{VM}$	$\mathcal{L}_{TM}$	AUROC↑	FPR95↓
✓	×	×	86.65	55.90
×	✓	×	86.08	57.15
×	×	✓	61.14	84.87
✓	✓	×	87.23	51.63
✓	×	✓	84.07	60.52
×	✓	✓	**91.62**	**38.87**
✓	✓	✓	91.07	39.93

Different learnable module

Prompt	Projection	DSFA	AUROC↑	FPR95↓
×	×	×	84.14	54.26
✓	×	×	84.99	55.69
×	×	✓	76.93	71.27
✓	✓	×	92.63	38.08
✓	×	✓	85.12	56.34
×	✓	✓	90.54	38.95
✓	✓	✓	**94.79**	**26.41**

Conclusion

Main contributions

- We propose a novel method DPM (Dual Pattern Matching) that effectively uses visual and textual modalities for efficient Out-of-Distribution (OOD) detection.
- We propose a domain-specific features aggregation module to refine the CLIP visual feature for better alignment with the textual features.
- Our DPM-F and DPM-T exhibit state-of-the-art performance on OOD detection benchmarks, demonstrating the superiority of our proposed approaches

Future Work

- Can DPM be applied to other OOD setting e.g. few-shot OOD detection?
- How to further improve the performance of DPM in OOD detection?

Thank you !