FSOD-VFM: Training-Free Few-Shot Object Detection with Vision Foundation Models and Graph Diffusion

Prject Page: https://intellindust-ai-lab.github.io/projects/FSOD-VFM

Chen-Bin Feng[12*], Youyang Sha[1*], Longfei Liu[1], Yongjun Yu[1], Chi Man Vong[2†], Xuanlong Yu[1†], Xi Shen[1†]

[1]Intellindust AI Lab & [2]University of Macau

* Eequal contribution † Corresponding Author

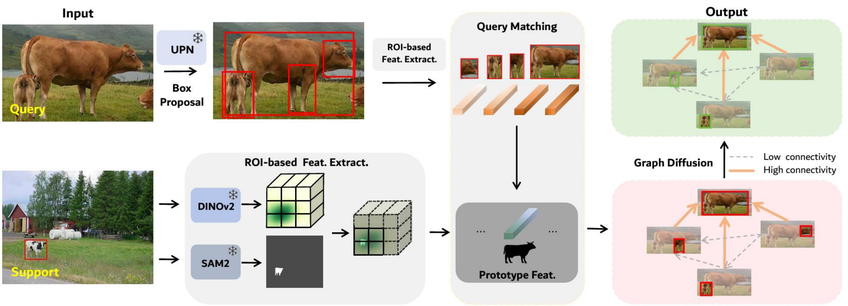
ICLR International Conference On Learning Representations

Key Idea

A training-free framework integrating vision foundation models (VFMs) and graph diffusion to address few-shot object detection:

- VFMs for proposal generation and feature extraction.
- Graph diffusion to refine proposal confidence

Overview



1. VFMs

- UPN: Generate object proposals.
- SAM2: Compute object masks.
- DINOv2: Extract visual features.

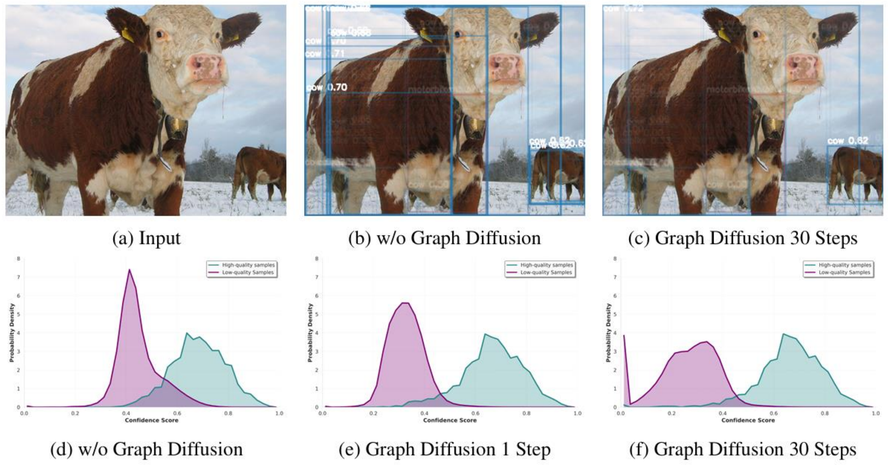
2. Prototype Matching

- Support Set: Aggregate features to form class prototypes.
- Query Matching: Predict classes for proposals via cosine similarity.

3. Graph Diffusion for Confidence Refinement

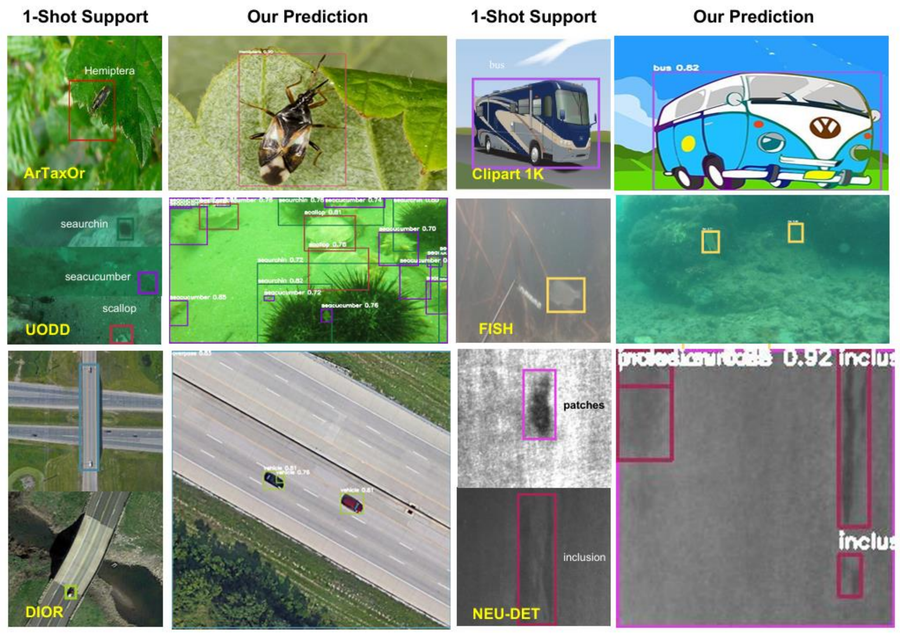
- Energy is diffused from low-confidence to high-confidence proposals.
- Fragmented bounding boxes are assigned as low confidence.

Effect of Graph Diffusion



- 1st Row: Score of high-quality object becomes more important after graph diffusion.
- 2nd Row: High-quality and low-quality boxes distributions before/after graph diffusion.

Visualization



Experimental Reuslts

Method	F.T. Novel.	Novel Split 1					Novel Split 2					Novel Split 3					Avg
		1	2	3	5	10	1	2	3	5	10	1	2	3	5	10	
FsDetView Xiao et al. (2022)	✓	25.4	20.4	37.4	36.1	42.3	22.9	21.7	22.6	25.6	29.2	32.4	19.0	29.8	33.2	39.8	29.2
TFA Wang et al. (2020)	✓	39.8	36.1	44.7	55.7	56.0	23.5	26.9	34.1	35.1	39.1	30.8	34.8	42.8	49.5	49.8	39.9
Retentive RCNN Fan et al. (2021)	✓	42.4	45.8	45.9	53.7	56.1	21.7	27.8	35.2	37.0	40.3	30.2	37.6	43.0	49.7	50.1	41.1
DiGeo Ma et al. (2023)	✓	37.9	39.4	48.5	58.6	61.5	26.6	28.9	41.9	42.1	49.1	30.4	40.1	46.9	52.7	54.7	44.0
HeteroGraph Han et al. (2021)	✓	42.4	51.9	55.7	62.6	63.4	25.9	37.8	46.6	48.9	51.1	35.2	42.9	47.8	54.8	53.5	48.0
Meta Faster R-CNN Han et al. (2022a)	✓	43.0	54.5	60.6	66.1	65.4	27.7	35.5	46.1	47.8	51.4	40.6	46.4	53.4	59.9	58.6	50.5
CrossTransformer Han et al. (2022b)	✓	49.9	57.1	57.9	63.2	67.1	27.6	34.5	43.7	49.2	51.2	39.5	54.7	52.3	57.0	58.7	50.9
LVC Kaul et al. (2022)	✓	54.5	53.2	58.8	63.2	65.7	32.8	29.2	50.7	49.8	50.6	48.4	52.7	55.0	59.6	59.6	52.3
NIFF Guirguis et al. (2023)	✓	62.8	67.2	68.0	70.3	68.8	38.4	42.9	54.0	56.4	54.0	56.4	62.1	61.2	64.1	63.9	59.4
Multi-Relation Det Fan et al. (2020)	✗	37.8	43.6	51.6	56.5	58.6	22.5	30.6	40.7	43.1	47.6	31.0	37.9	43.7	51.3	49.8	43.1
DE-ViT (ViT-S/14) Zhang et al. (2023)	✗	47.5	64.5	57.0	68.5	67.3	43.1	34.1	49.7	56.7	60.8	52.5	62.1	60.7	61.4	64.5	56.7
DE-ViT (ViT-B/14) Zhang et al. (2023)	✗	56.9	61.8	68.0	73.9	72.8	45.3	47.3	58.2	59.8	60.6	58.6	62.3	62.7	64.6	67.8	61.4
DE-ViT (ViT-L/14) Zhang et al. (2023)	✗	55.4	56.1	68.1	70.9	71.9	43.0	39.3	58.1	61.6	63.1	58.2	64.0	61.3	64.2	67.3	60.2
No-Time-To-Train Espinosa et al. (2025)	✗	70.8	72.3	73.3	77.2	79.1	54.5	67.0	76.3	75.9	78.2	61.1	67.9	71.3	70.8	72.6	71.2
FSOD-VFM	✗	**77.5**	**82.3**	**83.0**	**85.8**	**85.8**	**64.8**	**68.0**	**77.4**	**79.5**	**81.6**	**65.3**	**75.1**	**78.7**	**78.2**	**79.3**	**77.5**

Results for competing methods are taken from Zhang et al. (2023), with the best highlighted in **bold**.

Table 1: **Results on Pascal-5^i Everingham et al. (2010).** We report nAP50, i.e., the average precision at IoU 0.5 on novel classes.

Method	F.T. Novel.	10-shot			30-shot		
		nAP	nAP50	nAP75	nAP	nAP50	nAP75
TFA Wang et al. (2020)	✓	10.0	19.2	9.2	13.5	24.9	13.2
FSCE Sun et al. (2021)	✓	11.9	–	10.5	16.4	–	16.2
Retentive RCNN Fan et al. (2021)	✓	10.5	19.5	9.3	13.8	22.9	13.8
HeteroGraph Han et al. (2021)	✓	11.6	23.9	9.8	16.5	31.9	15.5
Meta F. R-CNN Han et al. (2022a)	✓	12.7	25.7	10.8	16.6	31.8	15.8
LVC Kaul et al. (2022)	✓	19.0	34.1	19.0	26.8	45.8	27.5
C. Transformer Han et al. (2022b)	✓	17.1	30.2	17.0	21.4	35.5	22.1
NIFF Guirguis et al. (2023)	✓	18.8	–	–	20.9	–	–
DiGeo Ma et al. (2023)	✓	10.3	18.7	9.9	14.2	26.2	14.8
CD-ViTO (ViT-L) Fu et al. (2024)	✓	35.3	54.9	37.2	35.9	54.5	38.0
FSRW Kang et al. (2019)	✗	5.6	12.3	4.6	9.1	19.0	7.6
Meta R-CNN Yan et al. (2019)	✗	6.1	19.1	6.6	9.9	25.3	10.8
DE-ViT (ViT-L) Zhang et al. (2023)	✗	34.0	53.0	37.0	34.0	52.9	37.2
No-Time-To-Train Espinosa et al. (2025)	✗	36.6	54.1	38.3	36.8	54.5	38.7
FSOD-VFM	✗	**44.0**	**59.4**	**47.6**	**45.8**	**61.9**	**49.4**

Results for competing methods are taken from Fu et al. (2024), with the best highlighted in **bold**.

Table 2: **Results on COCO-20^i Kang et al. (2019); Lin et al. (2014).** We report nAP (IoU thresholds 0.5–0.95), nAP50 (IoU 0.5), and nAP75 (IoU threshold 0.75) on novel classes.

Method	F.T. Novel.	ArTaxOr	Clip art1k	DIOR	Deep Fish	NEU DET	UODD	Avg
TFA w/cos Wang et al. (2020)○	✓	3.1 / 8.8 / 14.8	-	8.0 / 18.1 / 20.5	-	-	4.4 / 8.7 / 11.8	-
FSCE Sun et al. (2021)○	✓	3.7 / 10.2 / 15.9	-	8.6 / 18.7 / 21.9	-	-	3.9 / 9.6 / 12.0	-
DeFRCN Qiao et al. (2021)○	✓	3.6 / 9.9 / 15.5	-	9.3 / 18.9 / 22.9	-	-	4.5 / 9.9 / 12.1	-
Distill-cdfsd Xiong (2023)○	✓	5.1 / 12.5 / 18.1	7.6 / 23.3 / 27.3	10.5 / 19.1 / 26.5	- / 15.5 / 15.5	- / 16.0 / 21.1	5.9 / 12.2 / 14.5	- / 16.4 / 20.5
ViTDeT-FT† Li et al. (2022)†	✓	5.9 / 20.9 / 23.4	6.1 / 23.3 / 25.6	12.9 / 23.3 / 29.4	0.9 / 9.0 / 6.5	2.4 / 13.5 / 15.8	4.0 / 11.1 / 15.6	5.4 / 16.9 / 19.4
Detic-FT Zhou et al. (2022)†	✓	3.2 / 8.7 / 12.0	15.1 / 20.2 / 22.3	4.1 / 12.1 / 15.4	9.0 / 14.3 / 17.9	3.8 / 14.1 / 16.8	4.2 / 10.4 / 14.4	6.6 / 13.3 / 16.5
DE-ViT-FT Zhang et al. (2023)†	✓	10.5 / 38.0 / 49.2	13.0 / 38.1/40.8	14.7 / 23.4 / 25.6	19.3 / 21.2 / 21.3	0.6 / 7.8 / 8.8	2.4 / 5.0 / 5.4	10.1 / 22.3 / 25.2
CD-ViTO Fu et al. (2024)†	✓	21.0 / 47.9 / 60.5	17.7 / 41.1 / 44.3	17.8 / 26.9 / 30.8	20.3 / 22.3 / 22.3	3.6 /11.4 / 12.8	3.1 / 6.8 / 7.0	13.9 / 26.1 / 29.6
Mixture Liu et al. 2025	✓	26.1 / **63.3 / 71.3**	20.1 / **45.1 / 49.9**	**20.6 / 32.1 / 37.8**	24.2 / 29.5 / 34.1	**9.1/ 19.0 / 23.7**	9.0 / 19.6 / **22.1**	18.2 / **34.7 / 39.8**
Meta-RCNN Yan et al. (2019)○	✗	2.8 / 8.5 / 14.0	-	7.8 / 17.7 / 20.6	-	-	3.6 / 8.8 / 11.2	-
Detic Zhou et al. (2022)†	✗	0.6 / 0.6 / 0.6	11.4 / 11.4 / 11.4	0.1 / 0.1 / 0.1	0.9 / 0.9 / 0.9	0.0 / 0.0 / 0.0	0.0 / 0.0 /0.0	2.2 / 2.2 / 2.2
DE-ViT Zhang et al. (2023)†	✗	0.4 / 10.1 / 9.2	0.5 / 5.5 / 11.0	2.7 / 7.8 / 8.4	0.4 / 2.5 / 2.1	0.4 / 1.5 / 1.8	1.5 / 3.1 /3 .1	1.0 /5.1 / 5.9
No-Time-To-Train Espinosa et al. 2025	✗	**28.2** / 35.7 / 35.0	18.9 / 24.9 / 25.9	14.9 / 18.5 / 16.4	30.5 / 28.9 / 29.6	5.5 / 5.2 / 5.5	**10.0 / 20.2** / 16.0	18.0 / 22.4 / 21.4
FSOD-VFM	✗	**51.4** / 62.0 / 61.5	**29.1** / 43.7 / 46.5	18.3 / 23.5 / 22.5	**35.0 / 33.9 / 34.3**	5.9 / 7.4 / 7.2	**11.8** / 17.3 / 17.5	**25.3** / 31.3 / 31.6

○ indicates Distill-CD-FSOD Xiong (2023) results, and † denotes CD-ViTO Fu et al. (2024) results. Best results are in **bold**.

Table 3: **Results on the CD-FSOD benchmark Xiong (2023).** We report nAP (IoU thresholds 0.5–0.95) for all datasets, with each entry showing results for 1-shot, 5-shot, and 10-shot.

FSOD-VFM outperforms all the training-free methods and most of the fine-tuning methods on Pascal-5^i, COCO-20^i, and CD-FSOD.

Limitations

FSOD-VFM suffers from high inference latency and limited improvement with high-shot scenario.