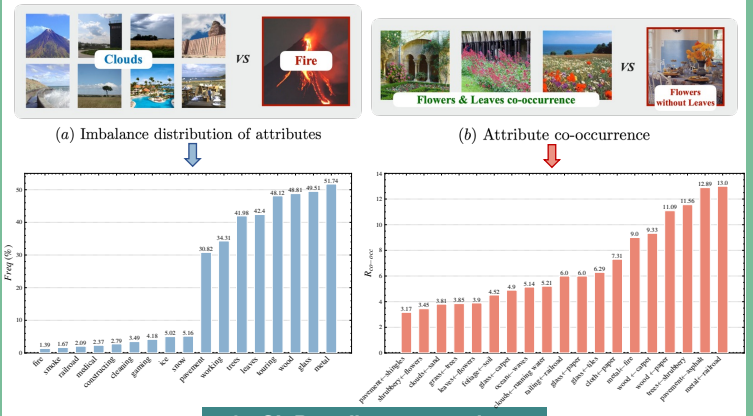
DUET: Cross-modal Semantic Grounding for Contrastive Zero-shot Learning

Zhuo Chen[1,5], Yufeng Huang[2,5], Jiaoyan Chen[3], Yuxia Geng[1,2], Wen Zhang[2,5], Yin Fang[1,5], Jeff Z. Pan[4], Huajun Chen[1,5*]

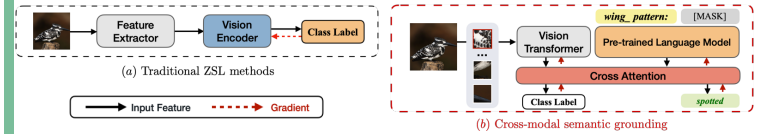
1. College of Computer Science and Technology, Zhejiang University
2. School of Software Technology, Zhejiang University
3. Department of Computer Science, The University of Manchester
4. School of Informatics, The University of Edinburgh
5. Alibaba-Zhejiang University Joint Institute of Frontier Technologies

1 Challenges



(a) Imbalance distribution of attributes (b) Attribute co-occurrence

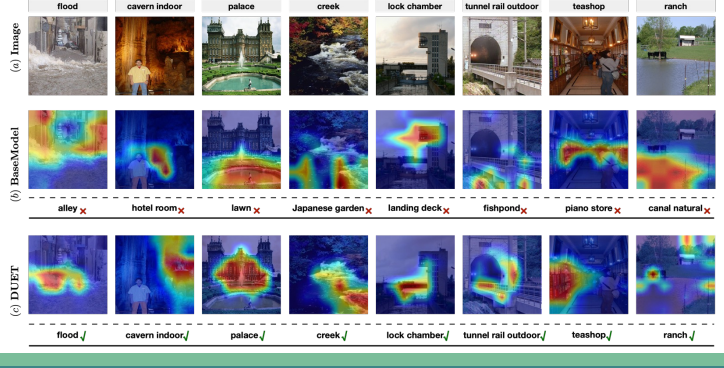
2 ZSL Paradigm comparison



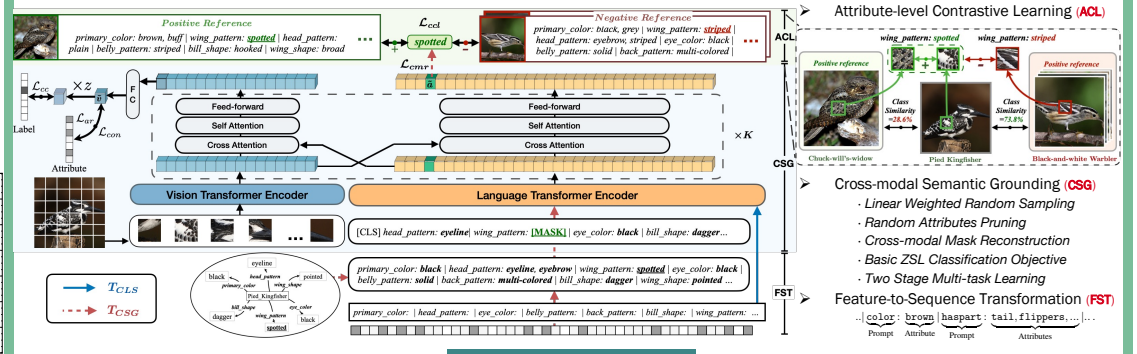
(a) Traditional ZSL methods (b) Cross-modal semantic grounding

We exploits the semantics of PLMs to augment the transformer-based vision encoder via reconstructing masked attributes with a cross-model attention mechanism.

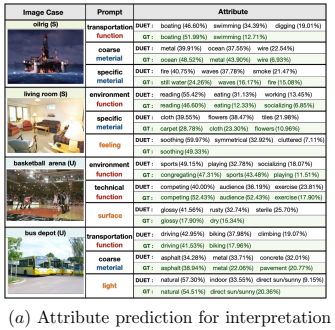
5 Semantic Grounding and Visualization



3 Propose DUET



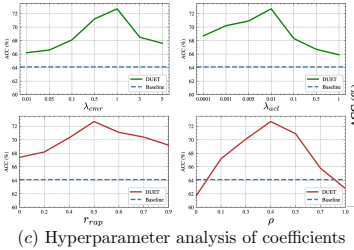
4 Experiments



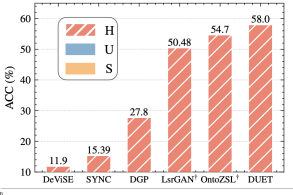
(a) Attribute prediction for interpretation

	Methods	CUB				SUN				AWA2			
		CZSL	GZSL			CZSL	GZSL			CZSL	GZSL		
		T1	U	S	H	T1	U	S	H	T1	U	S	H
†	TF-VAEGAN (ECCV) (2020)	64.9	52.8	64.7	58.1	66.0	45.6	40.7	43.0	72.2	59.8	75.1	66.6
	Composer (NeurIPS) (2020a)	69.4	56.4	63.8	59.9	62.6	55.1	22.0	31.4	71.5	62.1	77.3	68.8
	CE-GZSL (CVPR) (2021)	77.5	63.1	66.8	65.3	63.3	48.8	38.6	43.1	70.4	63.1	78.6	70.0
	GCM-CF (CVPR) (2021)	–	61.0	59.7	60.3	–	47.9	37.8	42.2	–	60.4	75.1	67.0
	FREE (ICCV) (2021c)	–	55.7	59.9	57.7	–	47.4	37.2	41.7	–	60.4	75.4	67.1
	HSVA (NeurIPS) (2021d)	62.8	52.7	58.3	55.3	63.8	48.6	39.0	43.3	–	59.3	76.6	66.8
	AGZSL (ICLR) (2021)	57.2	41.4	49.7	45.2	63.3	29.9	40.2	34.3	73.8	65.1	78.9	71.3
⋆	APN (NeurIPS) (2020)	72.0	65.3	69.3	67.2	61.6	41.9	34.0	37.6	68.4	57.1	72.4	63.9
	DVBE (CVPR) (2020)	–	53.2	60.2	56.5	–	45.0	37.2	40.7	–	63.6	70.8	67.0
	DAZLE (CVPR) (2020b)	66.0	56.7	59.6	58.1	59.4	52.3	24.3	33.2	67.9	60.3	75.7	67.1
	RGEN (ECCV) (2020)	76.1	60.0	73.5	66.1	63.8	44.0	31.7	36.8	73.6	67.1	76.5	71.5
	GEM-ZSL (CVPR) (2021a)	77.8	64.8	69.3	67.2	62.8	38.1	35.7	36.9	67.3	64.8	77.5	70.6
	MSDN (CVPR) (2022b)	76.1	68.7	67.5	68.1	65.8	52.2	34.2	41.3	70.1	62.0	74.5	67.7
	TransZero (AAAI) (2022a)	76.8	69.3	68.3	68.8	65.6	52.6	33.4	40.8	70.1	61.3	82.3	70.2
	DUET (Ours)	72.3	62.9	72.8	67.5	64.4	45.7	**45.8**	**45.8**	69.9	63.7	**84.7**	**72.7**

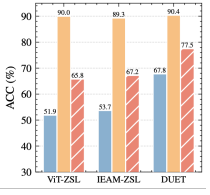
(b) Overall results on standard ZSL datasets compared with 14 representative methods



(c) Hyperparameter analysis of coefficients



(d) Performance on K-ZSL dataset



(e) ViT-based as ENC_{vis}

Methods	H	Δ
Only ENC_{vis}	64.1	8.6↓
1) $CSG_{freeze\ ENC_{lan}}$	66.5	6.2↓
2) $CSG_{w/\ only\ Prompt}$	61.7	11.0↓
3) $CSG_{w/o\ Prompt}$	64.9	7.8↓
4) $CSG_{w/o\ LWRS}$	66.9	5.8↓
5) $CSG_{w/o\ RAP}$	67.4	5.3↓
6) $CSG_{w/o\ \mathcal{L}_{con}}$	68.4	4.3↓
7) CSG	69.2	3.5↓
DUET (Full model)	**72.7**	-

(f) Component analysis