CRIS: CLIP-Driven Referring Image Segmentation

Zhaoqing Wang[1,2]*, Yu Lu[3]*, Qiang Li[4]*, Xunqiang Tao[2], Yandong Guo[2], Mingming Gong[5], Tongliang Liu[1]

[1]University of Sydney

[2]OPPO Research Institute

[3]Beijing University of Posts and Telecommunications

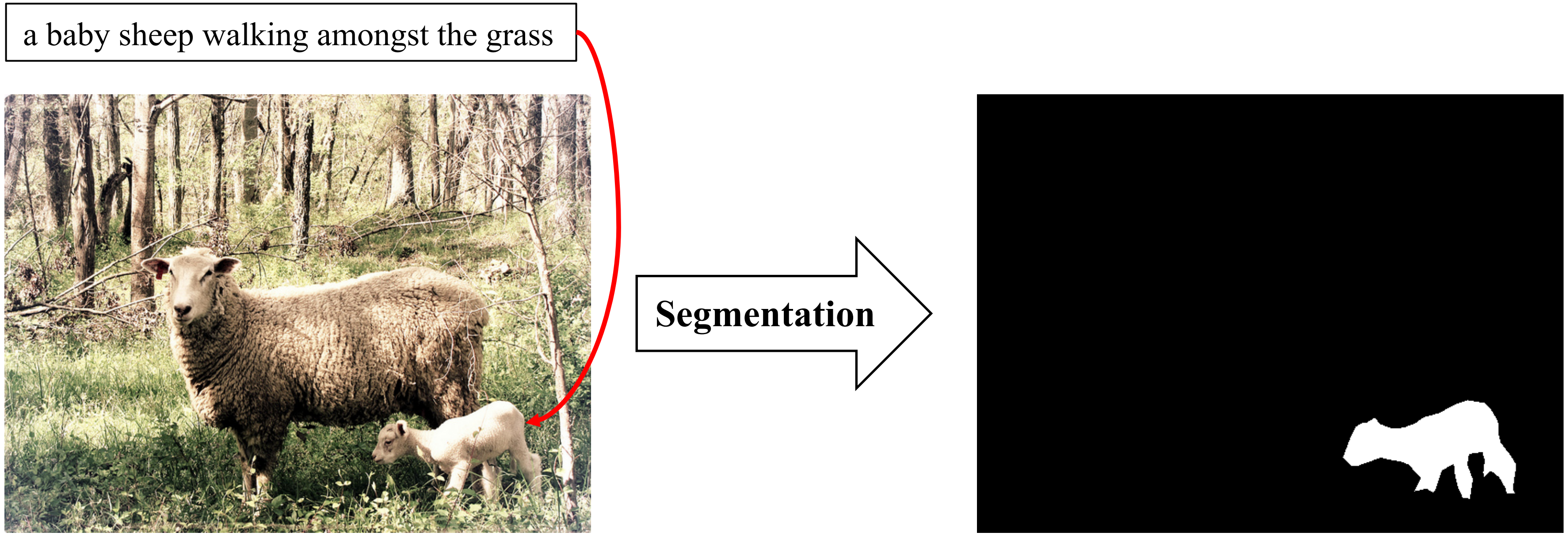
[4]Kuaishou Technology

[5]University of Melbourne

https://github.com/DerrickWang005/CRIS.pytorch.git

Referring Image Segmentation (RES)

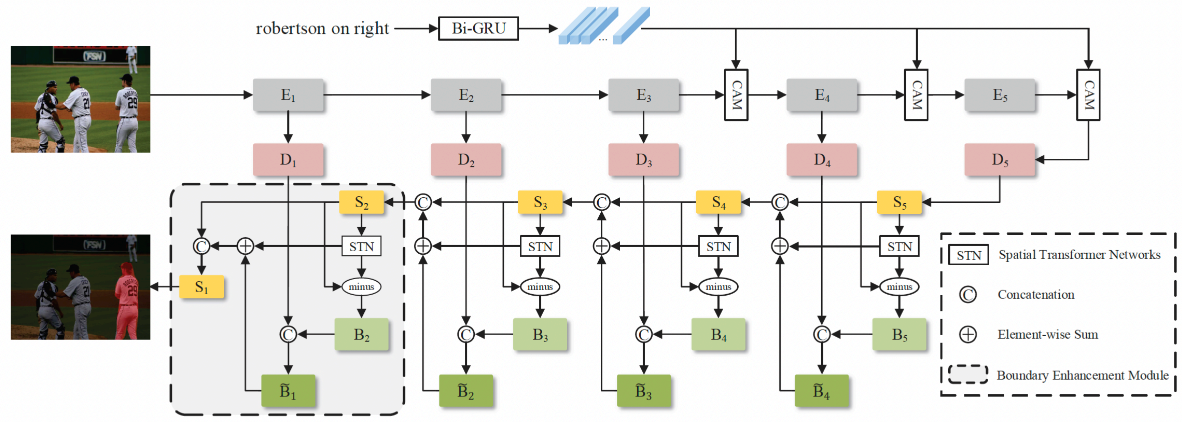
- RES is a fundamental and challenging task at the intersection of vision and language understanding.
- RES aims to segment a referent via a natural linguistic expression.



Previous Work

Encoder Fusion Network with Co-Attention Embedding for Referring Image Segmentation

- In most previous works, linguistic feature interacts with visual feature of each scale **separately**, which ignores the continuous guidance of language to multi-scale visual features.
- They propose an **encoder fusion network (EFN)**, which transforms the visual encoder into a multi-modal feature learning network and uses language to refine the multi-modal features progressively.
- They also propose a **boundary enhancement** module to make the network pay more attention to the fine structure.

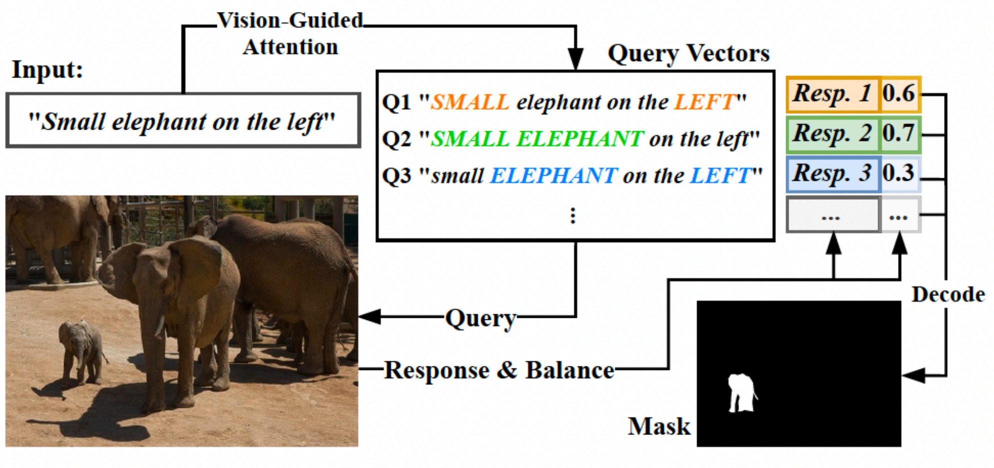


Feng, Guang, et al. "Encoder fusion network with co-attention embedding for referring image segmentation." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.

Previous Work

Vision-Language Transformer and Query Generation for Referring Segmentation

- The linguistic expression in RES can be treated as **a query**, which indicates the target object by describing its relationship with others. Then, RES is reformulated as **a direct attention problem**: finding the region in the image where the query is most attended to.
- They build a network with **an encoder-decoder attention mechanism architecture** that “queries” the given image with the language expression.
- They propose **a Query Generation Module (QGM)** that understands the language from different comprehension ways, and **a Query Balance Module (QBM)** to focus on the suitable ways.

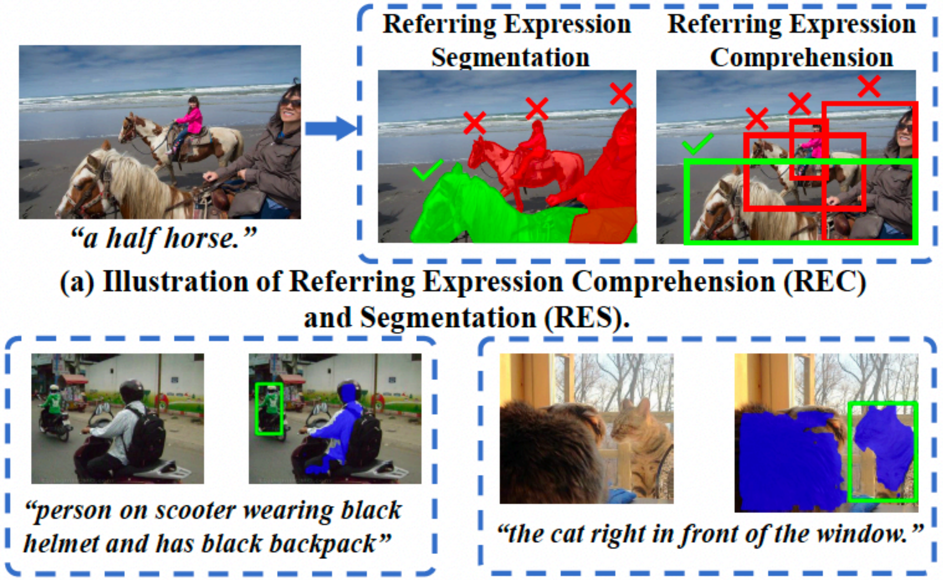


Ding, Henghui, et al. "Vision-language transformer and query generation for referring segmentation." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.

Previous Work

Multi-task Collaborative Network for Joint Referring Expression Comprehension and Segmentation

- Referring expression comprehension (REC) and segmentation (RES) are **two highly-related tasks**, which both aim at identifying the referent according to a natural language expression. RES can help REC to achieve better language-vision alignment, while REC can help RES to better locate the referent.
- To address **the prediction conflict**, they propose two innovative designs: **Consistency Energy Maximization (CEM)** and **Adaptive Soft Non-Located Suppression (ASNLS)**.
- CEM enables REC and RES to focus on similar visual regions by maximizing the consistency energy between two tasks.
- ASNLS suppresses the response of unrelated regions in RES based on the prediction of REC.



(a) Illustration of Referring Expression Comprehension (REC) and Segmentation (RES).

(b) Illustraion of the prediction conflict.

Luo, Gen, et al. "Multi-task collaborative network for joint referring expression comprehension and segmentation." *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*. 2020.

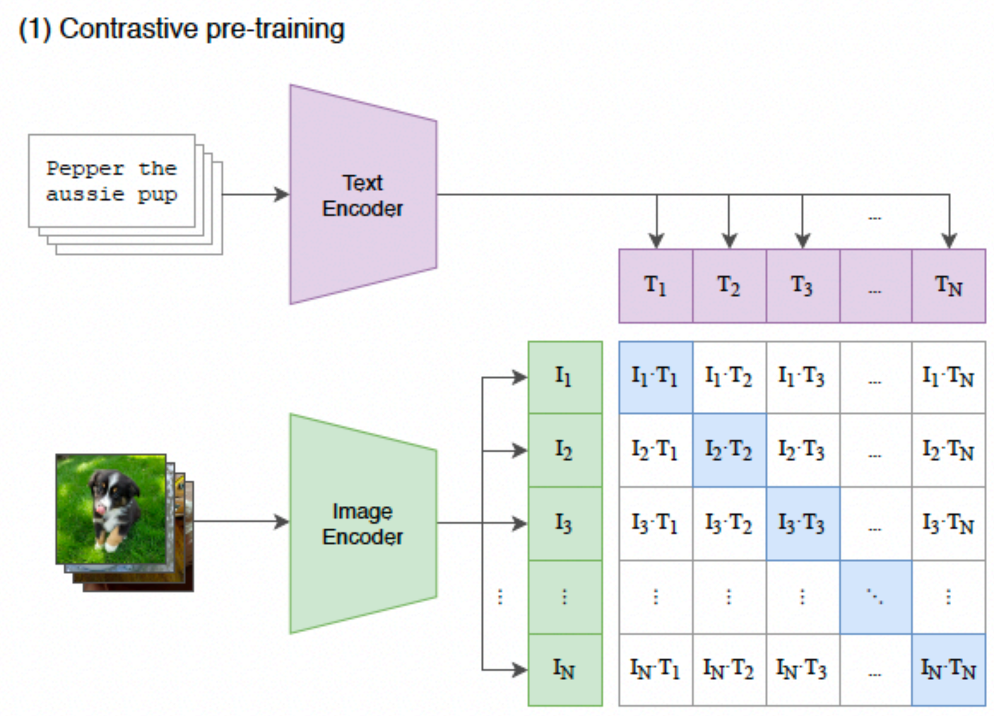
Motivation

- Due to **the distinct data properties** between text and image, it is challenging for a network to **well align** text and pixel-level features.
- Existing approaches use pretrained models to facilitate learning, yet **separately transfer** the language / vision knowledge from pretrained models, ignoring **the multi-modal corresponding information.**
- Overly **complex** model architectures and fusion strategies.

Introduction

Learning Transferable Visual Models From Natural Language Supervision

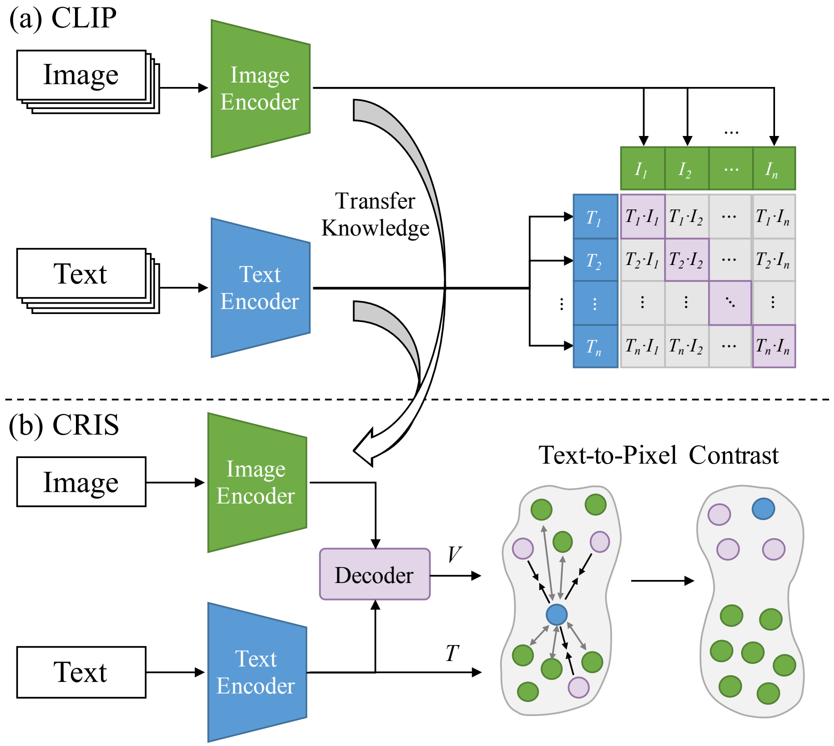
- State-of-the-art computer vision systems are trained to predict a fixed set of predetermined object categories. This restricted form of supervision limits their generality and usability since additional labeled data is needed to specify any other visual concept.
- Learning from a large-scale dataset of 400 million (image, text) pairs collected from the internet.
- Powerful vision-language alignment capability.



Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International Conference on Machine Learning*. PMLR, 2021.

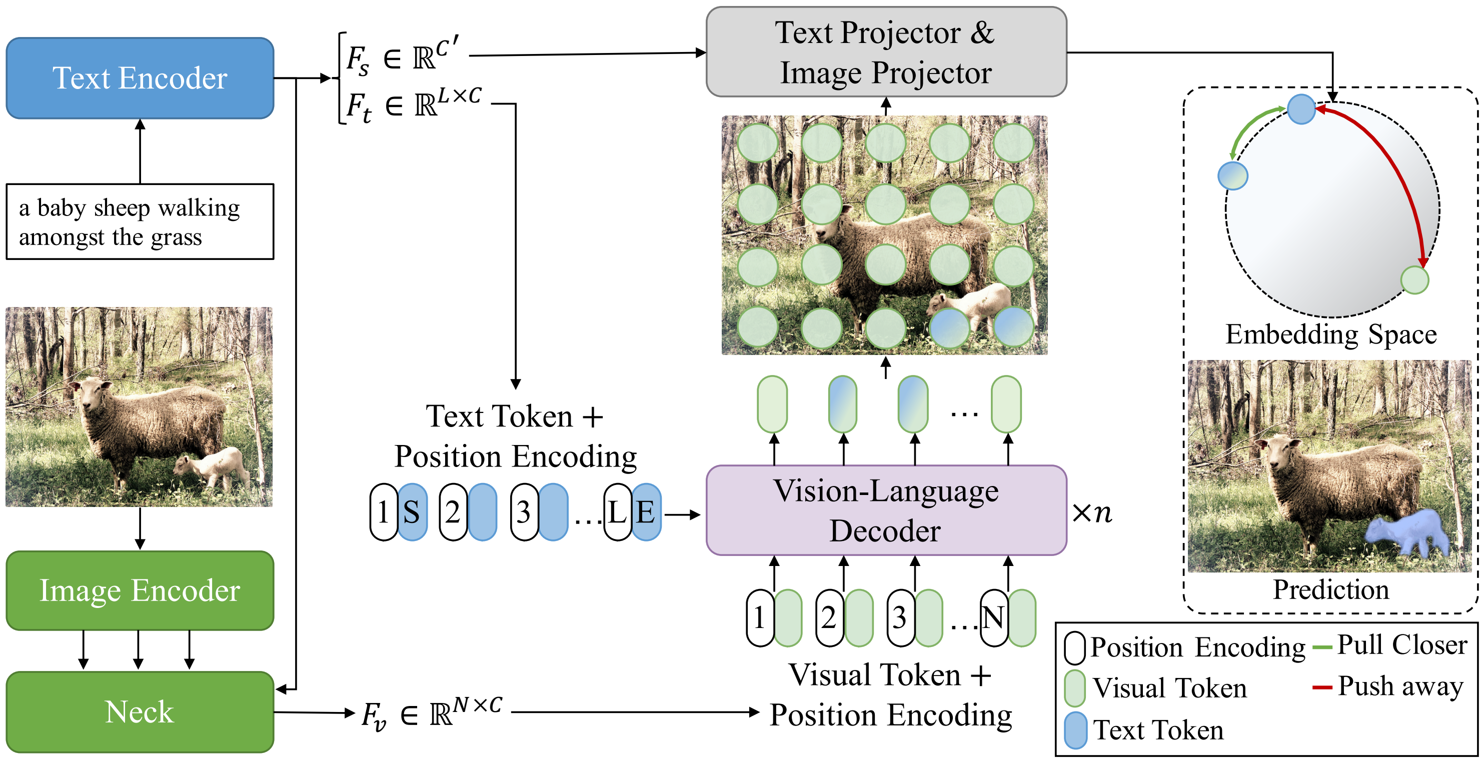
Introduction

- CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of image I and text T, which can capture the text-image information.
- To well transfer the powerful multi-modal knowledge of CLIP models, we propose a CLIP-Driven Referring Image Segmentation framework (CRIS).
- To generalize the multi-modal knowledge from image level to pixel level, CRIS resorts to vision-language decoding and contrastive learning for achieving the text-to-pixel alignment.



Method

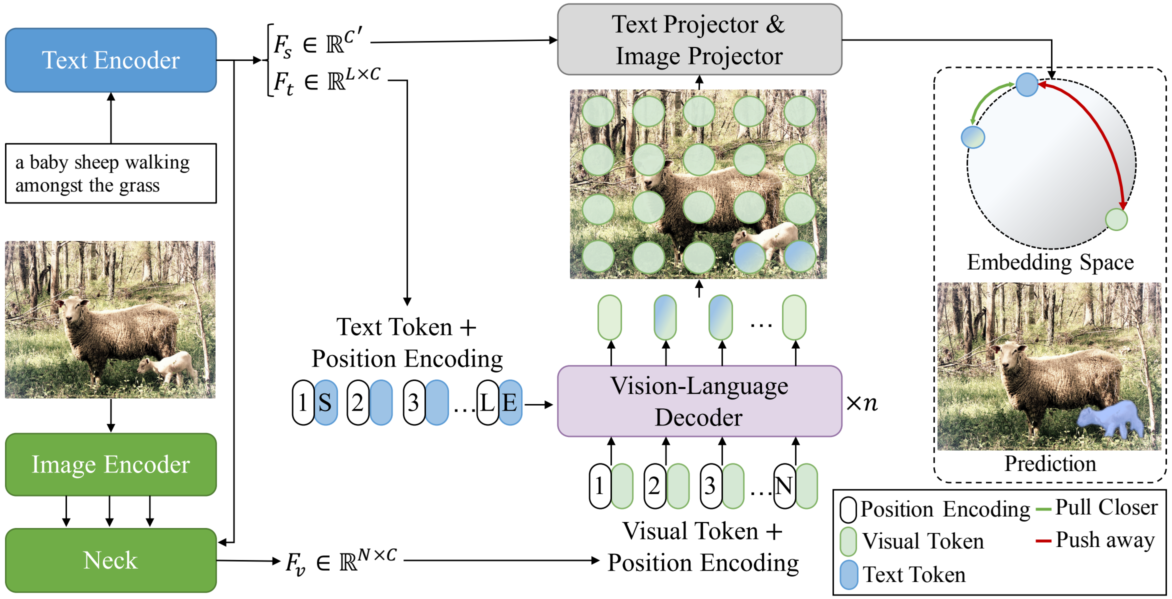
- CRIS mainly consists of a text encoder, an image encoder, a cross-modal neck, a vision-language decoder, and two projectors.



Method

Text Encoder:

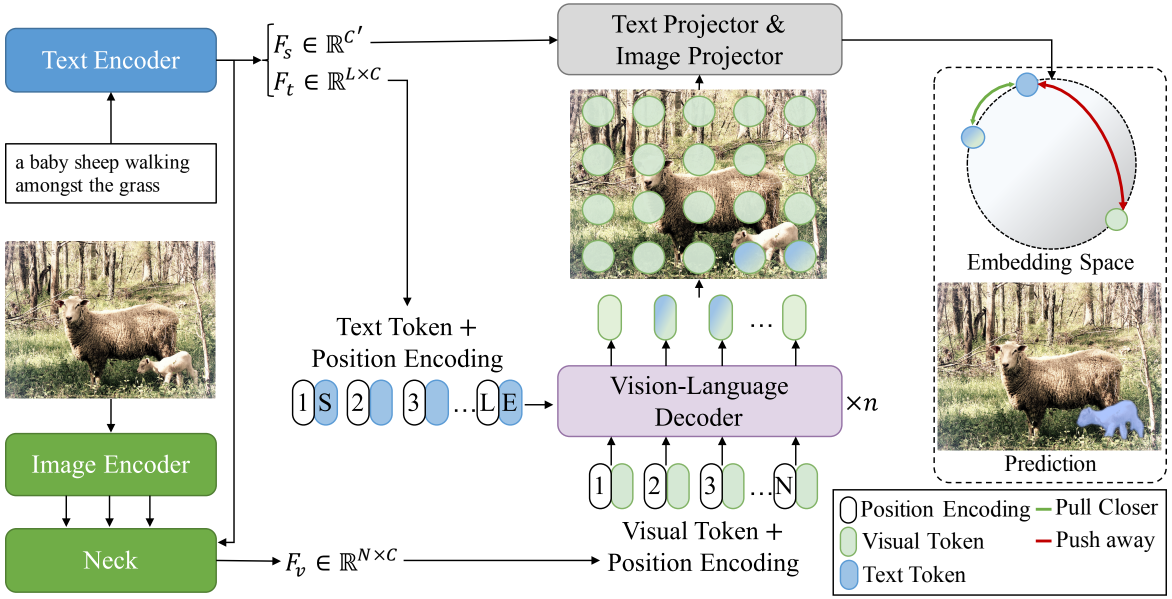
- A lower-cased byte pair encoding (BPE) representation of the text with a 49,152 vocab size.
- A modified transformer.
- Each text sequence is bracketed with [SOS] and [EOS] tokens.



Method

Image Encoder & Neck:

- A ResNet-50/101used in CLIP.
- To stablize training, we add a residual connect in the attention pooling layer of the ResNet.
- Following most previous methods, we adopt a cross-modal FPN to fuse the multi-level visual features and the sentence representation.



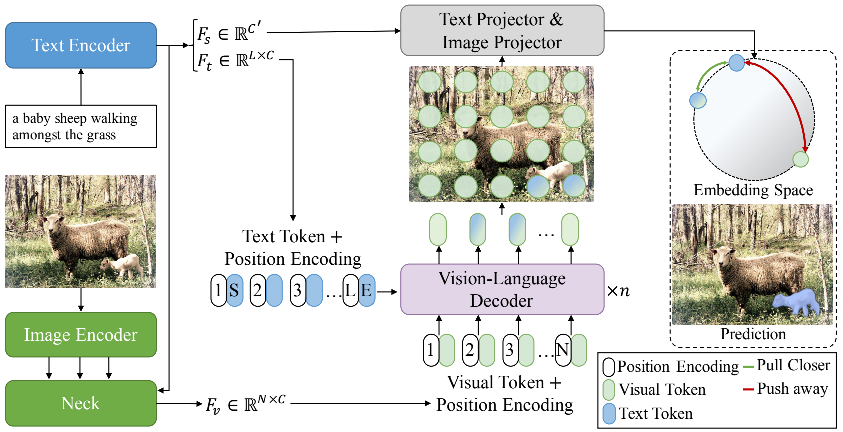
Method

Vision-Language Decoding:

- We design a vision-language decoder to adaptively propagate fine-grained semantic information from textual features to visual features.
- The vision-language decoder composed of n layers (n=3) is applied to generate a sequence of evolved multi-modal features F_c.
- Following the standard architecture of the transformer, each layer consists of a multi-head self-attention layer, a multi-head cross-attention layer, and a feed-forward network. In one decoder layer, F_v is first sent into the multi-head self-attention layer to capture global contextual information.

$$F'_v = MHSA(LN(F_v)) + F_v,$$

$$MHSA(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V.$$

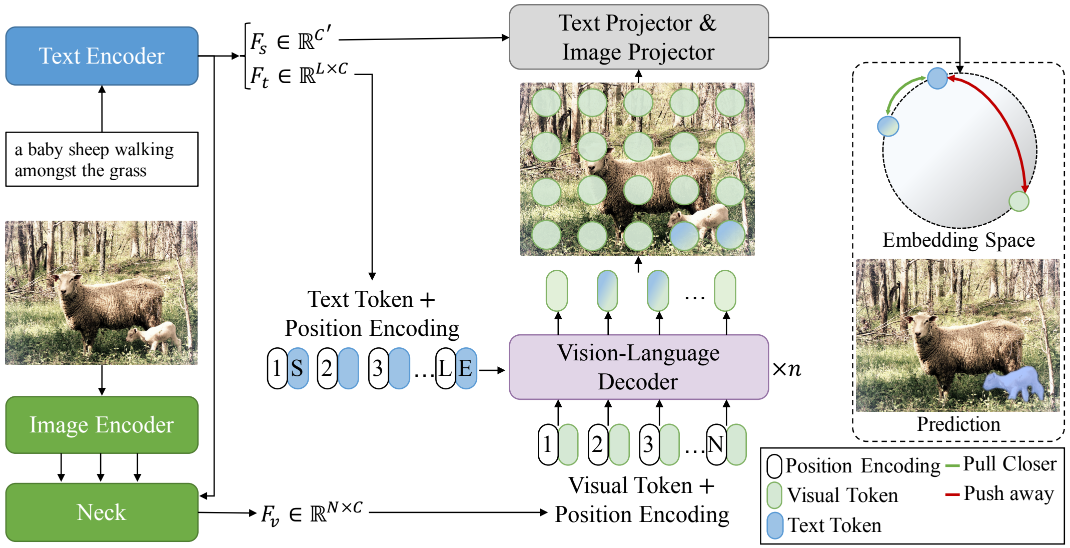


Method

Vision-Language Decoding:

- After that, the multi-head cross-attention layer and a MLP block of two linear layers with Layer Normalization and residual connections are adopted to propagate fine-grained semantic information into the evolved multi-modal features F_c.

$$F_c' = MHCA(LN(F_v'), F_t) + F_v',$$
$$F_c = MLP(LN(F_c')) + F_c' .$$



Method

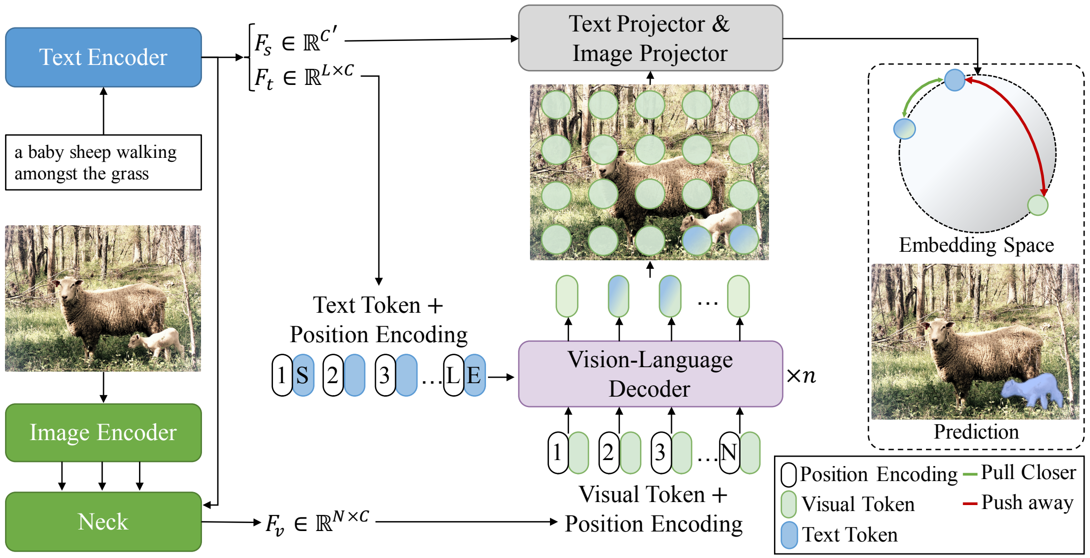
Text-to-pixel Contrastive Learning:

- A text projector and an image projector are adopted to transfer features into a multi-modal embedding space.
- We design a text-to-pixel contrastive loss to learn more fine-grained multi-model representations. A language expression and related pixel-wise features are pulled closer, while other irrelevances are pushed away.

$$z_v = F_c' W_v + b_v, \quad F_c' = Upsample(F_c),$$
$$z_t = F_s W_t + b_t,$$

$$L_{con}(z_t, z_v) = \frac{1}{|\mathcal{P} \cup \mathcal{N}|} \sum_{i \in \mathcal{P} \cup \mathcal{N}} L_{con}^i (z_t, z_v^i),$$

$$L_{con}^i(z_t, z_v^i) = \begin{cases} -log\sigma(z_t \cdot z_v^i), & i \in \mathcal{P} \\ -\log\left(1 - \sigma(z_t \cdot z_v^i)\right), & i \in \mathcal{N} \end{cases}$$

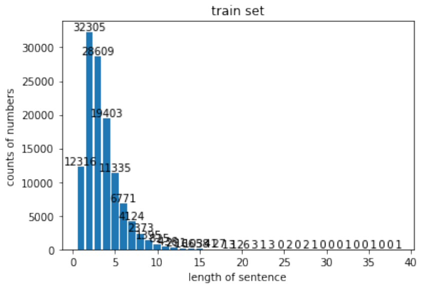


Experiments

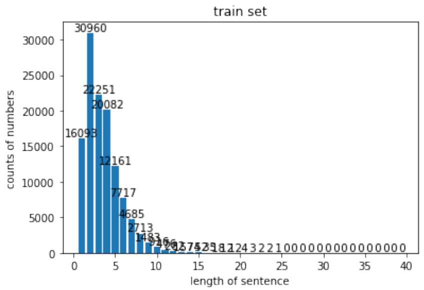
Dataset:

- RefCOCO
  a) Train / Val / TestA / TestB: 42404 / 3811 / 1975 / 1810
  b) Sentence: attribute, location…
  c) Length: min-1 / max-39 / mean-3.6
- RefCOCO+
  a) Train / Val / TestA / TestB: 42278 / 3805 / 1975 / 1798
  b) Sentence: No location information
  c) Length: min-1 / max-24 / mean-3.6
- G-Ref
  a) Train / Val / Test: 42226 / 2573 / 5023
  b) Sentence: More detailed descriptions
  c) Length: min-1 / max-46 / mean-8.4

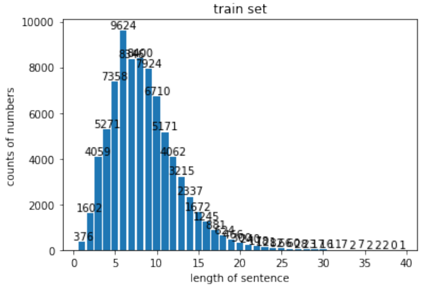
RefCOCO



RefCOCO+



G-Ref



Experiments

- To evaluate the effectiveness of each component in our method, we conduct extensive experiments on three benchmarks, including RefCOCO, RefCOCO+, and G-Ref.

Method	Backbone	RefCOCO			RefCOCO+			G-Ref	
		val	test A	test B	val	test A	test B	val	test
RMI* [24]	ResNet-101	45.18	45.69	45.57	29.86	30.48	29.50	-	-
DMN [32]	ResNet-101	49.78	54.83	45.13	38.88	44.22	32.29	-	-
RRN* [22]	ResNet-101	55.33	57.26	53.95	39.75	42.15	36.11	-	-
MAttNet [49]	ResNet-101	56.51	62.37	51.70	46.67	52.39	40.08	47.64	48.61
NMTree [25]	ResNet-101	56.59	63.02	52.06	47.40	53.01	41.56	46.59	47.88
CMSA* [48]	ResNet-101	58.32	60.61	55.09	43.76	47.60	37.89	-	-
Lang2Seg [5]	ResNet-101	58.90	61.77	53.81	-	-	-	46.37	46.95
BCAN* [16]	ResNet-101	61.35	63.37	59.57	48.57	52.87	42.13	-	-
CMPC* [17]	ResNet-101	61.36	64.53	59.64	49.56	53.44	43.23	-	-
LSCM* [18]	ResNet-101	61.47	64.99	59.55	49.34	53.12	43.50	-	-
MCN [29]	DarkNet-53	62.44	64.20	59.71	50.62	54.99	44.69	49.22	49.40
CGAN [28]	DarkNet-53	64.86	68.04	62.07	51.03	55.51	44.06	51.01	51.69
EFNet [8]	ResNet-101	62.76	65.69	59.67	51.50	55.24	43.01	-	-
LTS [19]	DarkNet-53	65.43	67.76	63.08	54.21	58.32	48.02	54.40	54.25
VLT [6]	DarkNet-53	65.65	68.29	62.73	55.50	59.20	49.36	52.99	56.65
CRIS (Ours)	ResNet-50	69.52	72.72	64.70	61.39	67.10	52.48	59.35	59.39
CRIS (Ours)	ResNet-101	**70.47**	**73.18**	**66.10**	**62.27**	**68.08**	**53.68**	**59.87**	**60.36**

Experiments

Ablation Study:

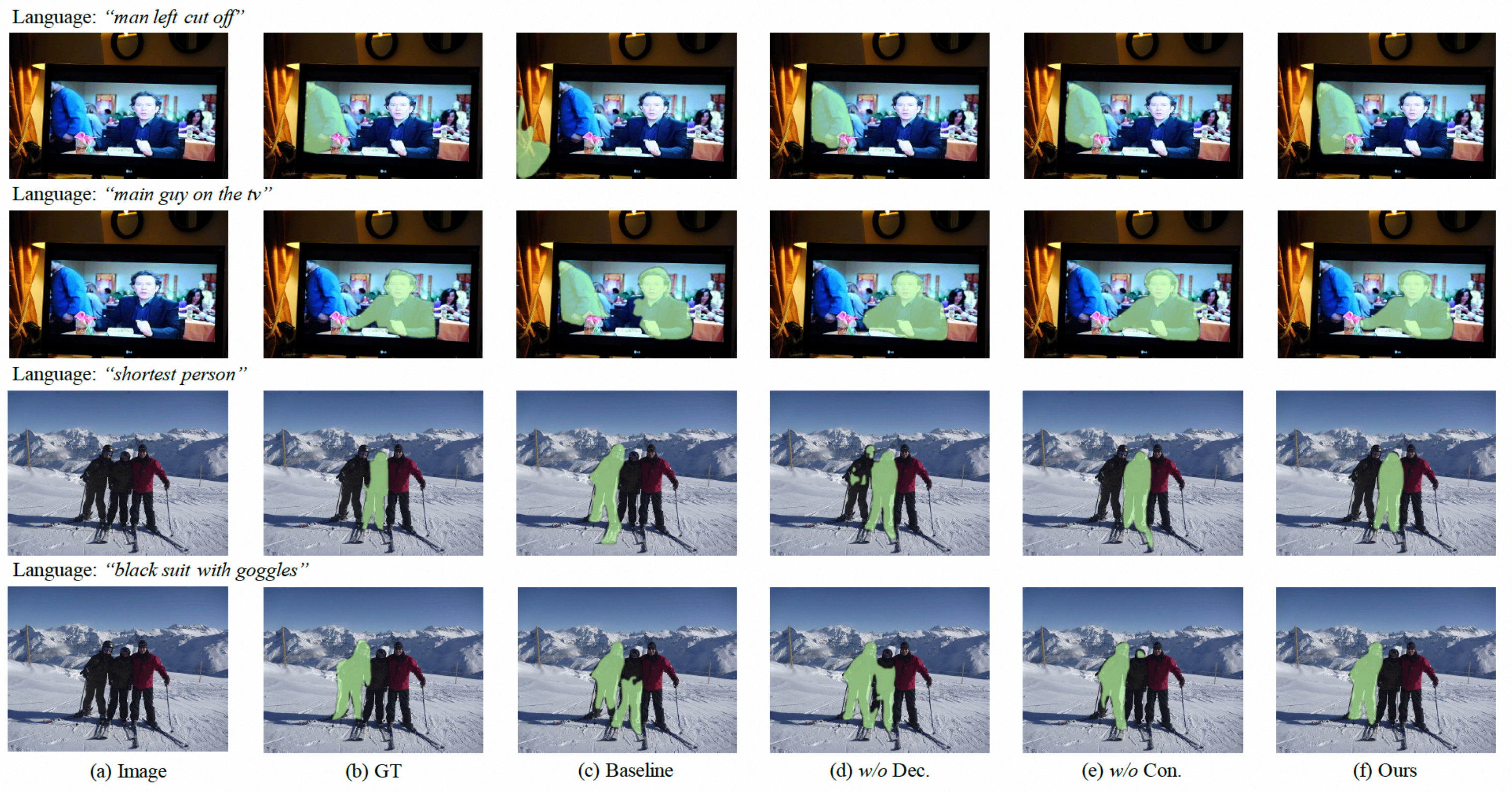
- Effectiveness of Contrastive Learning
- Effectiveness of Vision-Language Decoder
- Numbers of Layers in Decoder
- Efficiency analysis

Dataset	*Con.*	*Dec.*	*n*	IoU	Pr@50	Pr@60	Pr@70	Pr@80	Pr@90	Params	FPS
RefCOCO	-	-	-	62.66	72.55	67.29	59.53	43.52	12.72	131.86	27.30
	✓	-	-	64.64	74.89	69.58	61.70	45.50	13.31	134.22	25.79
	-	✓	1	66.31	77.66	72.99	65.67	48.43	14.81	136.07	23.02
	✓	✓	1	68.66	80.16	75.72	68.82	51.98	15.94	138.43	22.64
	✓	✓	2	69.13	80.96	76.60	69.67	52.23	16.09	142.64	20.68
	✓	✓	3	**69.52**	**81.35**	**77.54**	**70.79**	**52.65**	16.21	146.85	19.22
	✓	✓	4	69.18	80.99	76.74	69.32	52.57	**16.37**	151.06	18.26
RefCOCO+	-	-	-	50.17	54.55	47.69	40.19	28.75	8.21	131.86	27.30
	✓	-	-	53.15	58.28	53.74	46.67	34.01	9.30	134.22	25.79
	-	✓	1	54.73	63.31	58.89	52.46	38.53	11.70	136.07	23.02
	✓	✓	1	59.97	69.19	64.85	58.17	43.47	13.39	138.43	22.64
	✓	✓	2	60.75	70.69	66.83	60.74	45.69	13.42	142.64	20.68
	✓	✓	3	**61.39**	**71.46**	**67.82**	**61.80**	**47.00**	**15.02**	146.85	19.22
	✓	✓	4	61.15	71.05	66.94	61.25	46.98	14.97	151.06	18.26
G-Ref	-	-	-	49.24	53.33	45.49	36.58	23.90	6.92	131.86	25.72
	✓	-	-	52.67	59.27	52.45	44.12	29.53	8.80	134.22	25.33
	-	✓	1	51.46	58.68	53.33	45.61	31.78	10.23	136.07	22.57
	✓	✓	1	57.82	66.28	60.99	53.21	38.58	13.38	138.43	22.34
	✓	✓	2	58.40	67.30	61.72	54.70	39.67	13.40	142.64	20.61
	✓	✓	3	**59.35**	**68.93**	**63.66**	**55.45**	**40.67**	**14.40**	146.85	19.14
	✓	✓	4	58.79	67.91	63.11	55.43	39.81	13.48	151.06	17.84

Experiments

Qualitative Analysis:

- Effectiveness of Contrastive Learning
- Effectiveness of Vision-Language Decoder



Experiments

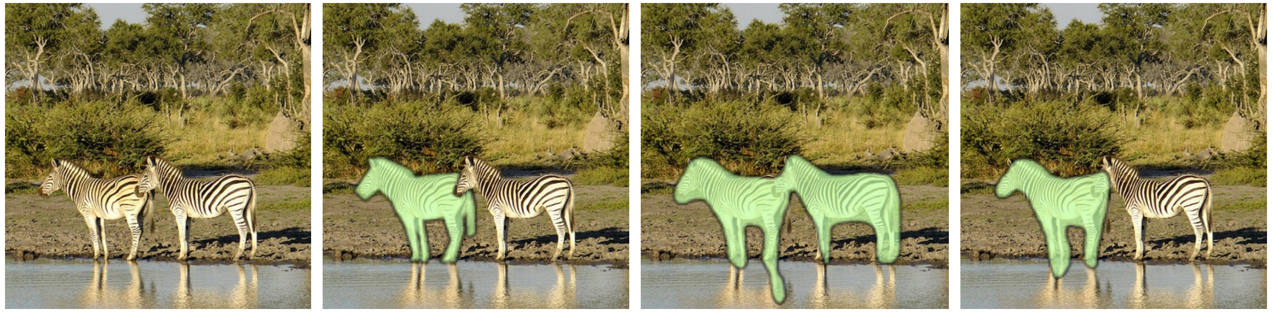
Qualitative Analysis:

- Comparison with Naïve finetuning

Language: *“a blond haired , blue eyed young boy in a blue jacket”*

(a) Image (b) GT (c) Naïve (f) Ours

Language: *“a zebra ahead of the other zebra”*



(a) Image (b) GT (c) Naïve (f) Ours

Experiments

Failure Cases:

Imperfect linguistic expressions:

- The expression of “yellow” is not enough to describe the region of the man in the yellow snowsuit.

Noisy annotation:

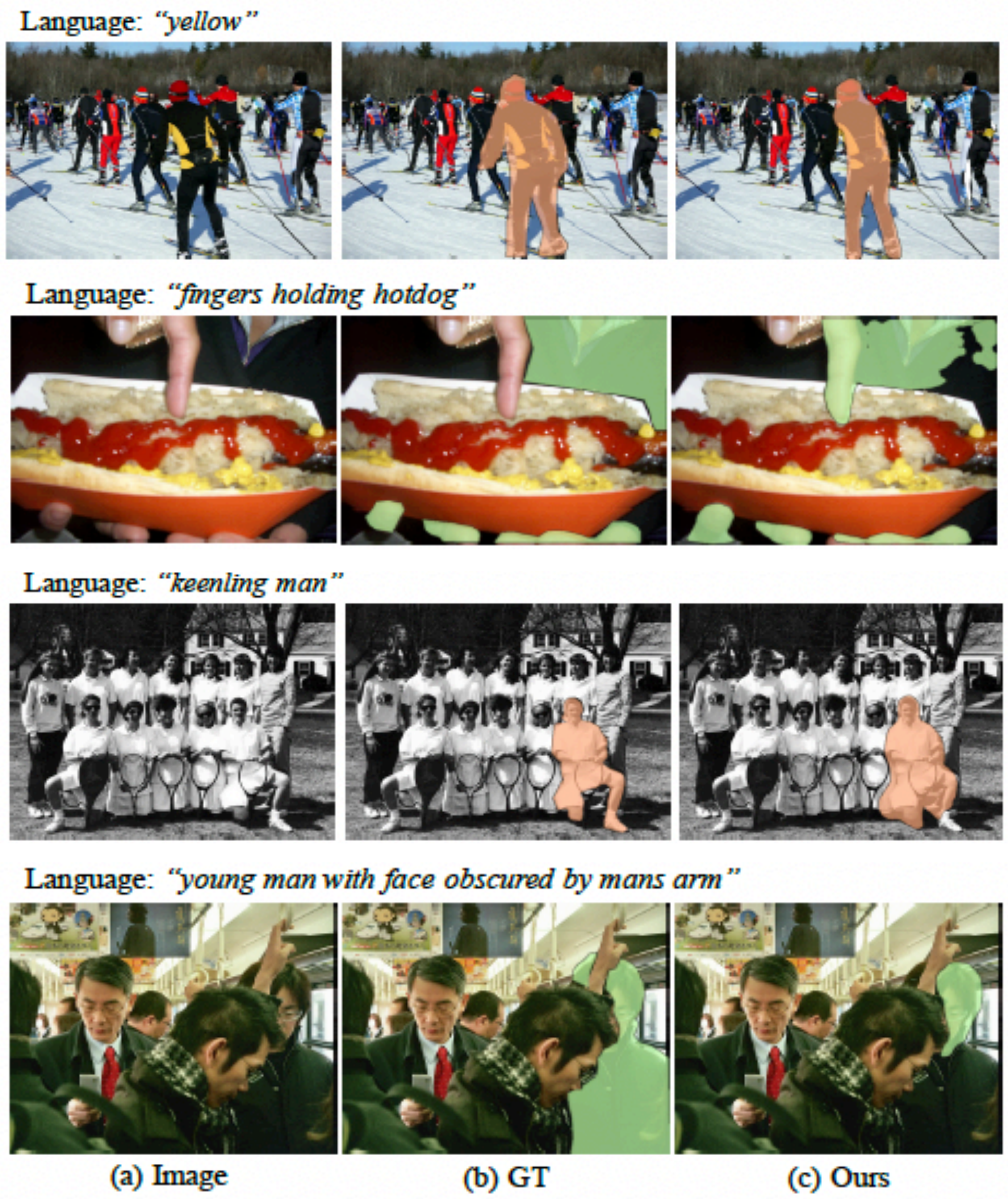
- Some failures are also caused by the wrong label. It is obvious that the top region is unrelated to “fingers”.

Boundary of masks:

- the boundaries of the referent cannot be accurately segmented, but this issue can be alleviated by introducing other technologies, such as the refine module.

Occlusion:

- occlusion could cause failure cases, which is a challenging problem in many vision tasks.



Thanks