Robust Region Feature Synthesizer for Zero-Shot Object Detection

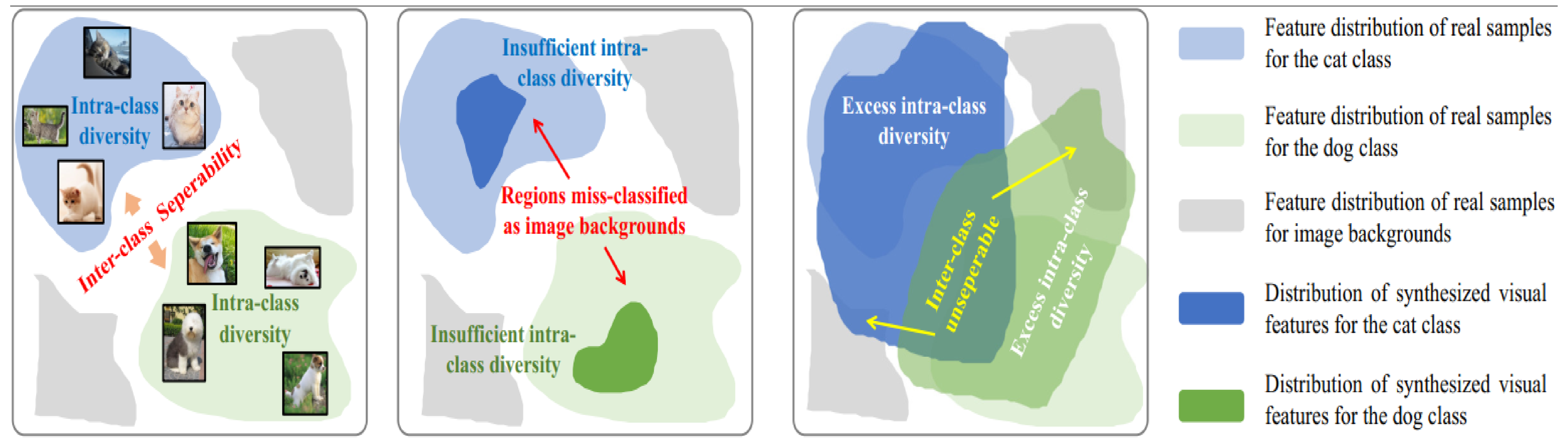
NORTHWESTERN POLYTECHNICAL UNIVERSITY

Peiliang Huang[1], Junwei Han[1], De Cheng[2], Dingwen Zhang[1] ([1]NPU, [2]XDU)

CVPR JUNE 19-24 2022 NEW ORLEANS • LOUISIANA

Motivation:

- **Intra-class diversity:** objects in real-world detection scenarios present high variation in pose, shape, texture, etc.
- **Inter-class separability:** each object category has easy-to-recognized characteristics that are distinct from other object categories
- Existing approaches did not jointly consider the intra-class diversity and inter-class separability.



(a) Feature distribution of real samples (b) Distribution of the synthesized visual features (Case I) (c) Distribution of the synthesized visual features (Case II)

Our approach:

We design a unified region feature synthesizer for feature synthesizing in real-world detection scenarios.

The contributions are:

- We reveal the key challenges, i.e., the intra-class diversity and inter-class separability, for feature synthesizing in real-world object detection scenarios.
- With the goal to synthesize robust region features for ZSD, we build a novel framework that contains an Intra-class Semantic Diverging component and an Inter-class Structure Preserving component.
- Comprehensive experiments on three datasets, including PASCAL VOC, COCO, and DIOR, demonstrate the effectives of the proposed approach. Notably, this is also the first attempt for implementing zero-shot object detection in remote sensing imagery.

Intra-class Semantic Diverging:

The visual features synthesized from adjacent noise vectors will be pulled closer while those synthesized from distinct noise vectors will be pushed away.

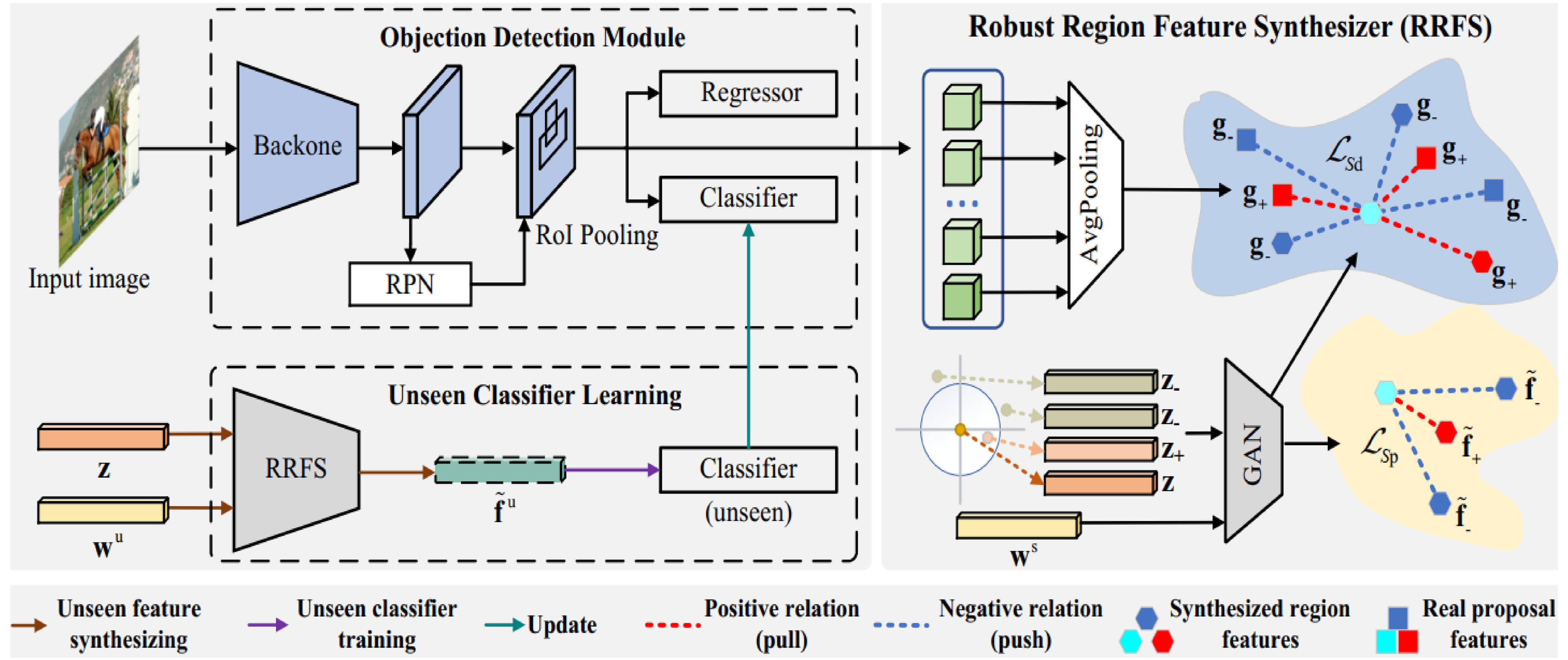
$$\mathcal{L}_{S_d} = \mathbb{E}[-\log \frac{\exp(\tilde{\mathbf{f}}^s \cdot \tilde{\mathbf{f}}^s_+/\tau)}{\exp(\tilde{\mathbf{f}}^s \cdot \tilde{\mathbf{f}}^s_+/\tau) + \sum_{i=1}^{N} \exp(\tilde{\mathbf{f}}^s \cdot \tilde{\mathbf{f}}^s_{i-}/\tau)}]$$

Inter-class Structure Preserving :

By pushing away the visual features from different categories this learning component can effectively enhance the discrimination of the synthesized visual features.

$$\mathcal{L}_{S_p} = \mathbb{E}[-\log \frac{\exp(\tilde{\mathbf{f}}^s \cdot \mathbf{g}_+/\tau)}{\exp(\tilde{\mathbf{f}}^s \cdot \mathbf{g}_+/\tau) + \sum_{j \in \Phi} \exp(\tilde{\mathbf{f}}^s \cdot \mathbf{g}_j/\tau)}]$$

Framework:



- Our method contains an object detection module and a unseen classifier learning module.
- The leaning objective function of RRFS is:

$$\min_G \max_D \mathcal{L}_{\mathrm{WGAN}} + \lambda_1 \mathcal{L}_{\mathrm{C_s}} + \lambda_2 \mathcal{L}_{\mathrm{S_d}} + \lambda_3 \mathcal{L}_{\mathrm{S_p}}$$

Results:

Performance on PASCAL VOC

Method	ZSD	GZSD		
		S	U	HM
SAN	59.1	48	37	41.8
HRE	54.2	**62.4**	25.5	36.2
BLC	55.2	58.2	22.9	32.9
SU	64.9	—	—	—
Ours	**65.5**	47.1	**49.1**	**48.1**

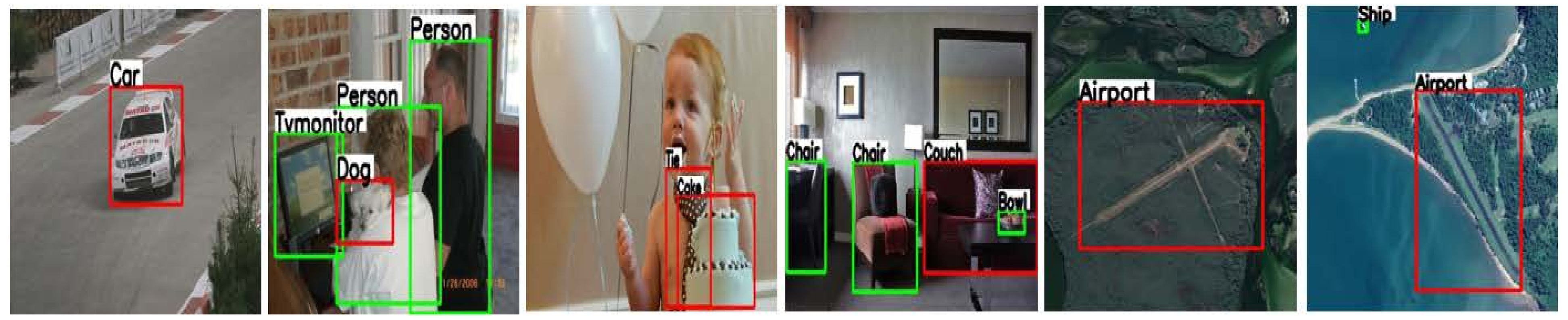
Performance on DIOR

Method	ZSD	GZSD		
		S	U	HM
PL	0.4	4.3	0	0
BLC	1.1	6.1	0.4	0.8
SU	10.5	30.9	2.9	5.3
Ours	**11.3**	**30.9**	**3.4**	**6.1**

Performance on PASCAL VOC

Method	Split	Recall			mAP		
		S	U	HM	S	U	HM
PL	48/17	38.2	26.3	31.2	35.9	4.1	7.4
BLC	48/17	57.6	46.4	51.4	42.1	4.5	8.2
Ours	48/17	**59.7**	**58.8**	**59.2**	**42.3**	**13.4**	**20.4**
PL	65/15	36.4	37.2	36.8	34.1	12.4	18.2
BLC	65/15	56.4	51.7	53.9	36.0	13.1	19.2
SU	65/15	57.7	53.9	55.8	36.9	19.0	25.1
Ours	65/15	**58.6**	**61.8**	**60.2**	**37.4**	**19.8**	**26.0**

Qualitative results on three datasets



Code:https://github.com/HPL123/RRFS

Contact with us:https://nwpu-brainlab.gitee.io/index_en.html