Conditional Prototype Learning for Few-Shot Object Detection

Zhenwei He Xinye Liao Xin Feng✉

School of Computer Science and Engineering, Chongqing University of Technology, Chongqing, China

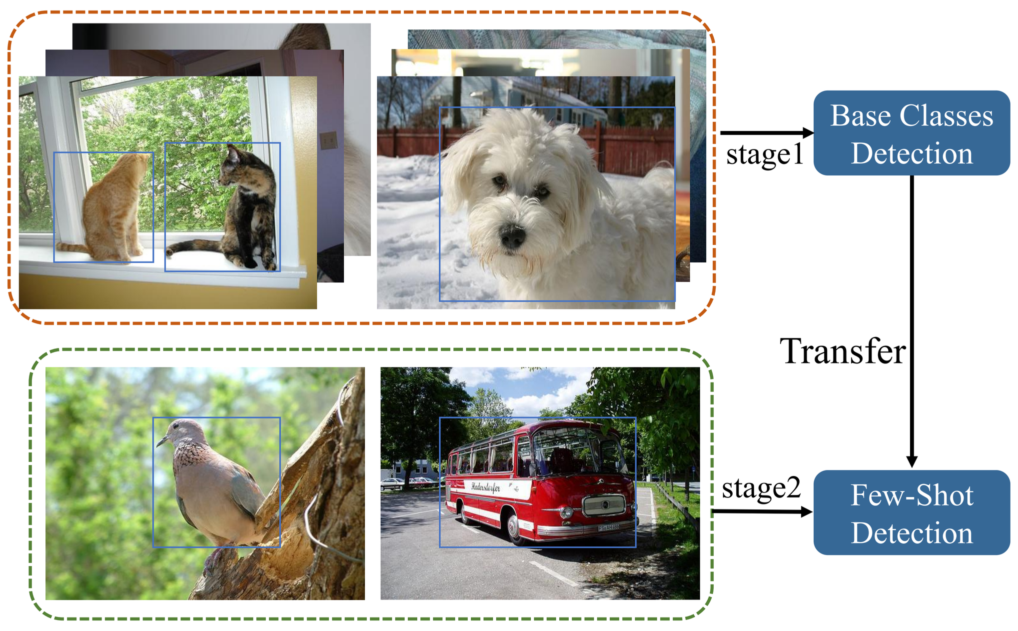
BMVC 2025

github

1.Background

What is Few-Shot Object Detection?

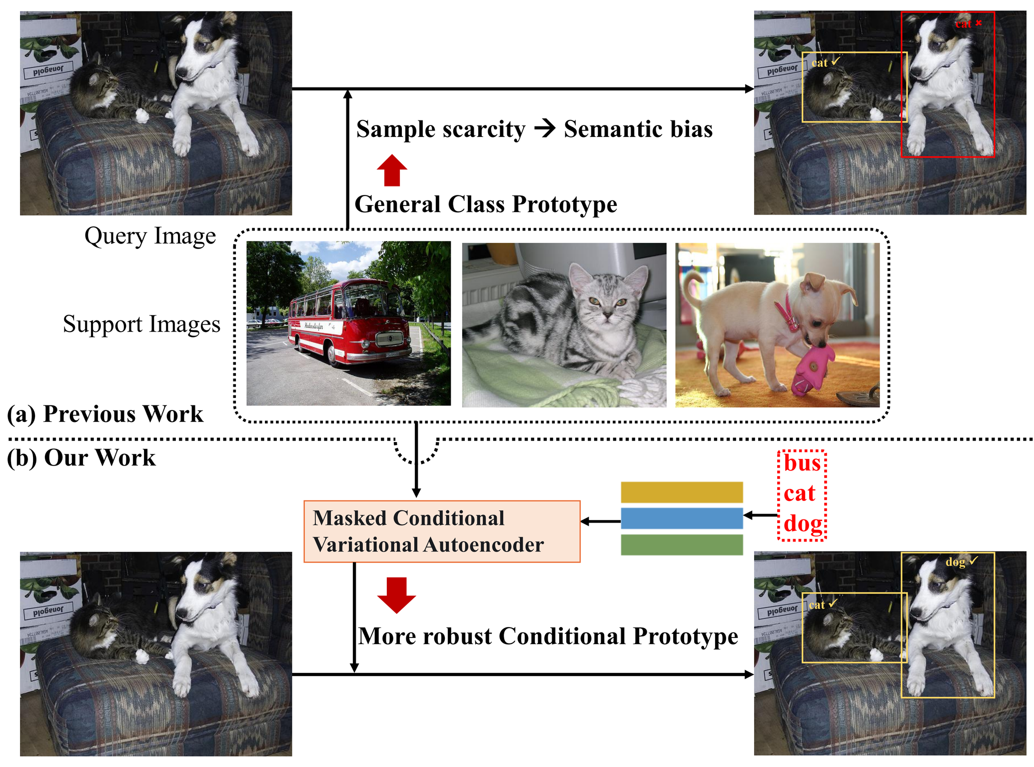
- **Detect with Few Examples**
  Few-Shot Object Detection aims to detect objects from **novel classes** using **only a handful of annotated samples.**
- **Handle Data Scarcity**
  In real-world scenarios, **collecting and labeling data is expensive**. FSOD tackles this by enabling detection for **rare or underrepresented categories** with limited supervision.
- **Learn to Generalize**
  FSOD uses **meta-learning** or **transfer learning** to extract knowledge from base classes and **generalize to novel classes quickly** with minimal data.



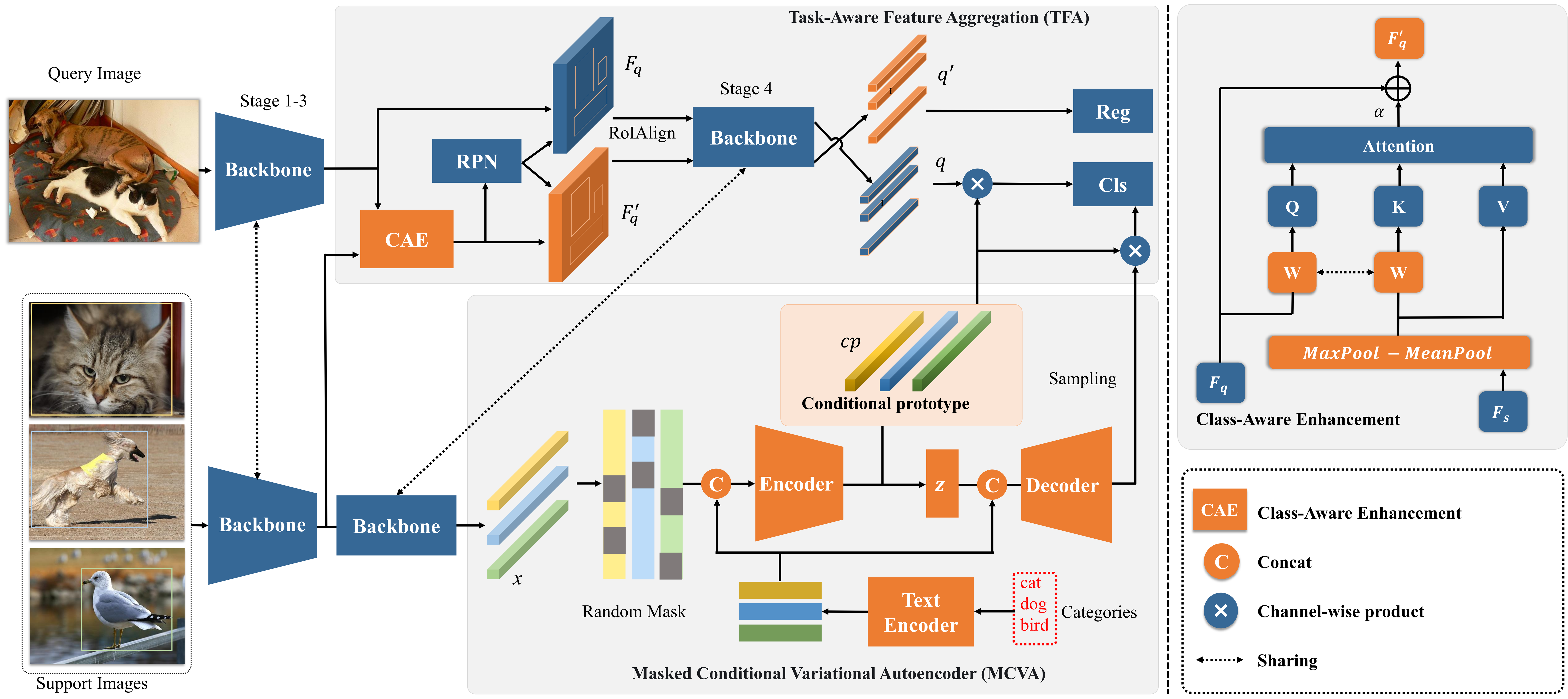
2.Motivation

In few-shot object detection (FSOD), prototypes guide learning. But with few samples, prototypes often overfit to sample-specific details rather than true class semantics → **semantic bias.**

To reduce semantic bias, we guide learning using category-specific semantics, enabling the model to generate **more robust prototypes** that better adapt to novel classes.



3.Overall Framework

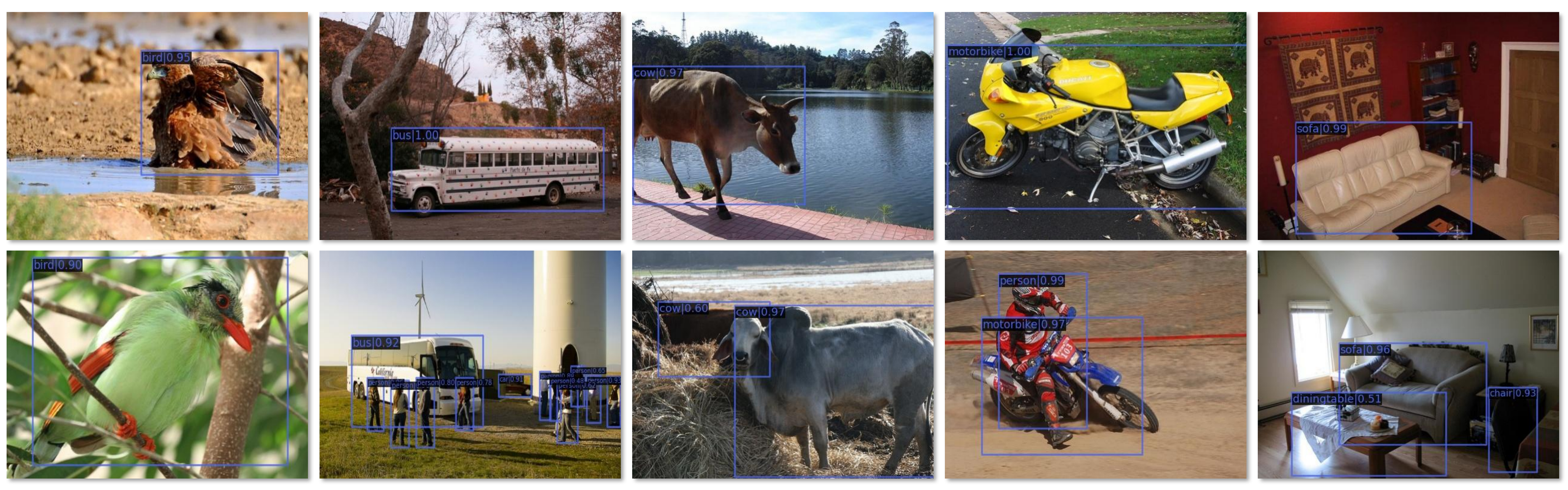


4.Experimental Results

Few-shot object detection performance (nAP50) on PASCAL VOC dataset. We evaluate the performance on three different splits.

Methods		split 1					split 2					split 3				
		1	2	3	5	10	1	2	3	5	10	1	2	3	5	10
FSRW [19]	ICCV19	14.8	15.5	26.7	33.9	47.2	15.7	15.3	22.7	30.1	40.5	21.3	25.6	28.4	42.8	45.9
Meta RCNN [42]	ICCV19	19.9	25.5	35.0	45.7	51.5	10.4	19.4	29.6	34.8	45.4	14.3	18.2	27.5	41.2	48.1
TFA w/ cos [35]	ICML20	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
MPSR [38]	ECCV20	41.7	-	51.4	55.2	61.8	24.4	-	39.2	39.9	47.8	35.6	-	42.3	48.0	49.7
DCNet [18]	CVPR21	33.9	37.4	43.7	51.1	59.6	23.2	24.8	30.6	36.7	46.6	32.3	34.9	39.7	42.6	50.7
QA-FewDet [12]	ICCV21	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
FSCE [34]	CVPR21	44.2	43.8	51.4	61.9	63.4	27.3	29.5	43.5	44.2	50.2	37.2	41.9	47.5	54.6	58.5
DeFRCN [27]	ICCV21	53.6	57.5	61.5	64.1	60.8	30.1	38.1	47.0	53.3	47.9	48.4	50.9	52.3	54.9	57.4
KFSOD [47]	CVPR22	44.6	45.2	54.4	60.9	65.8	37.8	38.4	43.1	48.1	50.4	34.8	42.7	44.1	52.7	53.9
MRSN [25]	ECCV22	47.6	48.6	57.8	61.9	62.6	31.2	38.3	46.7	47.1	50.6	35.5	30.9	45.6	54.4	57.4
Meta FR-CNN [13]	AAAI22	43.0	54.6	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
σ-ADP [6]	ICCV23	52.3	55.5	63.1	65.9	66.7	**42.7**	45.8	48.7	54.8	56.3	47.8	51.8	56.8	60.3	62.4
ICPE [24]	AAAI23	54.3	59.5	62.4	65.7	66.2	33.5	40.1	48.7	51.7	52.5	**50.9**	53.1	55.3	60.6	60.1
VFA [14]	AAAI23	**57.7**	**64.6**	**64.7**	67.2	67.4	41.4	**46.2**	**51.1**	51.8	51.6	48.9	**54.8**	56.6	59.0	58.9
FPD [37]	AAAI24	48.1	62.2	64.0	67.6	68.4	29.8	43.2	47.7	52.0	53.9	44.9	53.8	**58.1**	61.6	62.9
FM-FSOD [11]	CVPR24	40.1	53.5	57.0	**68.6**	**72.0**	33.1	36.3	48.8	**54.8**	**64.7**	39.2	50.2	55.7	**63.4**	**68.1**
CPL	Ours work	**60.6**	**68.2**	**69.5**	**70.9**	**70.2**	**43.0**	**51.5**	**55.5**	**55.9**	**56.9**	**54.0**	**58.5**	**60.9**	**64.1**	**63.0**

Visualization of the detection results on novel classes.



5.Ablation Study

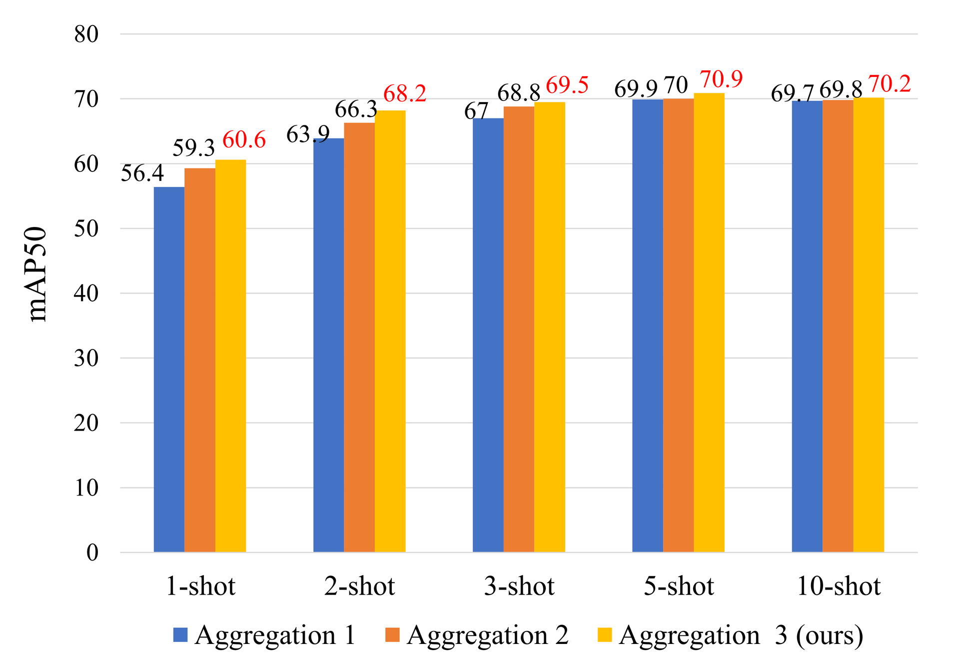
Ablation study of different components.

Method	MCVA	TFA	shot		
			1	3	5
Baseline			40.2	54.0	55.0
Ours	✓		53.0	67.5	69.9
		✓	53.2	65.2	68.3
	✓	✓	**60.6**	**69.5**	**70.9**

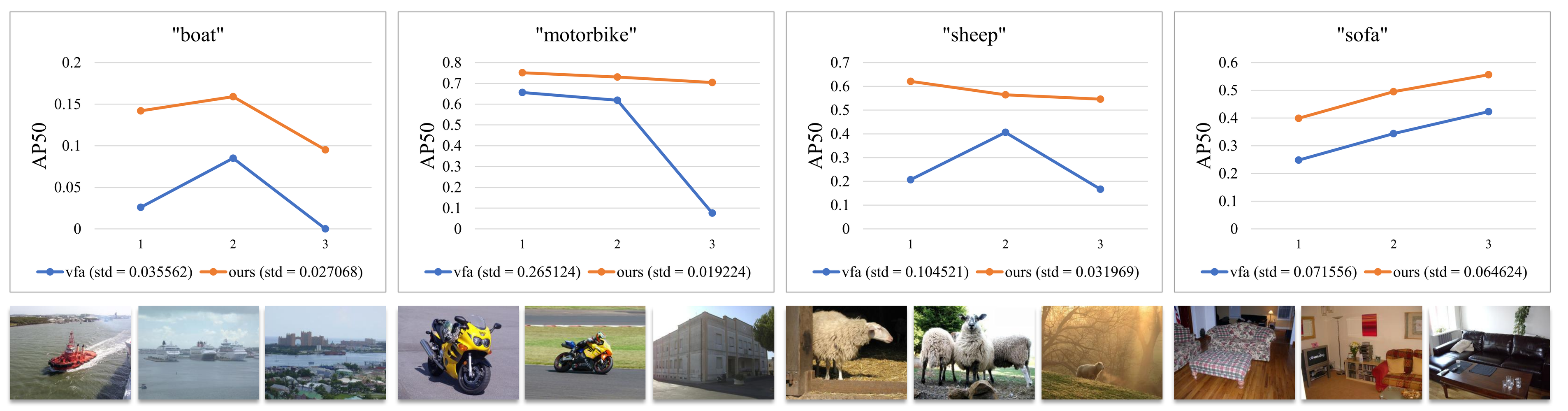
Analysis of different text encoders. Both the word embedding methods help the model achieve better performance, while CILP performs better.

Method	Text Encoder	shot				
		1	2	3	5	10
Ours	w/o reference	54.9	64.3	67.4	69.1	68.9
	CLIP [28]	**60.6**	**68.2**	**69.5**	**70.9**	**70.2**
	Word2Vec [26]	56.4	66.2	67.5	69.6	69.0

Analysis of different aggregation methods.



Detection performance varies with different training samples. We selected three different samples for "boat", "motorbike", "sheep", and "sofa". And conducted training under the 1-shot setting, using the Standard Deviation (std) to measure stability.



6. Conclusion

- We explore the issue of **semantic bias in class prototypes** for few-shot object detection (FSOD) under the meta-learning paradigm
- We introduce the **Masked Conditional Variational Autoencoder (MCVA)** to refine the semantic bias in class prototypes, generating more robust conditional prototypes.
- Considering that the classification and regression tasks need different kinds of features, we propose the **Task-Aware Feature Aggregation (TFA)** module, which separately enhances features for the two tasks.
- Extensive experiments on PASCAL VOC and MS COCO demonstrate that our approach achieves state-of-the-art performance.

7. Acknowledgement

- The key project of Humanities and Social Sciences under the Chongqing Ministry of Education(Grant No. 24SKD134),
- The National Natural Science Foundation of China Youth Program (Grant No. 62306053).
- The Graduate Innovation Project of Chongqing University of Technology (Grant No. gzlcx20242047).