CVPR22 – Open-Set

韩坤洋

Papers

- VGSE: Visually-Grounded Semantic Embeddings for Zero-Shot Learning
- Open-Vocabulary One-Stage Detection with Hierarchical Visual-Language Knowledge Distillation
- Decoupling Zero-Shot Semantic Segmentation
- Open-Vocabulary Instance Segmentation via Robust Cross-Modal Pseudo-Labeling
- ProposalCLIP: Unsupervised Open-Category Object Proposal Generation via Exploiting CLIP Cues
- Distinguishing Unseen from Seen for Generalized Zero-shot Learning

Open-set

- **Opposite to close-set**, which train on class set A, test on A.
- Anomaly Detection
  - Identify unexpected patterns. (K class → K+1 class)
- Zero-shot
  - Identify class of unseen object
  - Train with auxiliary information

- Open-world
  - Identify unknown object → human annotation → train with new class
- Open-vocabulary
  - Identify class of unseen object
  - Align embedding space between vision and text

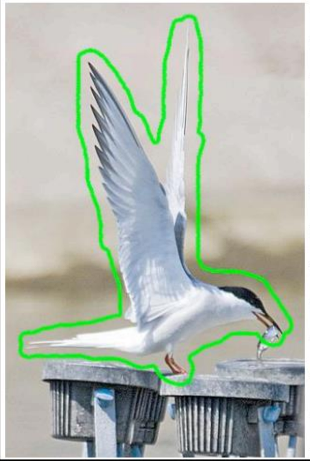
VGSE: Visually-Grounded Semantic Embeddings for Zero-Shot Learning

Wenjia Xu[1,7,8] Yongqin Xian[2] Jiuniu Wang[5,7,8] Bernt Schiele[3] Zeynep Akata[3,4,6]

[1] Beijing University of Posts and Telecommunications [2] ETH Zurich
[3] Max Planck Institute for Informatics [4] University of Tübingen [5] City University of Hong Kong
[6] Max Planck Institute for Intelligent Systems
[7] University of Chinese Academy of Sciences [8] Aerospace Information Research Institute, CAS

VGSE: Visually-Grounded Semantic Embeddings for Zero-Shot Learning

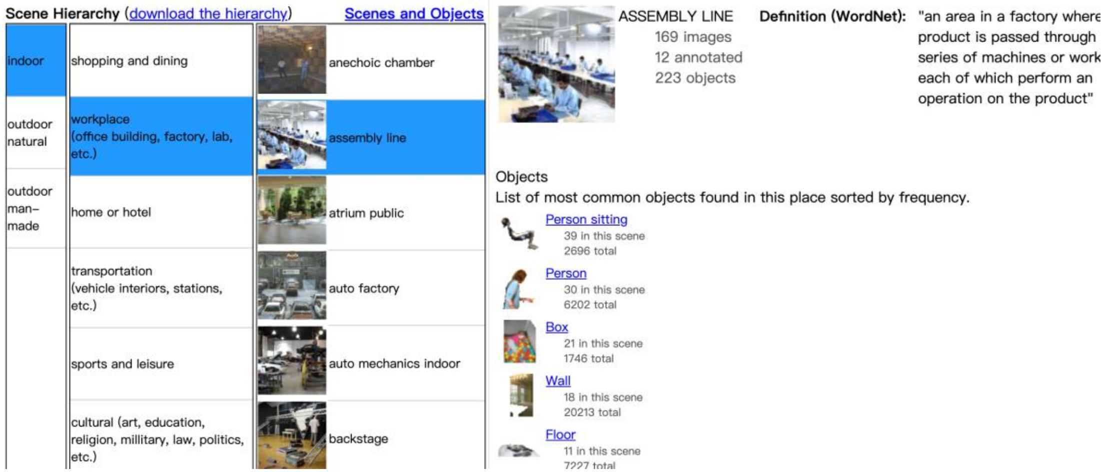
- Dataset
  - AWA2, 30_475 images, 50 animal classes, 40 seen
    - 85 numeric attribute values for each class
  - CUB, 11_788 images, 200 bird classes, 150 seen
    - 312 Binary Attributes
  - SUN, 14_340 images, 717 scene classes, 645 seen
    - 102 attributes



forehead_color	black	black	black
breast_pattern	solid	solid	solid
breast_color	white	white	white
head_pattern	plain	capped	plain
back_color	white	white	black
wing_color	grey/white	grey	white
leg_color	orange	orange	orange
size	medium	large	medium
bill_shape	needle	dagger	dagger
wing_shape	pointed	tapered	long
...	...	...	...
primary_color	white	white	white

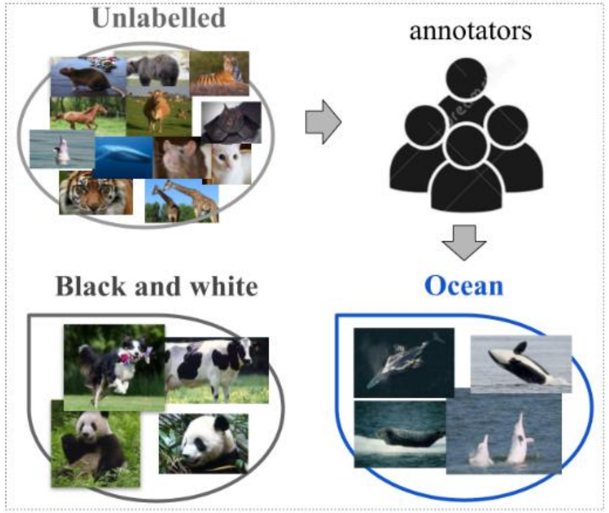
zebra

black:	yes
white:	yes
brown:	no
stripes:	yes
water:	no
eats fish:	no

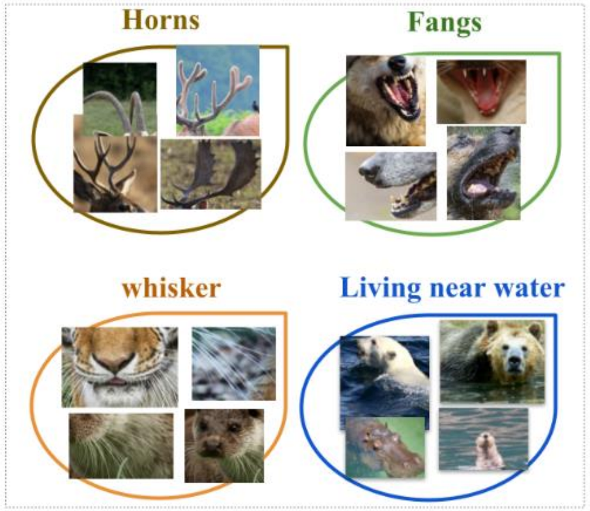


VGSE: Visually-Grounded Semantic Embeddings for Zero-Shot Learning

- Motivation
  - Semantic attributes annotation is labor-intensive. Recently, many works use word embedding instead.
  - **Word embedding not always reflect visual similarities**
  - Propose to discover semantic embeddings containing discriminative visual properties

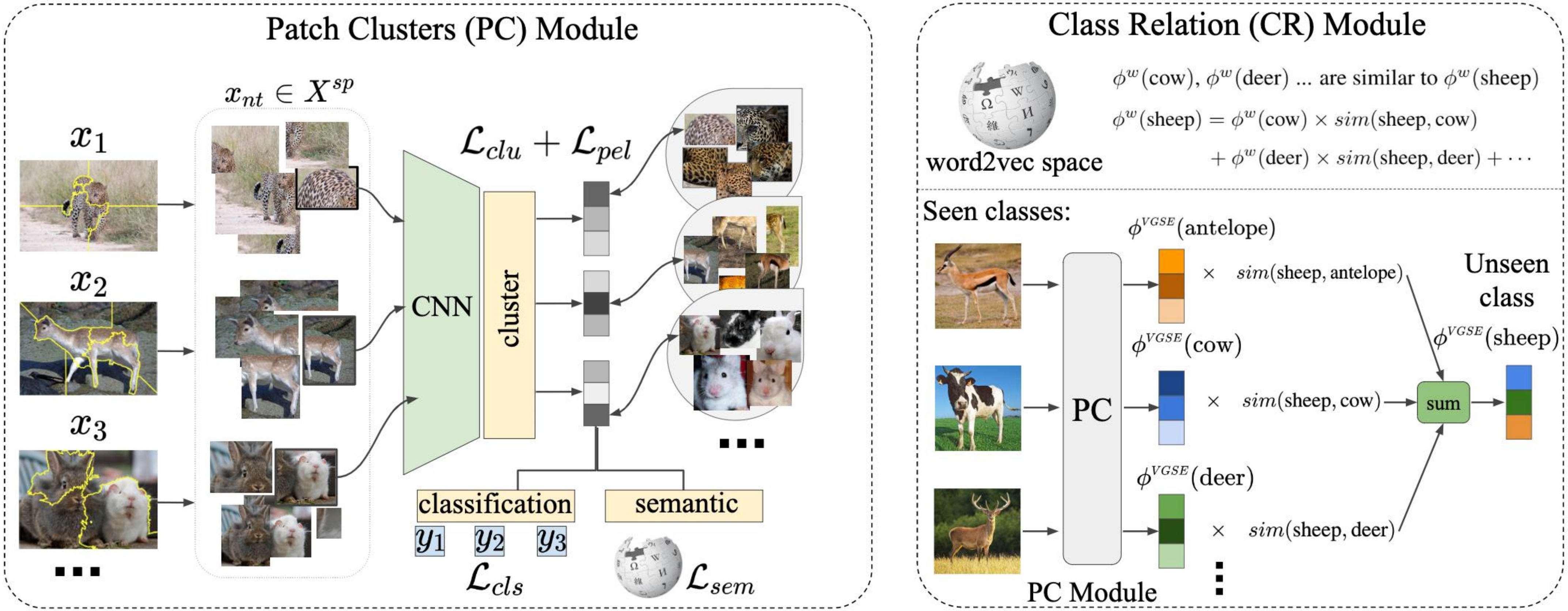


Human-Annotated Attributes

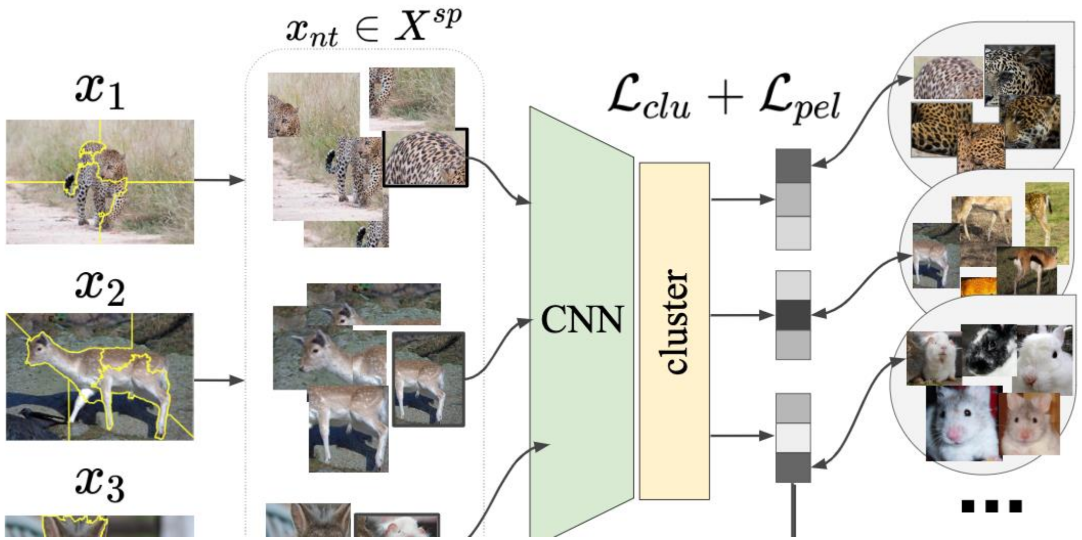


Semantic Embedding Discovery by VGSE

VGSE: Visually-Grounded Semantic Embeddings for Zero-Shot Learning



VGSE: Visually-Grounded Semantic Embeddings for Zero-Shot Learning



- Patch Clustering
  - Watershed segmentation algorithm.
    - Split into around 9 parts per image.
  - Cluster layer
    - Output k-dim vector, k clusters
  - Pretext task
    - Patch_i and its neighbors should in same cluster (L2 neighbor)
    - Avoid all image patches assigned to same cluster

$$\mathcal{L}_{clu} = -\sum_{x_{nt}\in X^{sp}} \sum_{x_i \in X^{sp}_{nb}} log(a_{nt}^T a_i)\,, \quad \mathcal{L}_{pel} = \sum_{k=1}^{D_v} \bar{a}_{nt}^k \log \bar{a}_{nt}^k\,, \quad \bar{a}_{nt}^k = \frac{1}{N_s N_t} \sum_{x_{nt}\in X^{sp}} a_{nt}^k\,,$$

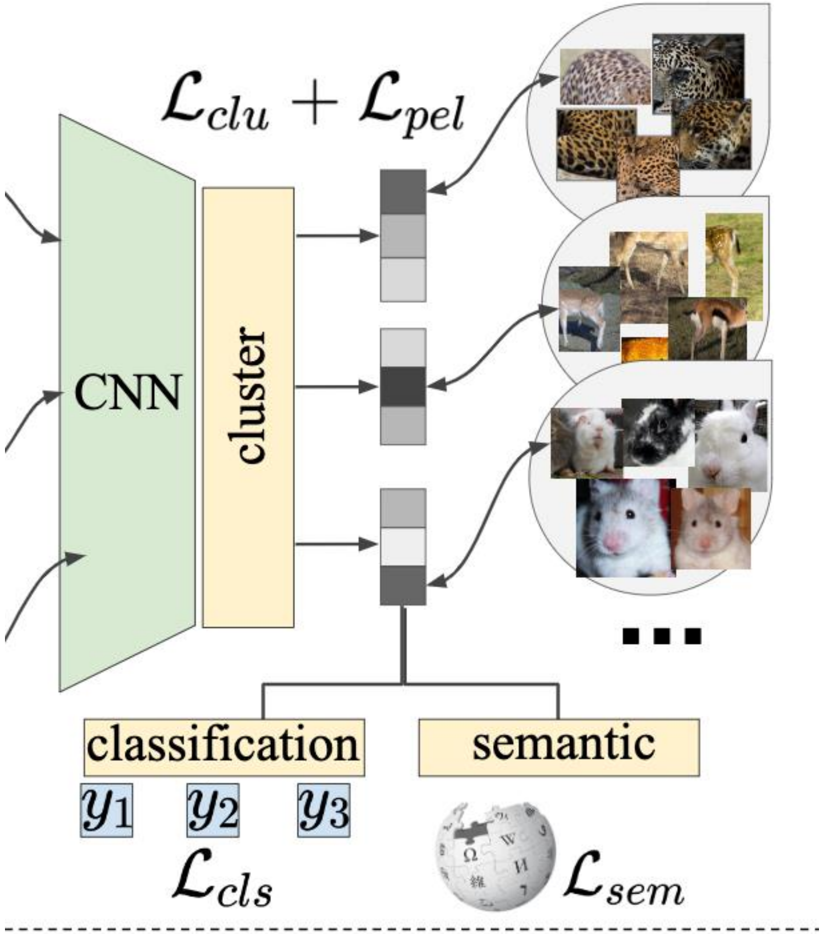
VGSE: Visually-Grounded Semantic Embeddings for Zero-Shot Learning

- Class discrimination
  - Cluster to class layer
    - CE loss, Seen classes

$$\mathcal{L}_{cls} = -\log \frac{\exp\left(p\left(y_n | x_{nt}\right)\right)}{\sum_{\hat{y} \in Y^s} \exp\left(p\left(\hat{y} | x_{nt}\right)\right)} .$$

- Semantic relatedness
  - Map cluster to semantic space
    - word2vec

$$\mathcal{L}_{sem} = \left\| S \circ a_{nt} - \phi^w(y_n) \right\|_2 ,$$



VGSE: Visually-Grounded Semantic Embeddings for Zero-Shot Learning

- Seen semantic embedding
  - Get image embedding by averaging all patch

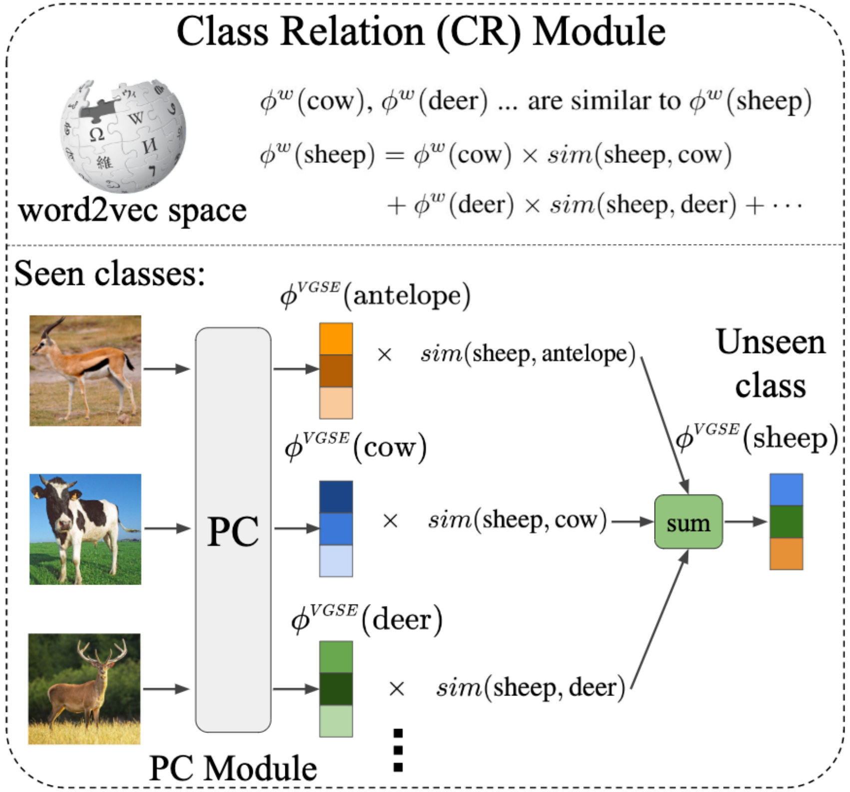
$$a_n = \frac{1}{N_t} \sum_{t=1}^{N_t} a_{nt} \, .$$

  - Get semantic embedding for class y_n by averaging all image belong to y_n

$$\phi^{VGSE}(y_n) = \frac{1}{|I_i|} \sum_{j \in I_i} a_j$$

VGSE: Visually-Grounded Semantic Embeddings for Zero-Shot Learning

- Class Relation Module
  - Formulate the similarity between seen and unseen
  - Use word2vec as external knowledge
- Solution
  - (1) directly averaging semantic embedding
  - (2) optimizing a similarity matrix



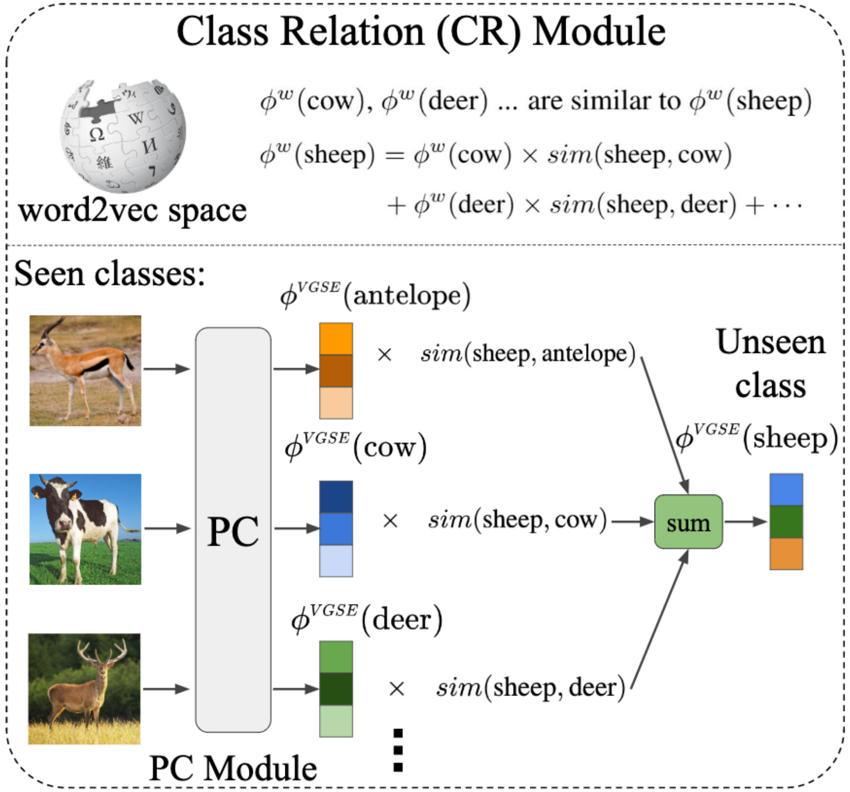
VGSE: Visually-Grounded Semantic Embeddings for Zero-Shot Learning

- Weighted Average
  - Retrieve nearest class neighbors in seen class
    - L2 distance over w2v embedding space

  - Semantic embedding:

$$\phi^{VGSE}(y_m) = \frac{1}{|Y^s_{nb}|} \sum_{\tilde{y} \in Y^s_{nb}} sim\left(y_m, \tilde{y}\right) \cdot \phi^{VGSE}(\tilde{y}) \,,$$

$$sim(y_m, \tilde{y}) = \exp(-\eta \left\|\phi^{w}\left(y_m\right) - \phi^{w}\left(\tilde{y}\right)\right\|_2) \,,$$



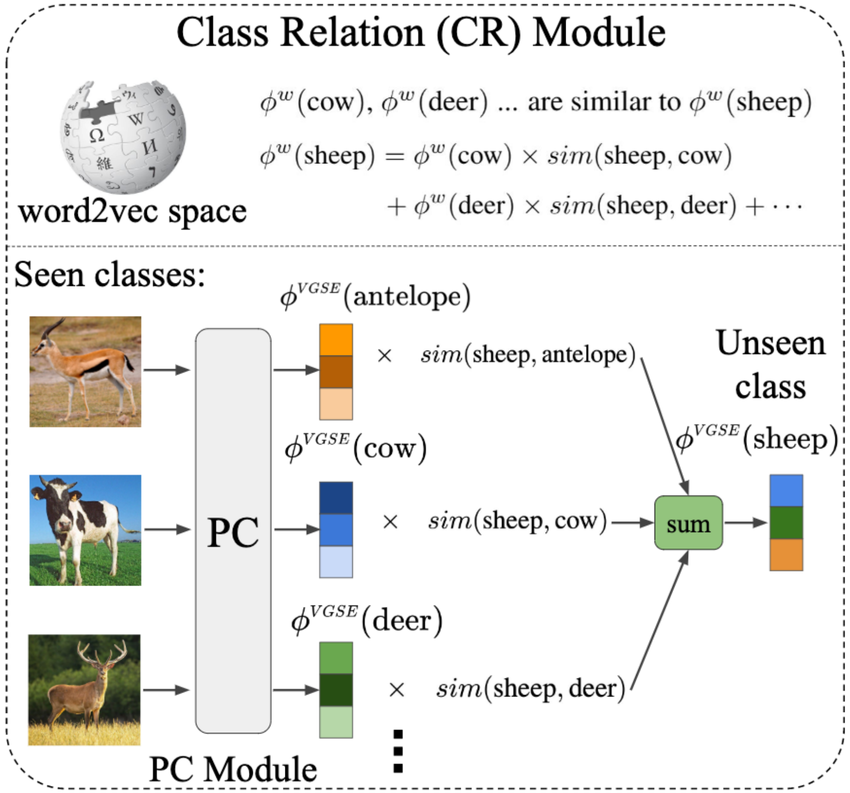
VGSE: Visually-Grounded Semantic Embeddings for Zero-Shot Learning

- Similarity Matrix Optimization
  - Learn a similarity matrix
  - y_m, unseen class
  - r_i, similarity between y_m and i-th seen class

$$\min_{r} \left\|\phi^{w}(y_m) - r^T\phi^{w}(Y^s)\right\|_2$$

$$\text{s.t.} \quad \alpha < r < 1 \quad and \quad \sum_{i=1}^{|Y^s|} r_i = 1.$$

*lower bound is 0 or -1



VGSE: Visually-Grounded Semantic Embeddings for Zero-Shot Learning

	ZSL Model	Semantic Embeddings	Zero-Shot Learning			Generalized Zero-Shot Learning								
			AWA2	CUB	SUN	AWA2			CUB			SUN		
			T1	T1	T1	u	s	H	u	s	H	u	s	H
Generative	CADA-VAE [43]	w2v [31]	49.0	22.5	37.8	38.6	60.1	47.0	16.3	39.7	23.1	26.0	28.2	27.0
		VGSE-SMO (Ours)	52.7	24.8	40.3	46.9	61.6	53.9	18.3	44.5	25.9	**29.4**	29.6	29.5
	f-VAEGAN-D2 [61]	w2v [31]	58.4	32.7	39.6	46.7	59.0	52.2	23.0	44.5	30.3	25.9	33.3	29.1
		VGSE-SMO (Ours)	61.3	**35.0**	**41.1**	45.7	66.7	54.2	**24.1**	45.7	**31.5**	25.5	**35.7**	**29.8**
Non-Generative	SJE [2]	w2v [31]	53.7	14.4	26.3	39.7	65.3	48.8	13.2	28.6	18.0	19.8	18.6	19.2
		VGSE-SMO (Ours)	62.4	26.1	35.8	46.8	72.3	56.8	16.4	44.7	28.3	28.7	25.2	26.8
	GEM-ZSL [28]	w2v [31]	50.2	25.7	-	40.1	80.0	53.4	11.2	**48.8**	18.2	-	-	-
		VGSE-SMO (Ours)	58.0	29.1	-	49.1	78.2	60.3	13.1	43.0	20.0	-	-	-
	APN [62]	w2v [31]	59.6	22.7	23.6	41.8	75.0	53.7	17.6	29.4	22.1	16.3	15.3	15.8
		VGSE-SMO (Ours)	**64.0**	28.9	38.1	**51.2**	**81.8**	**63.0**	21.9	45.5	29.5	24.1	31.8	27.4

Table 1. Comparing our VGSE-SMO, with w2v semantic embedding over state-of-the-art ZSL models. In ZSL, we measure Top-1 accuracy (**T1**) on unseen classes, in GZSL on seen/unseen (**s/u**) classes and their harmonic mean (**H**). Feature Generating Methods, i.e., f-VAEGAN-D2, and CADA-VAE generating synthetic training samples, and SJE, APN, GEM-ZSL using only real image features.

VGSE: Visually-Grounded Semantic Embeddings for Zero-Shot Learning

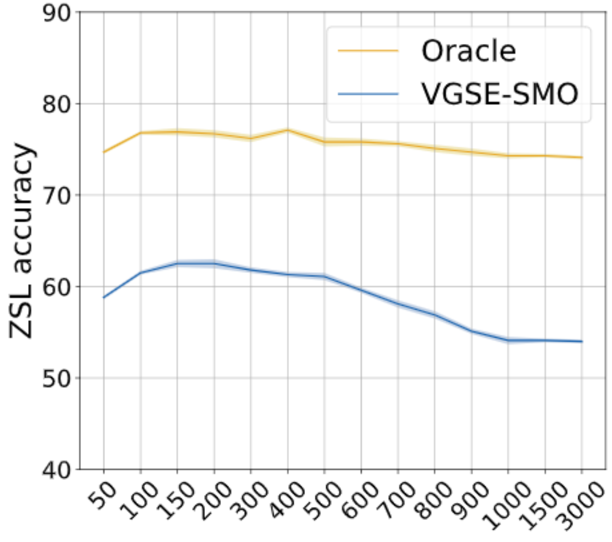
Semantic Embeddings	External knowledge	Zero-shot learning		
		AWA2	**CUB**	**SUN**
w2v [31]	w2v	58.4	32.7	39.6
ZSLNS [39]	T	57.4	27.8	-
GAZSL [67]	T	-	34.4	-
Auto-dis [3]	T	52.0	-	-
CAAP [5]	T and H	55.3	31.9	35.5
VGSE-SMO (Ours)	w2v	**61.3** $\pm$ 0.3	**35.0** $\pm$ 0.2	**41.1** $\pm$ 0.3

Table 2. Comparing with state-of-the-art methods for learning semantic embeddings with less human annotation (T: online textual articles, H: human annotation) using same image features and ZSL model (f-VAEGAN-d2 [61]).

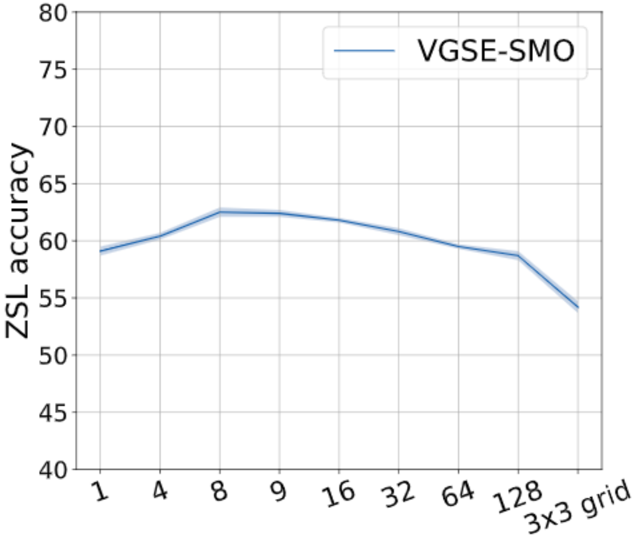
Semantic Embeddings	Zero-shot learning		
	AWA2	CUB	SUN
k-means-SMO	54.5 $\pm$ 0.4	15.0 $\pm$ 0.5	25.2 $\pm$ 0.4
ResNet-SMO	55.3 $\pm$ 0.2	15.4 $\pm$ 0.1	25.1 $\pm$ 0.1
$\mathcal{L}_{clu} + \mathcal{L}_{pel}$ (baseline + SMO)	56.6 $\pm$ 0.2	16.7 $\pm$ 0.2	26.3 $\pm$ 0.3
$+\mathcal{L}_{cls}$	61.2 $\pm$ 0.1	23.7 $\pm$ 0.2	30.5 $\pm$ 0.2
$+\mathcal{L}_{sem}$ (VGSE-SMO)	**62.4** $\pm$ 0.3	**26.1** $\pm$ 0.3	**35.8** $\pm$ 0.2
VGSE-WAvg	57.7 $\pm$ 0.2	25.8 $\pm$ 0.3	35.3 $\pm$ 0.2

Table 3. Ablation study over the PC module reporting ZSL T1 on AWA2, CUB, and SUN (mean accuracy and std over 5 runs). The baseline is the PC module with the cluster loss $\mathcal{L}_{clu}$ and $\mathcal{L}_{pel}$. Our full model VGSE-SMO is trained with two additional losses $\mathcal{L}_{cls}$, $\mathcal{L}_{sem}$. Two kinds of semantic embeddings learned from k-means clustering and pretrained ResNet are listed below for comparison.

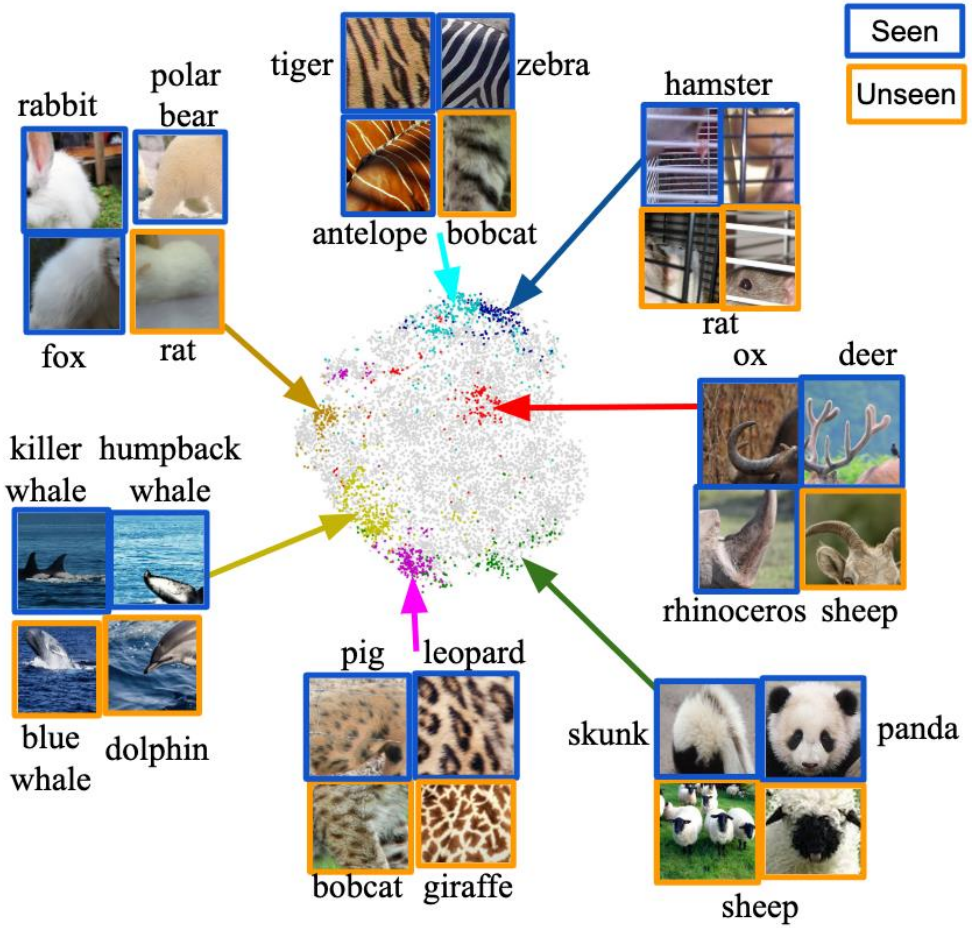
VGSE: Visually-Grounded Semantic Embeddings for Zero-Shot Learning



(a) cluster number D_v



(b) patch number N_t



Open-Vocabulary One-Stage Detection with Hierarchical Visual-Language Knowledge Distillation

Zongyang Ma[1,2], Guan Luo[1,2], Jin Gao[1,2,†], Liang Li[3,†], Yuxin Chen[1,2], Shaoru Wang[1,2]
Congxuan Zhang[4], and Weiming Hu[1,2,5]

[1]NLPR, Institute of Automation, Chinese Academy of Sciences

[2]School of Artificial Intelligence, University of Chinese Academy of Sciences

[3]Brain Science Center, Beijing Institute of Basic Medical Sciences [4]Nanchang Hangkong University

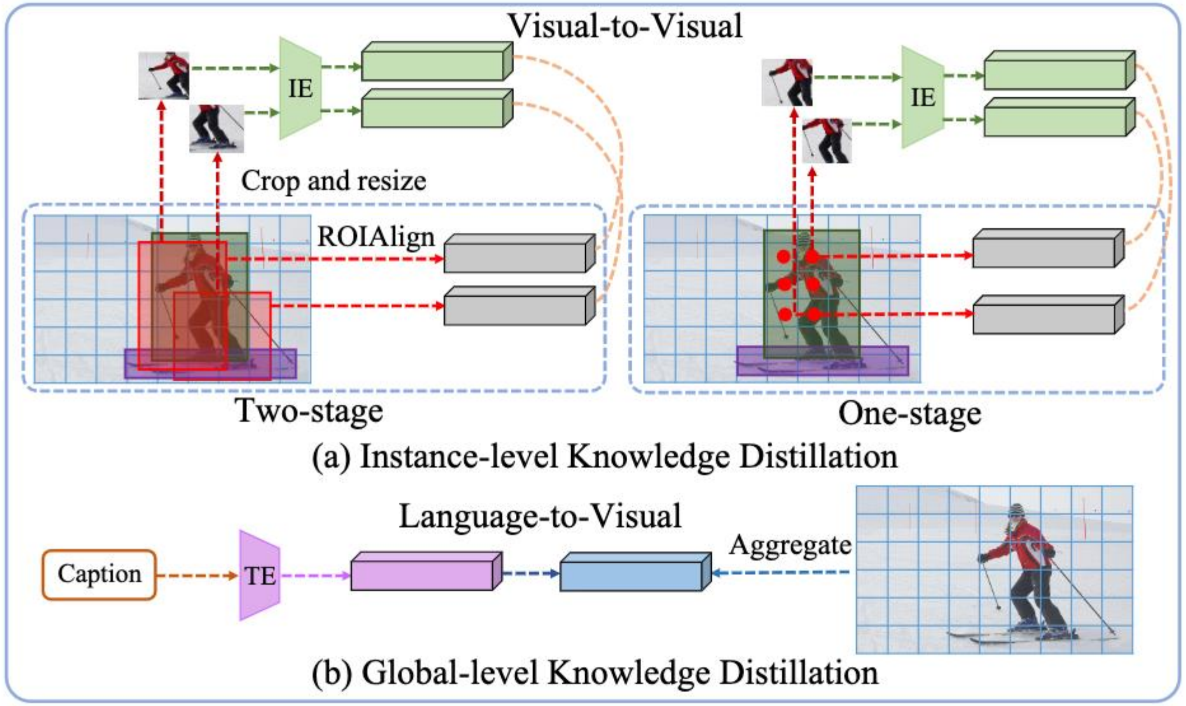
[5]CAS Center for Excellence in Brain Science and Intelligence Technology

mazongyang2020@ia.ac.cn, {gluo,jin.gao}@nlpr.ia.ac.cn, liang.li.brain@aliyun.com

Open-Vocabulary One-Stage Detection with Hierarchical Visual-Language Knowledge Distillation

- Motivation
  - One-stage detector is more efficient
  - Absence of class-agnostic object proposals ←→ Find unseen objects

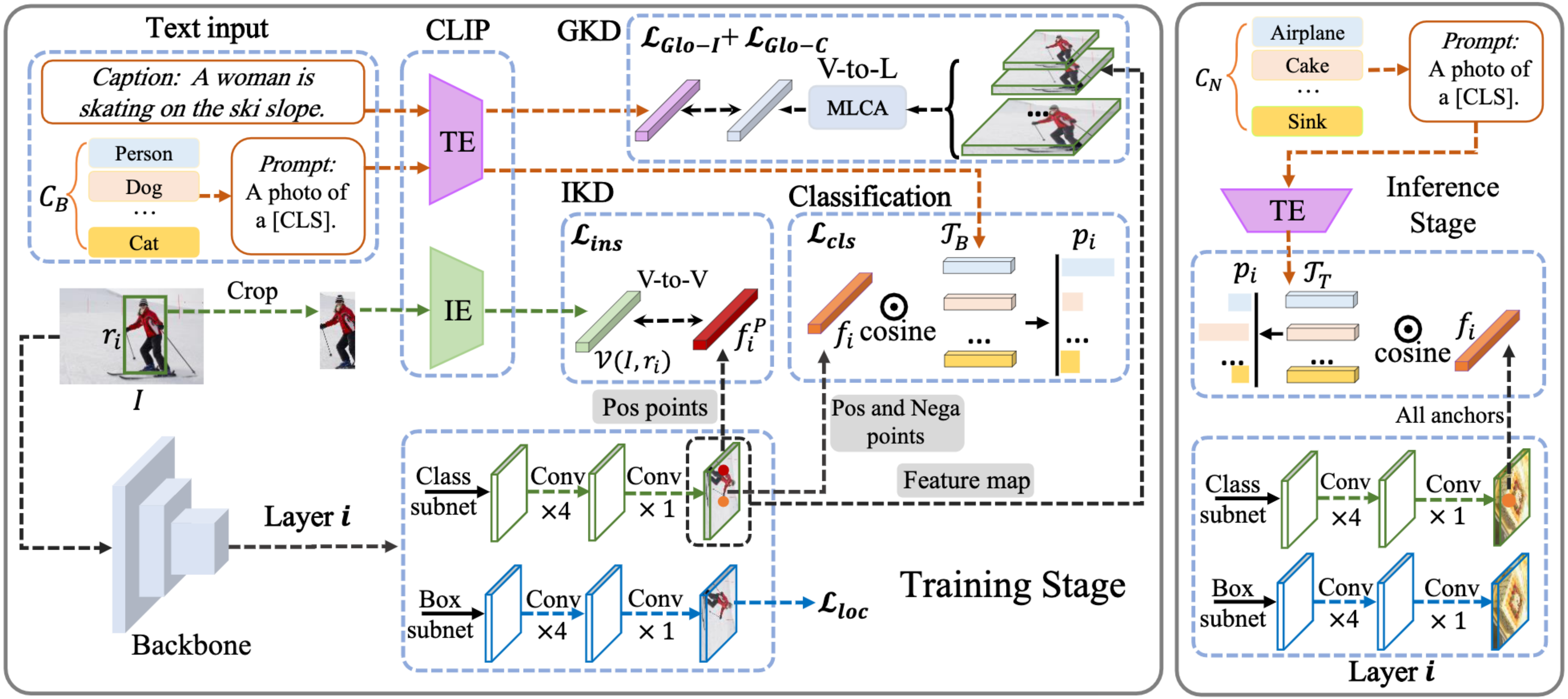
  - Open-vocabulary one-stage detection



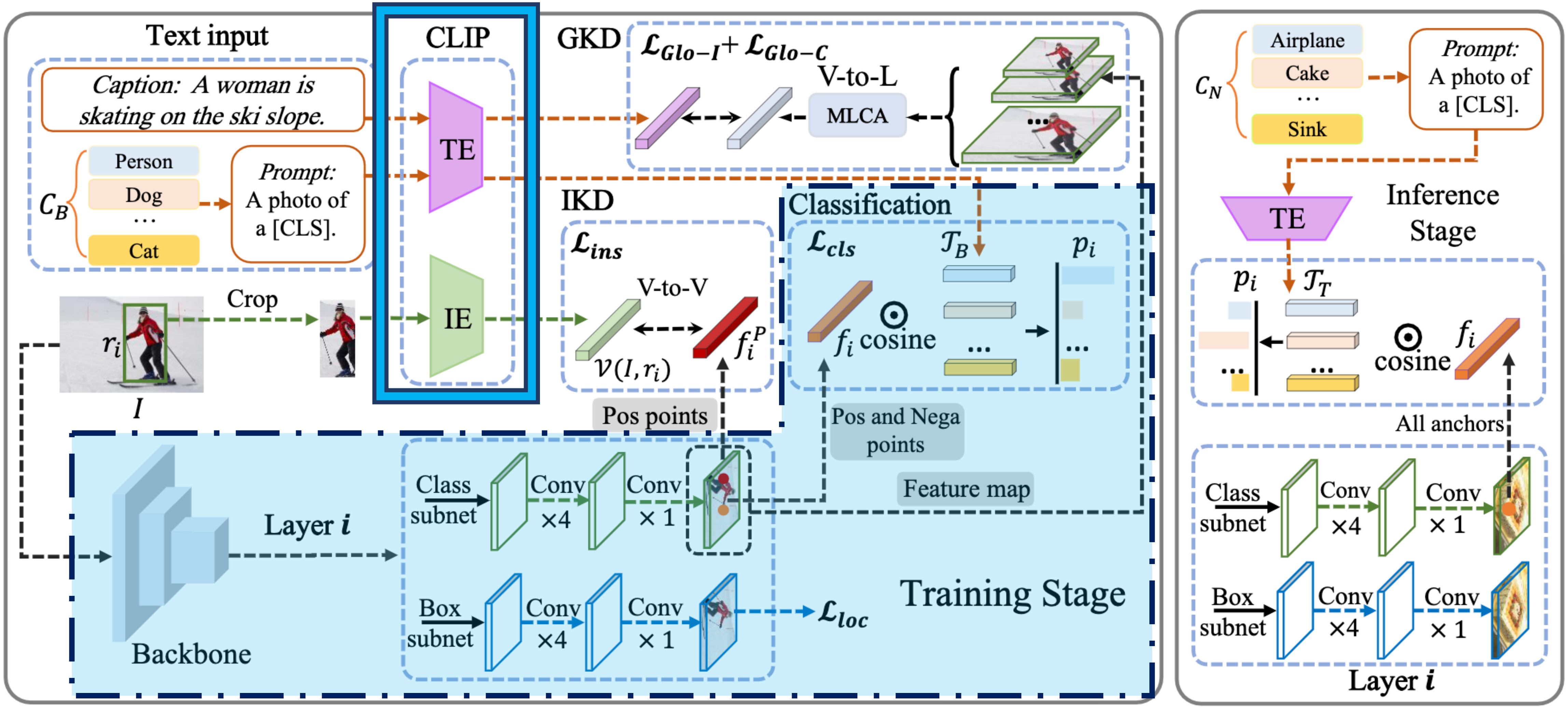
(a) Instance-level Knowledge Distillation

(b) Global-level Knowledge Distillation

Open-Vocabulary One-Stage Detection with Hierarchical Visual-Language Knowledge Distillation

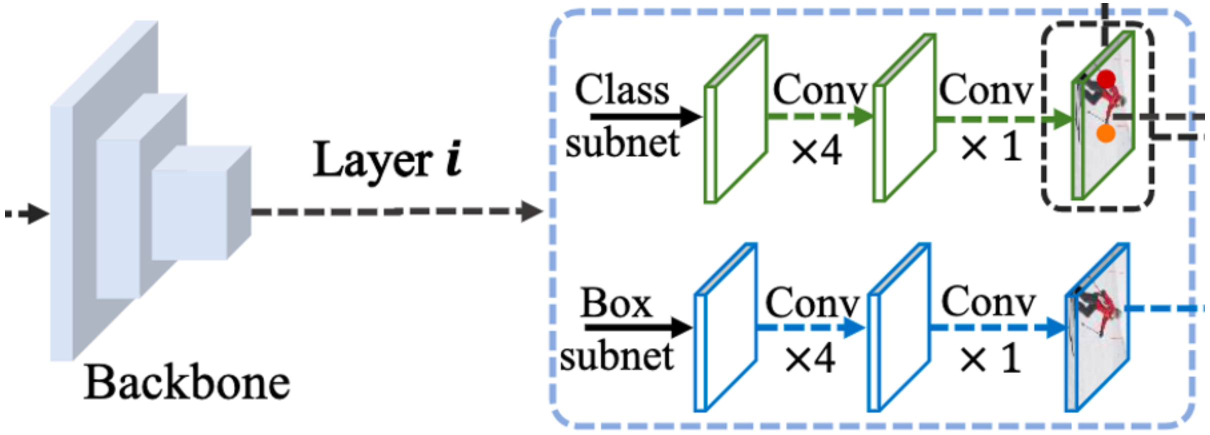


Open-Vocabulary One-Stage Detection with Hierarchical Visual-Language Knowledge Distillation



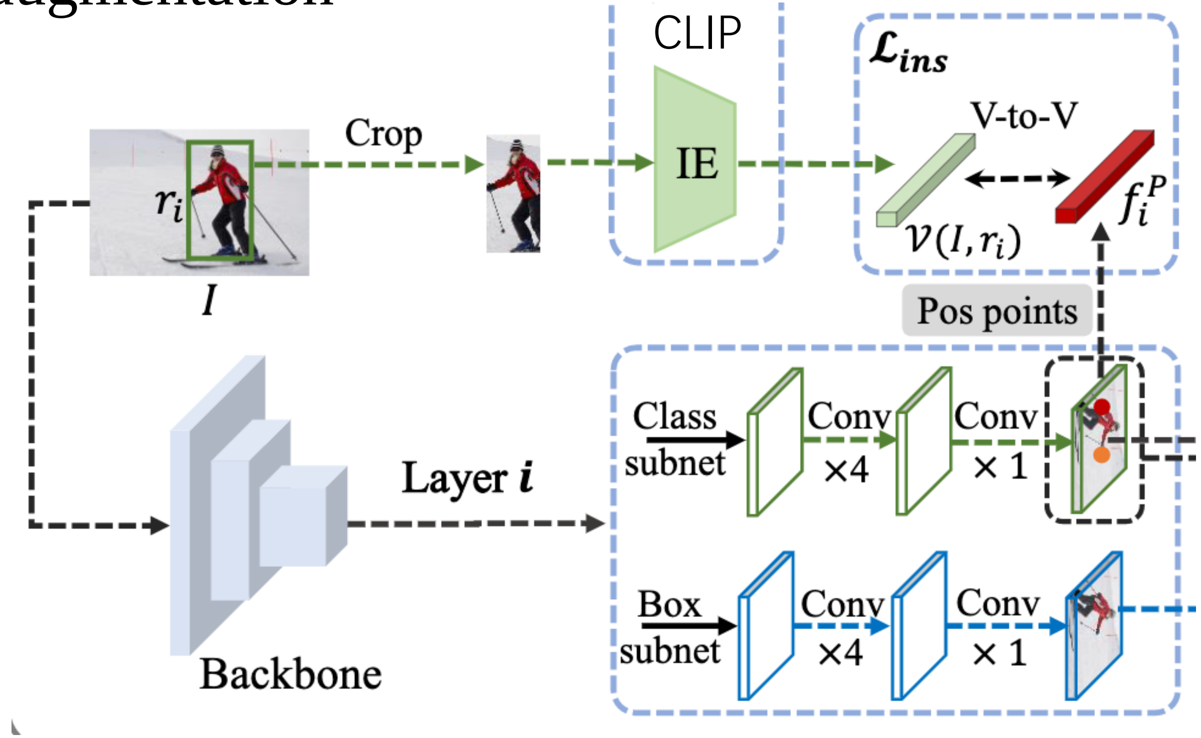
Open-Vocabulary One-Stage Detection with Hierarchical Visual-Language Knowledge Distillation

- Network
  - Based on ATSS, which based on FCOS
    - Per-pixel prediction
  - Class subnet, (hid_dim, cls_num) → (hid_dim, emb_dim)



Open-Vocabulary One-Stage Detection with Hierarchical Visual-Language Knowledge Distillation

- IKD
  - Select positive samples
  - Crop image based on box pred, data augmentation
  - L1 loss supervision

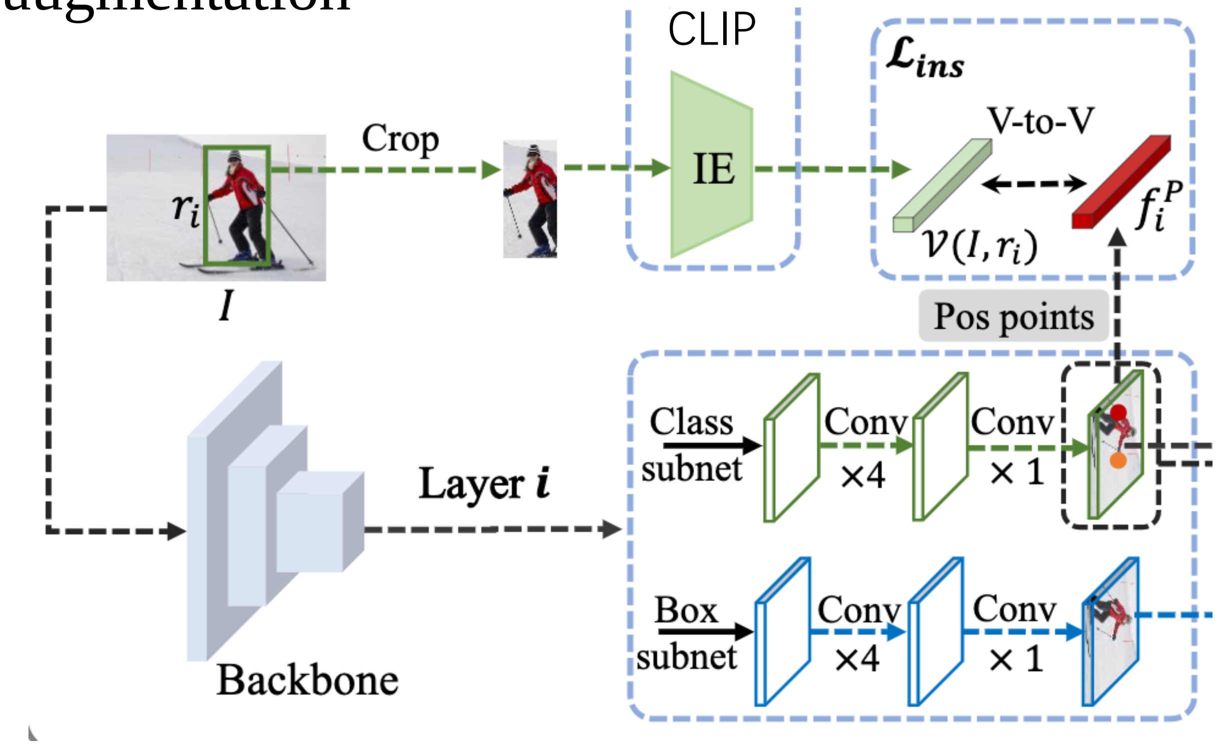


Open-Vocabulary One-Stage Detection with Hierarchical Visual-Language Knowledge Distillation

- IKD
  - Select positive samples
  - Crop image based on box pred, data augmentation
  - L1 loss supervision

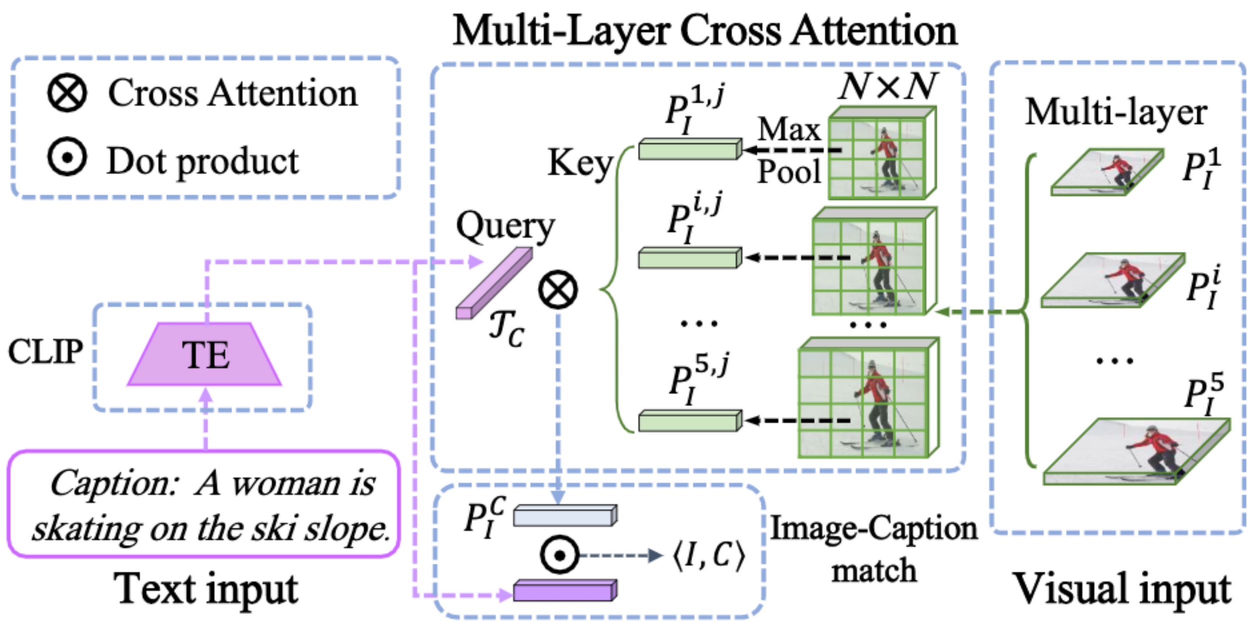
Norm	Weight	Region	Area	AR_{50}	AP_{50}
L_1	1	*pred*	1×	62.4	14.6
L_2	1	*pred*	1×	**65.1**	12.8
L_2	10	*pred*	1×	63.6	14.6
L_1	1	*GT*	1×	62.8	14.5
L_1	1	*pred*	1.5×	64.5	**15.3**

Table 2. Comparisons between different sub-module options in IKD. *pred* and *GT* mean cropping regions from prediction boxes and ground-truth boxes, respectively. 1× and 1.5× represent cropping the original box and its 1.5× center expansion respectively.



Open-Vocabulary One-Stage Detection with Hierarchical Visual-Language Knowledge Distillation

- GKD
  - Whole caption is sent into CLIP text encoder
  - Full image feature (5 FPN layers, $\frac{HW}{NN}$ patches), pool into vector
  - Query: text, Key: vision
  - Caption-Image score
    - <Text, Output>
  - N*N, Contrastive loss



Open-Vocabulary One-Stage Detection with Hierarchical Visual-Language Knowledge Distillation

- GKD
  - Whole caption is sent into CLIP text encoder
  - Full image feature (5 FPN layers, $\frac{HW}{NN}$ patches), pool into vector
  - Query: text, Key: vision
  - Caption-Image score
    - <Text, Output>
  - N*N, Contrastive loss

Patch	*Pool*	*Loss*	*bs/gpu*	AR_{50}	AP_{50}
4	*Ave*	*CL*	*8*	59.2	12
4	*Max*	*CL*	*8*	64.2	20.1
3	*Max*	*CL*	*8*	61.1	**20.7**
8	*Max*	*CL*	*8*	60.8	13.7
3	*Max*	*PL*	*8*	60.9	17.9
3	*Max*	*CL*	*4*	**65.6**	20.5

Table 3. Comparisons between different sub-module options in GKD. *Ave* and *Max* represent using Average Pooling and Max Pooling to obtain patch features respectively. *CL* denotes training with the contrastive learning loss, while *PL* only considers the cosine similarities between positive pairs. *bs/gpu* is the batch size on each GPU during training.

Open-Vocabulary One-Stage Detection with Hierarchical Visual-Language Knowledge Distillation

- Negative samples
  - Sampling negative samples will boost the performance
  - Identify novel foreground region as background

  - Sampling 10% negative samples

Negative samples	IKD	GKD	Base AR_{50}	Base AP_{50}	Novel AR_{50}	Novel AP_{50}
1:1	√		71.0	37.0	**63.0**	**16.8**
10%	√		**75.9**	44.3	62.4	14.6
100%	√		74.5	**44.4**	60.3	9.0
1:1		√	69.2	34.9	60.2	19.3
10%		√	**74.0**	**42.7**	**61.1**	**20.7**
100%		√	72.4	42.6	56.4	18.7

Table 5. Comparisons between different sampling strategies for negative samples. 1:1, 10%, 100% mean sampling the same number of negative samples as the positive samples, sampling 10 % of the negative samples, and using all the negative samples.

Open-Vocabulary One-Stage Detection with Hierarchical Visual-Language Knowledge Distillation

Method			Base/Novel	ZSD	GZSD		
				Novel	Base	Novel	All
TS	ZS	SB [1]	48/17	0.70	29.2	0.31	24.9
		LAB [1]	48/17	0.27	20.8	0.22	18.0
		DESE [1]	48/17	0.54	26.7	0.27	22.1
		BLC [39]	48/17	9.9	42.1	4.50	32.3
		ZSI* [40]	48/17	11.4	46.5	4.83	35.6
	OV	OVR-CNN [36]	48/17	16.7	-	-	34.3
		ViLD* [8]	48/17	-	59.5	27.6	51.3
OS	ZS	PL* [25]	48/17	10.0	35.9	4.12	27.9
		DELO [42]	48/17	7.6	13.8	3.41	13.0
	OV	ZSD-YOLO* [35]	48/17	13.4	31.7	13.6	27.0
		HierKD(ours)	48/17	**25.3**	**51.3**	**20.3**	**43.2**
TS	ZS	BLC [39]	65/15	13.1	36.0	13.1	31.7
		ZSI* [40]	65/15	13.6	38.7	13.6	34.0
OS	ZS	PL* [25]	65/15	12.4	34.1	12.4	30.0
	OV	ZSD-YOLO* [35]	65/15	18.3	31.7	17.9	29.2
		HierKD(ours)	65/15	**27.4**	**48.9**	**20.4**	**43.6**

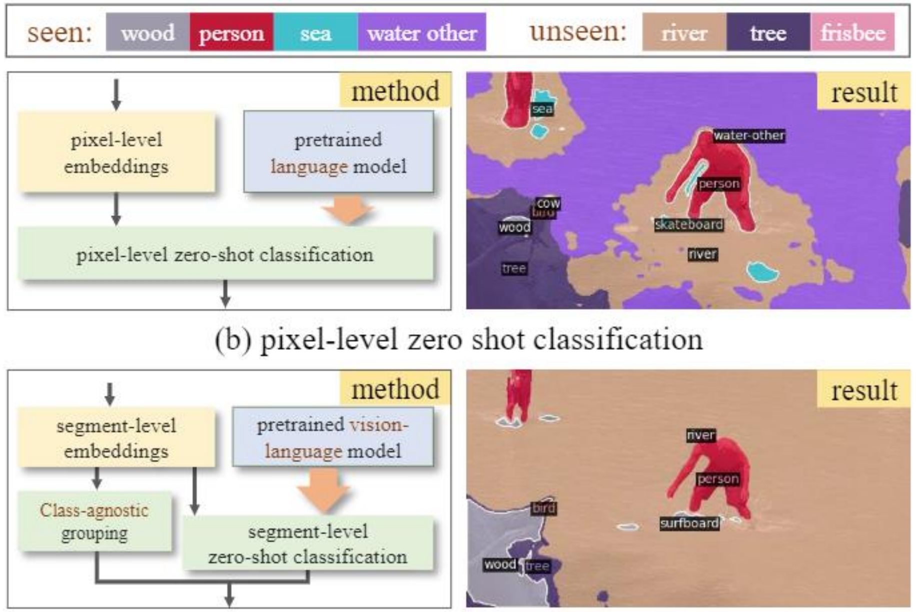
Table 7. **Comparison with other state-of-the-art methods:** * denotes the state-of-the-art methods in various settings. "TS" and "OS" are abbreviation of two-stage and one-stage detectors, respectively. Note that we classify Cascade R-CNN based detectors as generalized two-stage methods. "ZS" and "OV" indicate that the models belong to zero-shot and open-vocabulary detectors, respectively.

Decoupling Zero-Shot Semantic Segmentation

Jian Ding[1,2], Nan Xue[1], Gui-Song Xia[1,*], Dengxin Dai[2]
[1]CAPTAIN, Wuhan University, China [2]MPI for Informatics, Germany
{jian.ding, xuenan, guisong.xia}@whu.edu.cn, ddai@mpi-inf.mpg.de

Decoupling Zero-Shot Semantic Segmentation

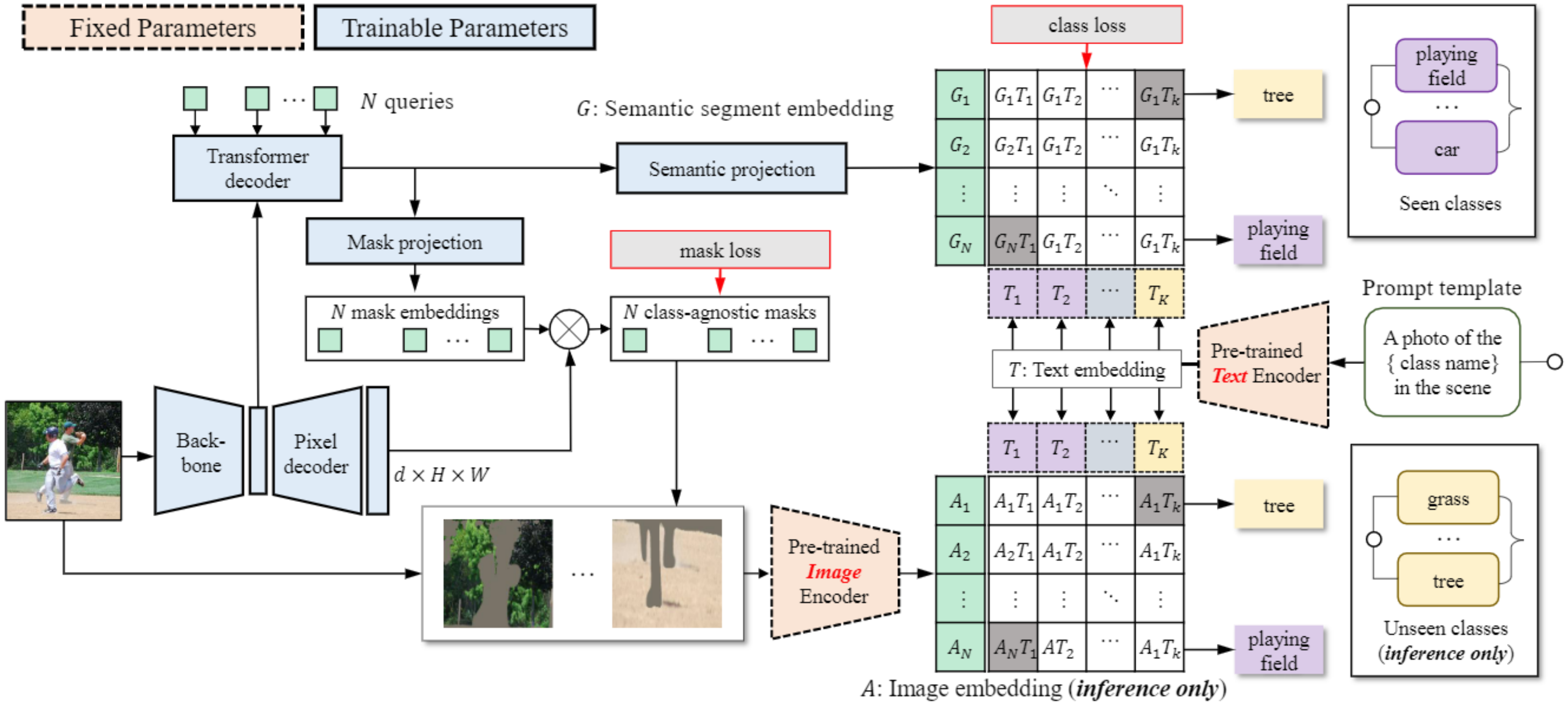
- Motivation
  - Pixel-level zero-shot classification problem ✗
    - Limited capability to integrate vision-language models
  - Decoupling Zero-Shot Segmentation ☑
    - Class-agnostic grouping task
    - Zero-shot classification task



(b) pixel-level zero shot classification

(c) decoupled zero shot segmentation (ours)

Decoupling Zero-Shot Semantic Segmentation



Decoupling Zero-Shot Semantic Segmentation

(3) *ZegFormer*. This variant is our full model. We first fuse $p_q(c)$ and $p_q^{'}(c)$ for each query as:

$$p_{q,\text{fusion}}(c) = \begin{cases} p_q(c)^{\lambda} \cdot p_{q,\text{avg}}^{(1-\lambda)} & \text{if } c \in S \\ p_q(c)^{(1-\lambda)} \cdot p_q^{'}(c)^{\lambda} & \text{if } c \in U, \end{cases} \quad (3)$$

	preprocess	Seen	Unseen	Harmonic
ZegFormer-seg	-	**37.4**	21.4	27.2
ZegFormer	crop	36.6	19.7	25.6
	mask	36.0	31.0	33.3
	crop and mask	35.9	**33.1**	**34.4**

Figure 3. Comparison between three preprocess for a segment.

Open-Vocabulary Instance Segmentation via Robust Cross-Modal Pseudo-Labeling

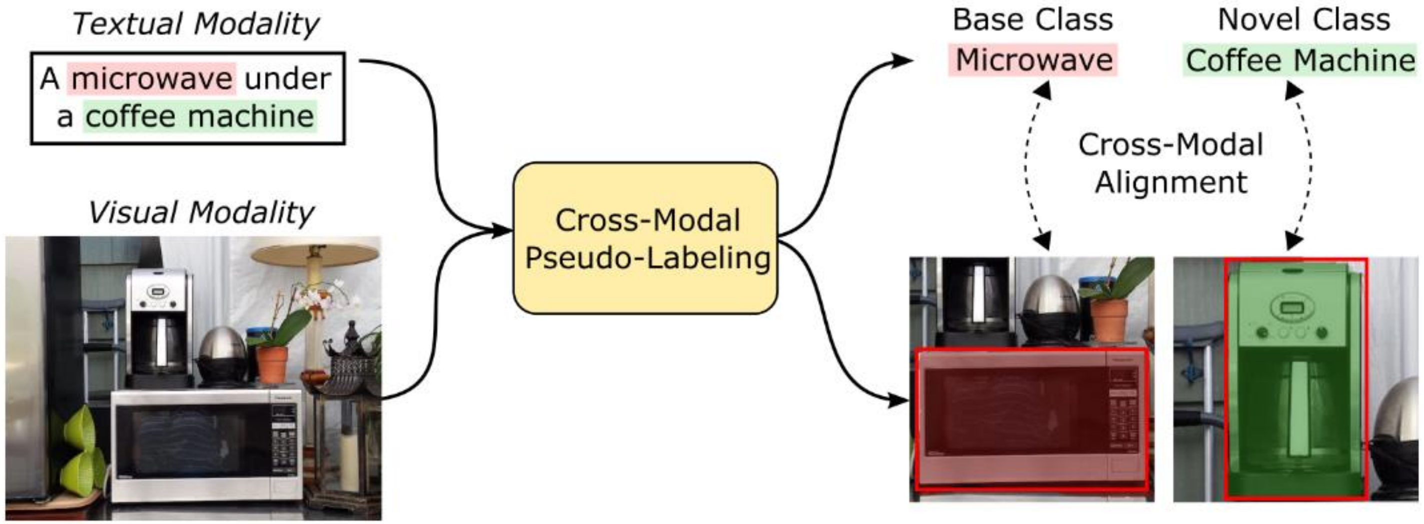
Dat Huynh[1*] Jason Kuen[2] Zhe Lin[2] Jiuxiang Gu[2] Ehsan Elhamifar[1]
[1]Northeastern University [2]Adobe Research
[1]{huynh.dat,e.elhamifar}@northeastern.edu [2]{kuen, zlin, jigu}@adobe.com

Instance segmentation, Open-vocabulary, **Weak supervision**

Open-Vocabulary Instance Segmentation via Robust Cross-Modal Pseudo-Labeling

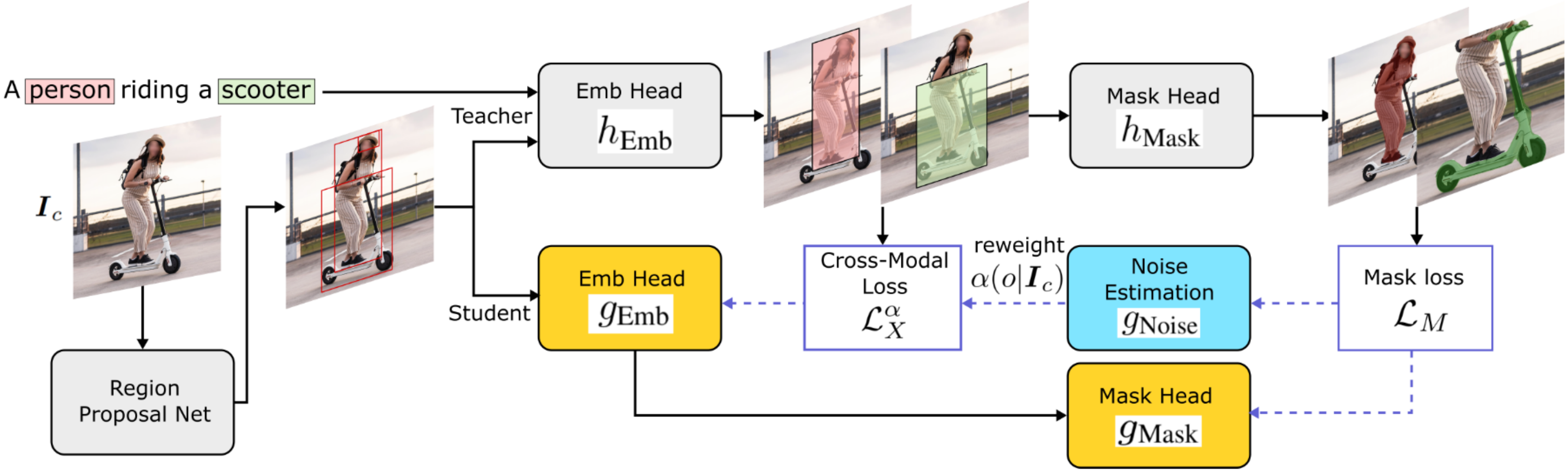
- Motivation
  - Aims at segment novel classes without mask annotations
  - Caption can’t provide details required in pixel-wise segmentation

  - Propose a cross-modal pseudo-labeling framework



Open-Vocabulary Instance Segmentation via Robust Cross-Modal Pseudo-Labeling

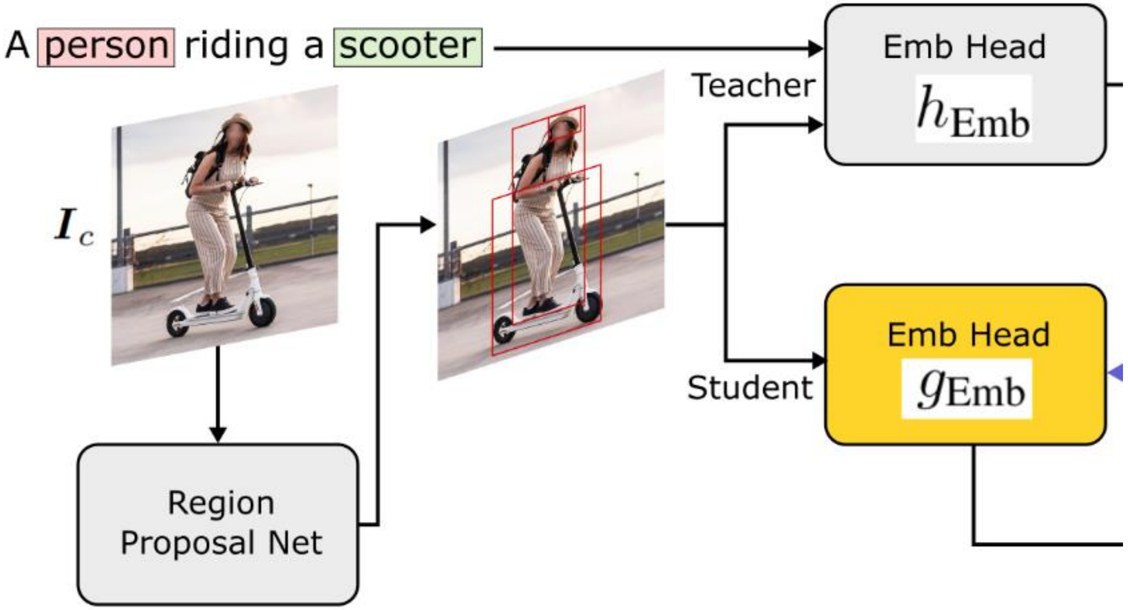
- Framework



Open-Vocabulary Instance Segmentation via Robust Cross-Modal Pseudo-Labeling

- Input
  - Region proposals (ROI Align)
  - Nouns from image caption
- Target
  - Mask
  - Class (Embedding similarity)

*background embedding is full of 0



Open-Vocabulary Instance Segmentation via Robust Cross-Modal Pseudo-Labeling

- Mask supervision
  - ~~Simple BCE~~, bad teacher prediction

$$\sum_{o \in \mathcal{O}_c} \sum_{x,y} \mathcal{L}_{\text{BCE}}\big(\boldsymbol{M}_o^{xy} | g_{\text{Mask}}^{xy}(\boldsymbol{f}_{\boldsymbol{b}_o})\big),$$

  - Assumption: pseudo mask is corrupted by a Gaussian noise, var can be estimated

$$\mathcal{L}_M(\mathcal{Y}_c | \boldsymbol{I}_c, g) = \sum_{o \in \mathcal{O}_c} \sum_{x,y} \mathcal{L}_{\text{BCE}}\big(\boldsymbol{M}_o^{xy} | g_{\text{Mask}}^{xy}(\boldsymbol{f}_{\boldsymbol{b}_o}) + \epsilon_o^{xy}\big)$$

$$\epsilon_o^{xy} \sim \mathcal{N}\big(0, g_{\text{Noise}}^{xy}(\boldsymbol{f}_{\boldsymbol{b}_o})\big),$$

*Noise is per-pixel predicted

*Noise is predicted based on object region

ProposalCLIP: Unsupervised Open-Category Object Proposal Generation via Exploiting CLIP Cues

Hengcan Shi, Munawar Hayat, Yicheng Wu, Jianfei Cai
Department of Data Science and AI, Monash University, Australia
{hengcan.shi, munawar.hayat, yicheng.wu, jianfei.cai}@monash.edu

ProposalCLIP: Unsupervised Open-Category Object Proposal Generation via Exploiting CLIP Cues

- Motivation
  - Require a large number of bounding box annotations
  - Can only generate proposals for limited object categories

ProposalCLIP: Unsupervised Open-Category Object Proposal Generation via Exploiting CLIP Cues

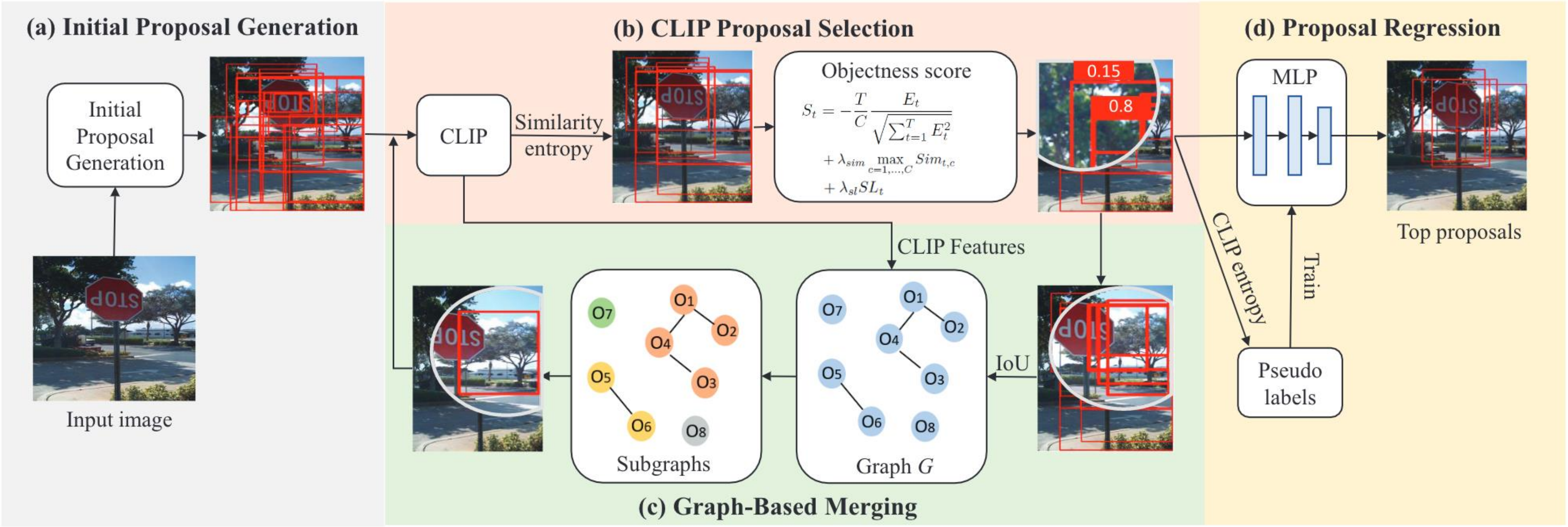
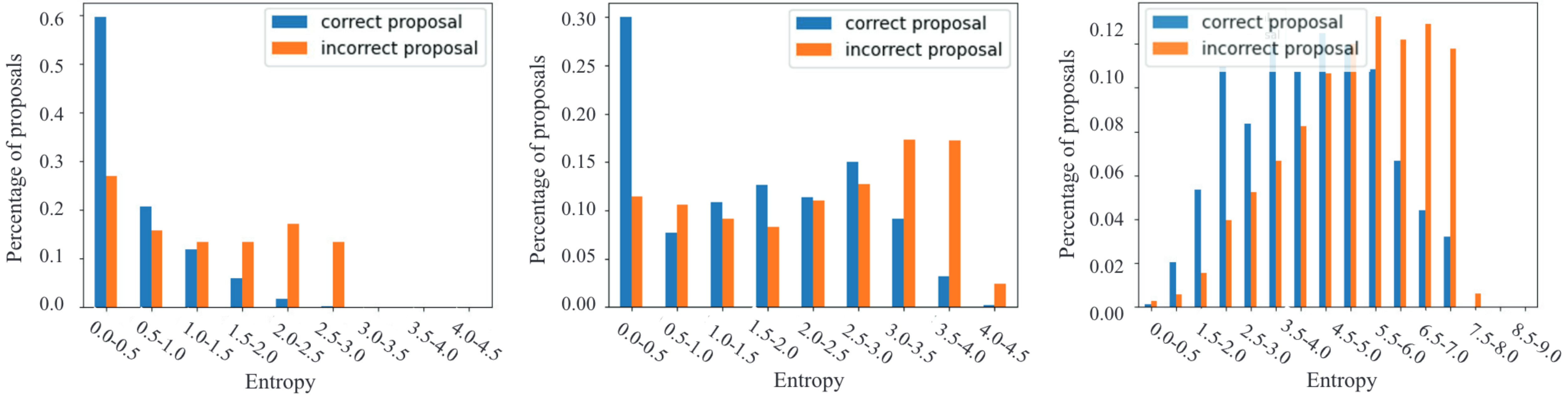


Figure 4. Illustration of our ProposalCLIP. (a) The initial proposal generation model extracts initial proposals. (b) The CLIP proposal selection model selects and re-scores proposals based on CLIP cues. (c) The graph-based proposal merging model corrects fragmented proposals based on CLIP features. (d) The proposal regression model refines proposals.

ProposalCLIP: Unsupervised Open-Category Object Proposal Generation via Exploiting CLIP Cues

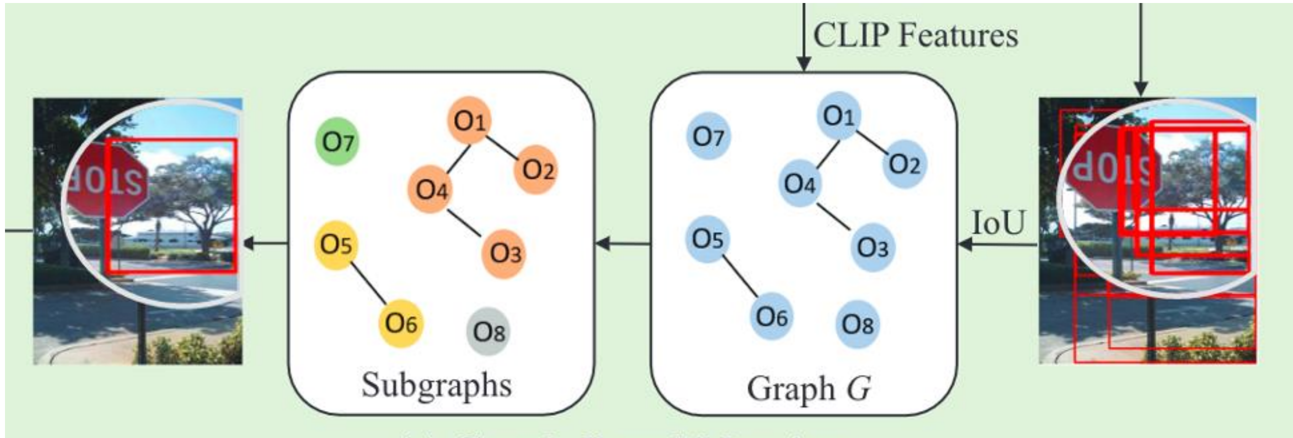
- Select 60% low-similarity-entropy initial proposal



(a) VOC 2007 (20 Categories) (b) COCO (80 Categories) (c) Visual Genome (1600 Categories)

ProposalCLIP: Unsupervised Open-Category Object Proposal Generation via Exploiting CLIP Cues

- Merging



(c) Graph-Based Merging

- Semi-supervise
  - intersection between top 1% low-entropy proposals and top 5% high-initial-score proposals

$$S_t = -\frac{T}{C}\frac{E_t}{\sqrt{\sum_{t=1}^{T} E_t^2}} + \lambda_{sim} \max_{c=1,\ldots,C} Sim_{t,c} + \lambda_{sl} SL_t$$

Distinguishing Unseen from Seen for Generalized Zero-shot Learning

Hongzu Su[1] Jingjing Li[12*] Zhi Chen[3] Lei Zhu[4] Ke Lu[1]
[1]University of Electronic Science and Technology of China
[2] Institute of Electronic and Information Engineering of UESTC in Guangdong
[3]The University of Queensland [4]Shandong Normal University

Zero-shot, GAN

Distinguishing Unseen from Seen for Generalized Zero-shot Learning

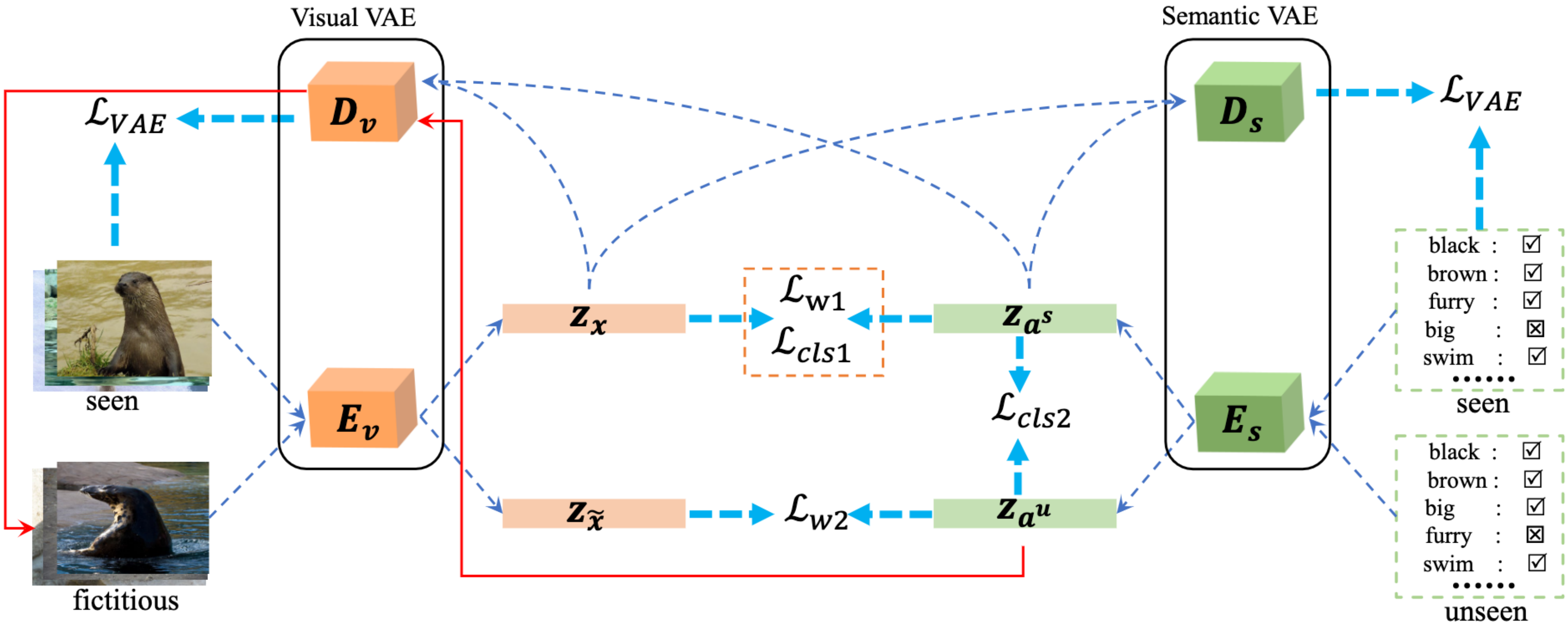


Figure 2. Illustration of our framework. E_v, D_v, E_s and D_s refer to visual encoder, visual decoder, semantic encoder and semantic decoder, respectively. Notations $z_x, z_{a^s}, z_{\tilde{x}}$ and z_{a^u} denote latent representations of seen visual samples, seen semantic descriptions, fictitious visual samples and unseen semantic descriptions, respectively. Notice that fictitious classes are generated with semantic encoder and visual decoder in our method. Red lines indicate the generation of fictitious classes.