JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Learning to Generate Text-grounded Mask for Open-world Semantic Segmentation from Only Image-Text Pairs

Junbum Cha, Jonghwan Mun, Byungseok Roh

kakaobrain

Poster: WED-AM-279

Target Task

Open-world Semantic Segmentation

learning an universal model to segment **arbitrary concepts** beyond pre-defined categories

from Only Image-Text Pairs

using only image-text pairs **without any segmentation annotation**

Open-world Semantic Segmentation

- Open-world segmentation model can segment **different dog species**,

Open-world Semantic Segmentation

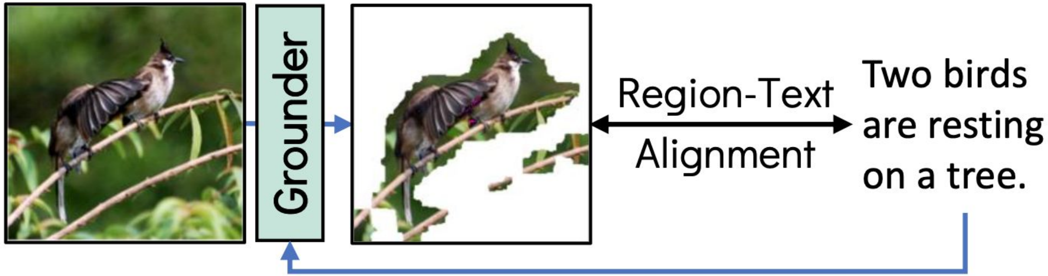
- Open-world segmentation model can segment different dog species, **bananas by color**,

Open-world Semantic Segmentation

- Open-world segmentation model can segment different dog species, bananas by color, and **even proper nouns** such as Frodo, Gollum, and Samwise

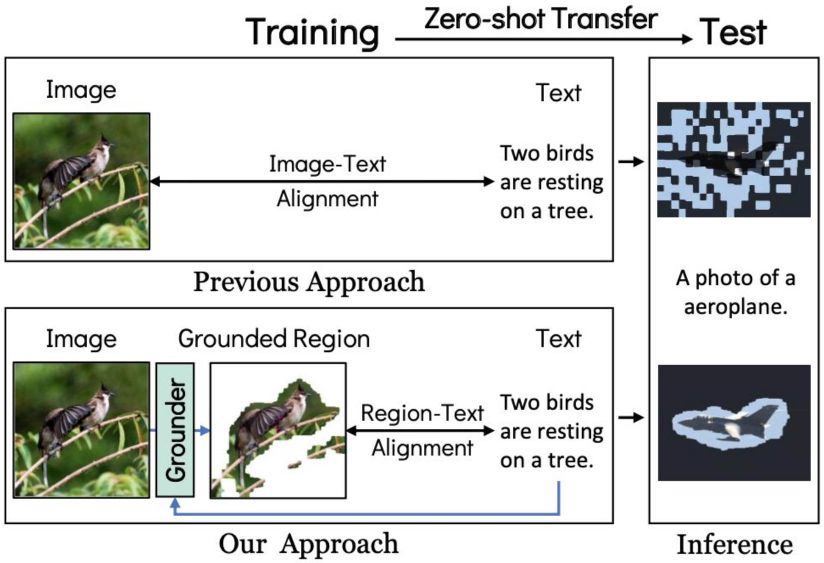
Region-Text Alignment

- Open-world segmentation is conducted by region-text alignment
- In inference, the model identifies regions in the image that align with the given text queries



Train-Test Discrepancy in Existing Works

All of the existing methods train the model via **image-text alignment**, even though the target task requires **region-text alignment**



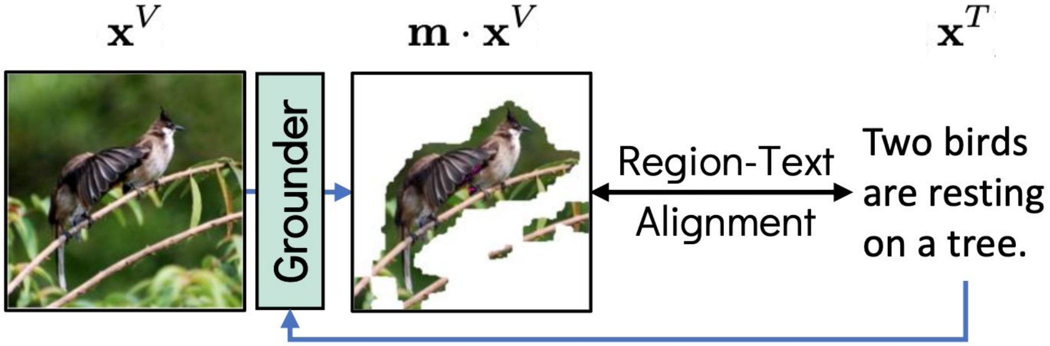
TCL: Text-grounded Contrastive Learning

We propose Text-grounded Contrastive Learning (TCL) to let the model learn region-text alignment instead of image-text alignment in training time

CL $\quad \arg\max_{\theta} I_{\theta}(\mathbf{x}^{V}; \mathbf{x}^{T})$

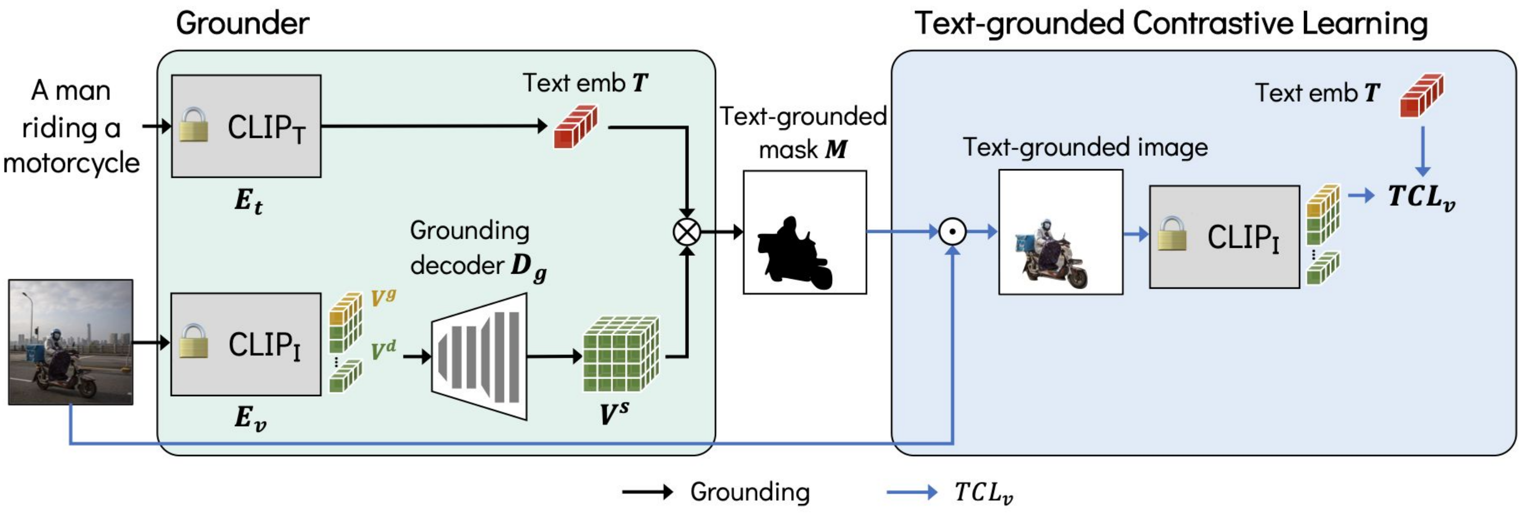
$\arg\max_{\theta} I_{\theta}(\mathbf{m} \cdot \mathbf{x}^{V}; \mathbf{x}^{T}) \quad$ TCL

Text-grounded mask



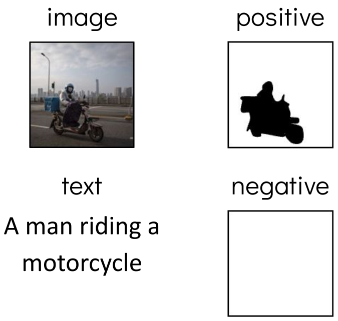
TCL FRAMEWORK

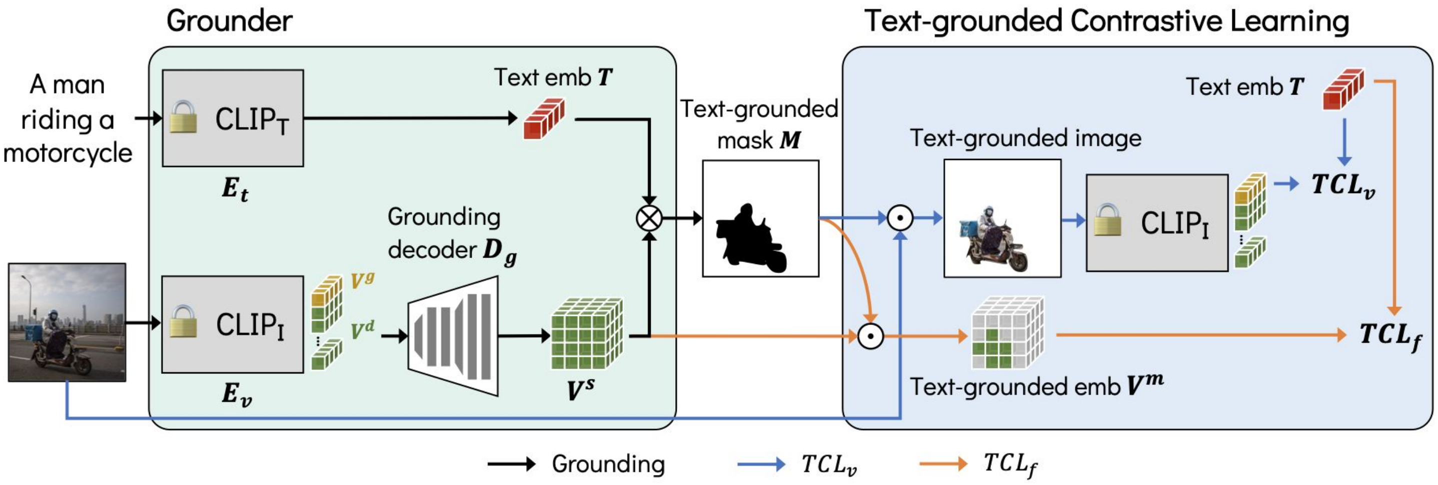
1. Grounder generates text-grounded mask M
2. TCL loss is computed by incorporating text-grounded image into contrastive learning



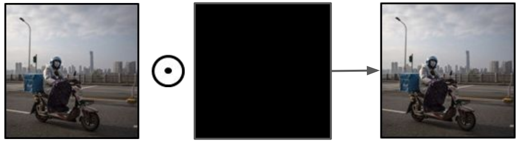
Feature-level TCL Loss

- (Image-level) TCL loss only can consider "positive" text-grounded masks
- We introduce feature-level TCL loss to incorporate "negative" masks into our objective





PREVENTING TRIVIAL SOLUTION



- There is a trivial solution in this framework: generating full-mask independent of the text
- To prevent this trivial solution, we introduce area TCL loss:

A man riding a motorcycle

$$\mathcal{L}_{\text{area}} = \left\| p^+ - \mathbb{E}\left[\overline{\mathbf{M}^+}\right] \right\|_1 + \left\| p^- - \mathbb{E}\left[\overline{\mathbf{M}^-}\right] \right\|_1$$

Expectation of positive text-grounded mask area

Expectation of negative text-grounded mask area

Smooth Regularization

Finally, we introduce a smooth regularization loss via total variation (TV) loss:

$$\mathcal{L}_{\text{tv}} = \underbrace{\|\mathbf{M}\|_{\text{TV}}}_{\text{text-grounded mask}} + \underbrace{\|\mathbf{V}^s\|_{\text{TV}}}_{\text{dense image embedding}}$$

where

$$\|y\|_{\text{TV}} = \sum_{i,j} \sqrt{|y_{i+1,j} - y_{i,j}|^2} + \sqrt{|y_{i,j+1} - y_{i,j}|^2} = \sum_{i,j} |y_{i+1,j} - y_{i,j}| + |y_{i,j+1} - y_{i,j}|$$

SUM UP

We introduce TCL loss to let the model learn **region-text alignment**, instead of image-text alignment

$$\mathcal{L}_{\text{final}} = \overbrace{\lambda_{\text{TCL}}(\mathcal{L}_{\text{TCL}_v} + \mathcal{L}_{\text{TCL}_f}) + \lambda_{\text{area}}\mathcal{L}_{\text{area}}}^{\text{TCL losses}} + \overbrace{\lambda_{\text{tv}}\mathcal{L}_{\text{tv}}}^{\text{regularization}}$$

image-level TCL

feature-level TCL

area TCL

smooth regularization

Zero-shot Evaluation Protocol

- Since the open-world semantic segmentation task is introduced recently, evaluation protocols vary across studies
- For a fair comparison, we present a unified evaluation protocol

Prohibited examples

$\left.\begin{array}{l}\text{floor-other}\\ \text{floor-tile}\\ \text{floor-wood}\\ \text{floor-stone}\end{array}\right\}$ floor

Class integration

Crop

vegetation -> tree

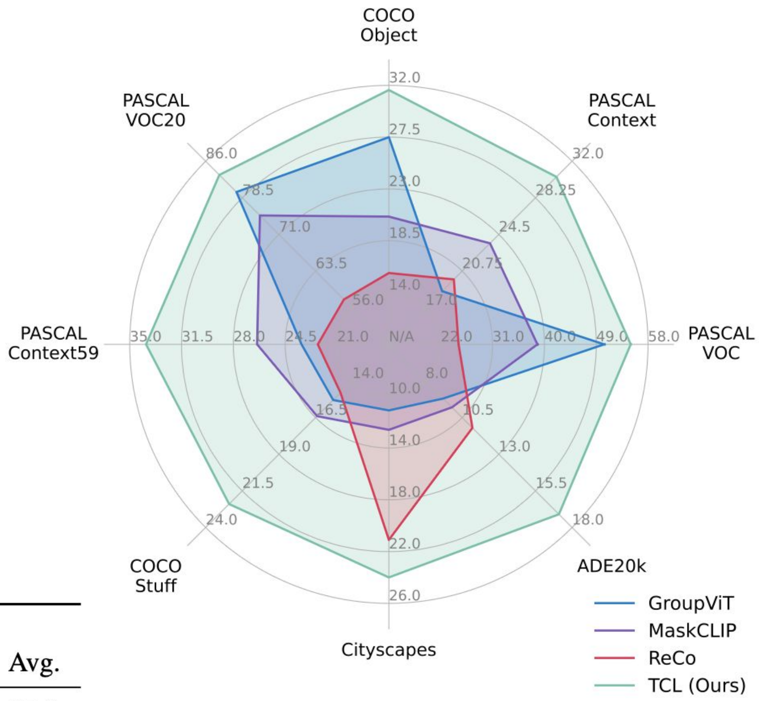
Rephrasing

VOC: threshold=0.95
Context: threshold=0.35
…

Dataset-specific HPs

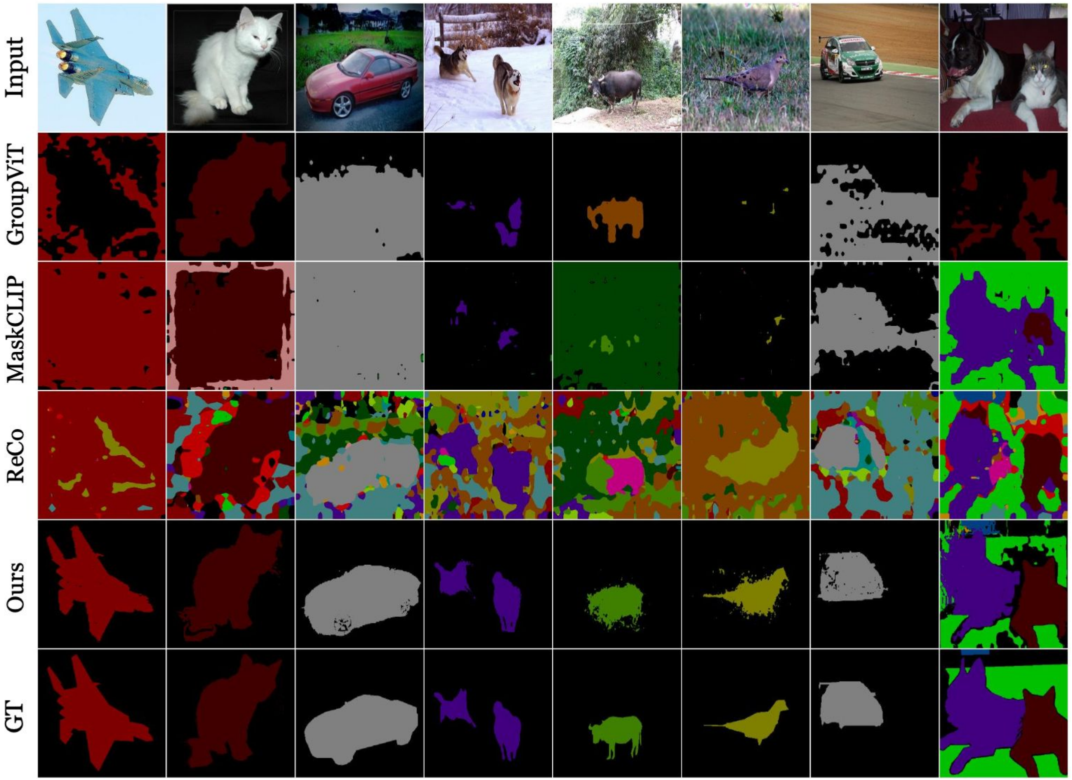
QUANTITATIVE RESULTS

TCL remarkably outperforms previous methods in every dataset



Methods	with background class VOC	Context	Object	without background class VOC20	Context59	Stuff	City	ADE	Avg.
GroupViT (YFCC)	49.5	19.0	24.3	74.1	20.8	12.6	6.9	8.7	27.0
GroupViT (RedCaps)	<u>50.4</u>	18.7	<u>27.5</u>	<u>79.7</u>	23.4	15.3	11.1	9.2	<u>29.4</u>
MaskCLIP†	29.3	21.1	15.5	53.7	23.3	14.7	<u>21.6</u>	10.8	23.7
MaskCLIP	38.8	<u>23.6</u>	20.6	74.9	<u>26.4</u>	<u>16.4</u>	12.6	9.8	27.9
ReCo	25.1	19.9	15.7	57.7	22.3	14.8	21.1	<u>11.2</u>	23.5
TCL (Ours)	**55.0** (+4.6)	**30.4** (+6.8)	**31.6** (+4.1)	**83.2** (+3.5)	**33.9** (+7.5)	**22.4** (+6.0)	**24.0** (+2.4)	**17.1** (+5.9)	**37.2** (+7.8)

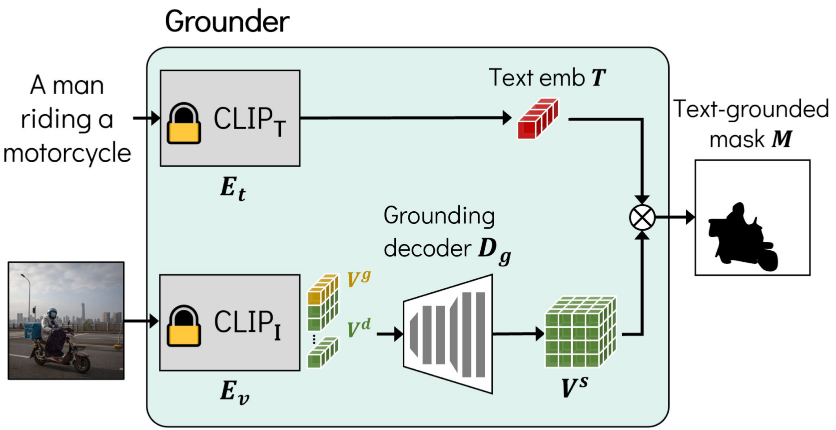
Qualitative Comparison on Pascal VOC

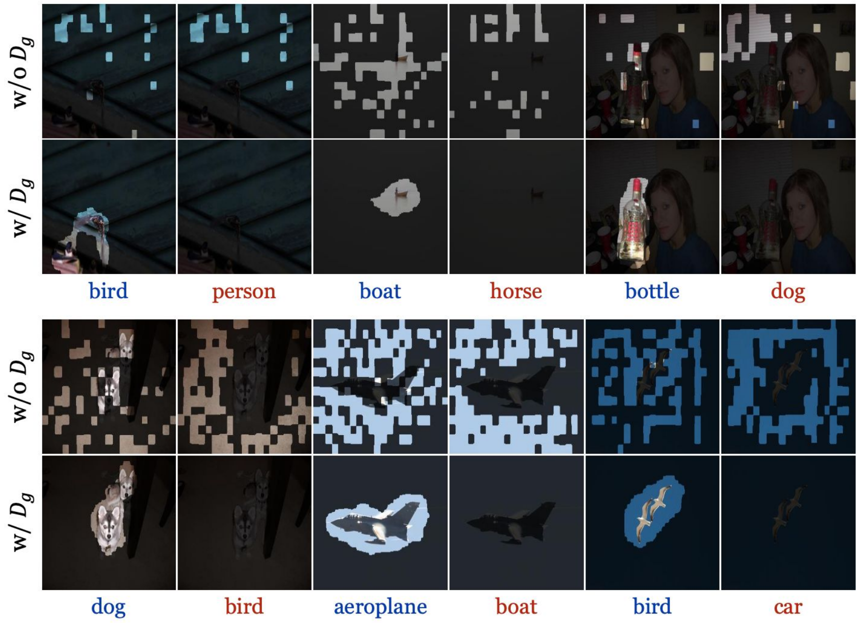


Qualitative Examples in the Wild

Grounding Visualization

Grounding decoder lets the model learn region-text alignment





ABLATION STUDIES

Table (a) indicates that TCL loss significantly improves segmentation performance by learning region-text alignment

[B] Training the grounding decoder without TCL loss does not improve performance

	Method	VOC20
A	Baseline	53.2
B	+ Decoder	52.3
C	+ TCL	**77.4**

(a) **Baseline to TCL.**

	TCL_v	TCL_f	$\mathcal{L}_{\text{area}}$	CL	VOC20
D				✔	61.1
E	✔		✔		74.6
F		✔	✔		76.0
C	✔	✔	✔		**77.4**
G	✔	✔	✔	✔	75.6
H	✔	✔			67.1

(b) **TCL losses.**

THANK YOU!

Contact: junbum.cha@kakaobrain.com
Code: https://github.com/kakaobrain/tcl
Demo: https://huggingface.co/spaces/khanrc/tcl

Poster: WED-AM-279