Finding NeMo : Negative-mined Mosaic Augmentation for Referring Image Segmentation

Seongsu Ha[1,2*], Chaeyun Kim[1*], Donghwa Kim[1*], Junho Lee[1], Sangho Lee[3], Joonseok Lee[1,4]

Seoul National University [1] Twelve Labs [2] Ai2 [3] Google Research [4]

Referring Image Segmentation

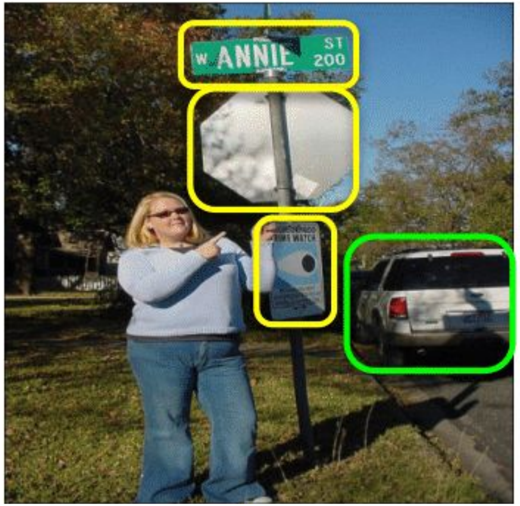
- **Given an image and a text, RIS predicts a segmentation mask of the object referred.**
- **The key to RIS is to discern the referent among visually similar objects via textual cues.**

a young woman in blue shirt and striped pants sitting in the snow

a skier in an orange jacket bending over

What makes RIS difficult?

- **The difficulty of each RIS scenario can be affected by the degree of visual ambiguity in the scene given the linguistic complexity of the referring expression.**



(1) *"a sign lettered 'ANNIE' between woman and SUV"*
(2) *"a parked white FORD SUV"*

Motivation - the gap between easy & hard scenarios

- **We manually pick 100 easy and hard samples depending on the number of negative objects.**
- **A huge performance gap exists between easy & hard examples in current models.**

Table 2: mIoU & oIoU on 100 easy and hard samples from G-Ref UMD test set

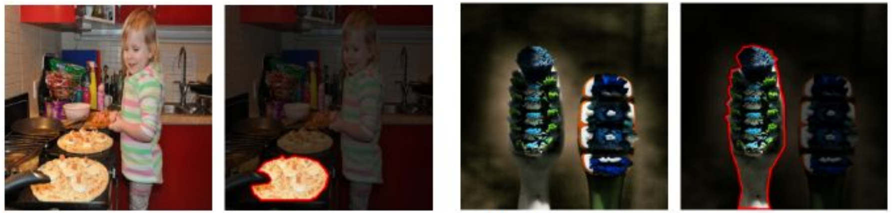
	mIoU		oIoU	
Models	**Easy**	**Hard**	**Easy**	**Hard**
LAVT [50]	78.26	54.61	79.16	47.40
CRIS [45]	76.89	52.97	78.81	43.20
CGFormer [43]	79.86	61.22	79.95	53.27

Fig. I: Easy samples from G-Ref test split

"A little girl in a blue dress"

"A yellow train with black trim"

Fig. II: Hard samples from G-Ref test split



"An uncooked pizza with four hotdogs"

"A white toothbrush with green, blue and white bristles"

Motivation - training data challenging enough?

- **Variant Inter and even Intra-dataset grounding difficulty levels exist in training data as well.**
- **We ask if these samples are challenging enough to discern subtle visual and textual nuance for RIS.**

Mosaic Image Augmentation for RIS

- **In this work, we propose a data augmentation method that generates ambiguous examples where a model is encouraged to concretely understand the scene and the query.**

"A woman standing in front of the wall"   "A woman standing in front of the wall"

Figure 6. expected augmentation effect of our method

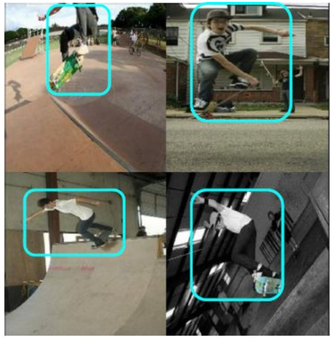
Overall Pipeline

- **NeMo: Negative Mining + Mosaic Augmentation**
- **Filtering is necessary for the right level of ambiguity, and to avoid invalid mosaics.**

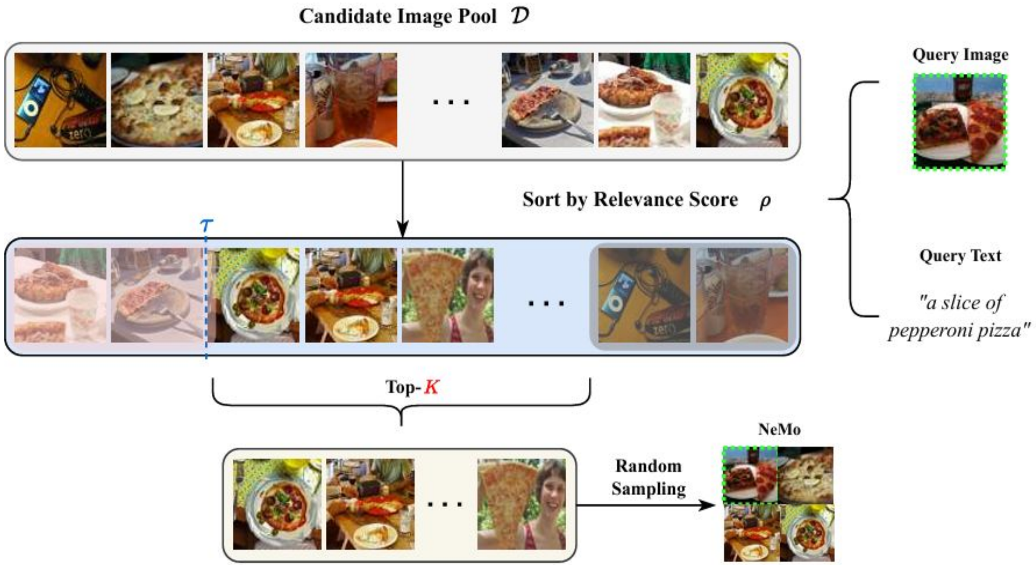
easy

“the rightmost pizza on a paper plate”

invalid



“a man jumping with a skateboard”



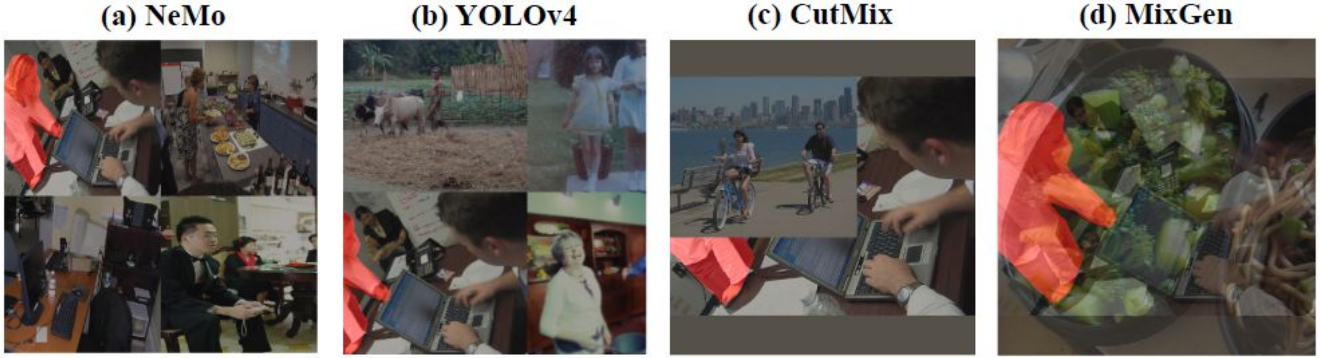
Overall Comparison

- **Overall RIS performance (oIoU) comparison w/ and w/o NeMo**
  - We observe a larger performance boost on more complex datasets.
  - Harder datasets benefit more because of its intricate referring expressions and visually dense scenes.

RIS model	NeMo	RefCOCO (UNC) Val	RefCOCO (UNC) TestA	RefCOCO (UNC) TestB	RefCOCO+ (UNC) Val	RefCOCO+ (UNC) TestA	RefCOCO+ (UNC) TestB	G-Ref (UMD) Val	G-Ref (UMD) Test	GRES Val	Average Gain
LAVT [50]	✗	72.73	75.82	68.79	62.14	68.38	55.10	61.24	62.09	57.64	$+1.92_{\pm 2.34}$
	✓	**73.25**	**76.12**	**69.67**	**62.52**	**69.95**	**56.02**	**63.40**	**64.95**	**65.35**	
CRIS [45]	✗	66.68	70.62	59.93	56.94	64.20	46.97	55.91	58.50	54.55	$+1.73_{\pm 0.82}$
	✓	**68.66**	**72.82**	**63.06**	**57.94**	**65.25**	**48.41**	**58.47**	**59.07**	**56.23**	
ReLA [28]	✗	73.67	76.18	70.39	63.82	68.70	55.78	65.22	65.29	63.10	$+0.97_{\pm 0.82}$
	✓	**74.24**	**77.11**	**70.39**	**65.35**	**70.55**	**56.68**	**65.32**	**65.73**	**65.54**	
CGFormer [43]	✗	72.53	75.12	70.09	63.55	68.58	56.05	62.92	64.63	64.77	$+1.04_{\pm 0.67}$
	✓	**73.52**	**76.07**	**70.92**	**64.30**	**69.58**	**57.85**	**65.31**	**65.07**	**65.00**	
VPD [58]	✗	73.46	75.31	70.23	61.41	67.98	54.99	63.12	63.59	62.38	$+1.47_{\pm 0.85}$
	✓	**74.48**	**76.32**	**71.51**	**62.86**	**69.92**	**55.56**	**64.40**	**64.80**	**65.89**	
Average Gain		$+1.11_{\pm 0.79}$			$+1.21_{\pm 0.48}$			$+1.55_{\pm 0.99}$		$+3.11_{\pm 2.83}$	

Table 3: Overall RIS performance (in oIoU) comparison with and without NeMo

Comparison to other augmentation methods



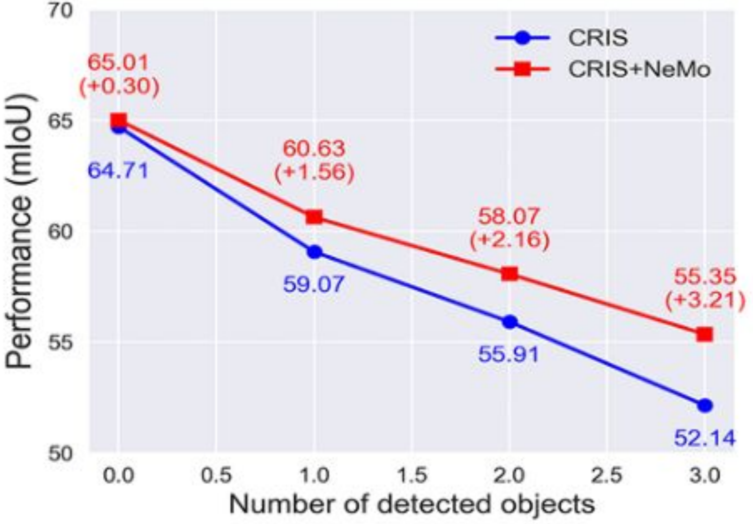
Query : A woman in a white shirt looking down at a laptop

Augmentation Method	oIoU Val	oIoU Test	Prec (Val) 0.5	Prec (Val) 0.7
CRIS	55.91	58.50	67.95	54.84
+YOLOv4 [3]	56.22	58.55	66.94	53.54
+CutMix [56]	56.50	58.34	66.63	53.11
+MixGen [13]	53.62	55.85	64.37	51.28
+NeMo (Ours)	**58.47**	**59.07**	**70.01**	**56.60**

Detailed Analysis (1)

- **Performance on Visually Challenging Scenarios**
  - better in challenging cases with more negative objects.
- **Performance w.r.t Query Complexity**
  - robust at sentence lengths, even with longer complex ones.

(1) Performance on Visually Challenging Scenarios

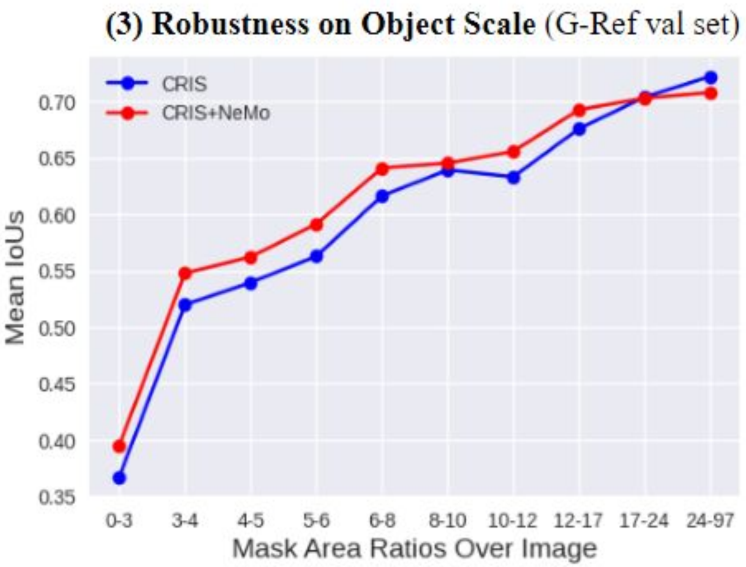


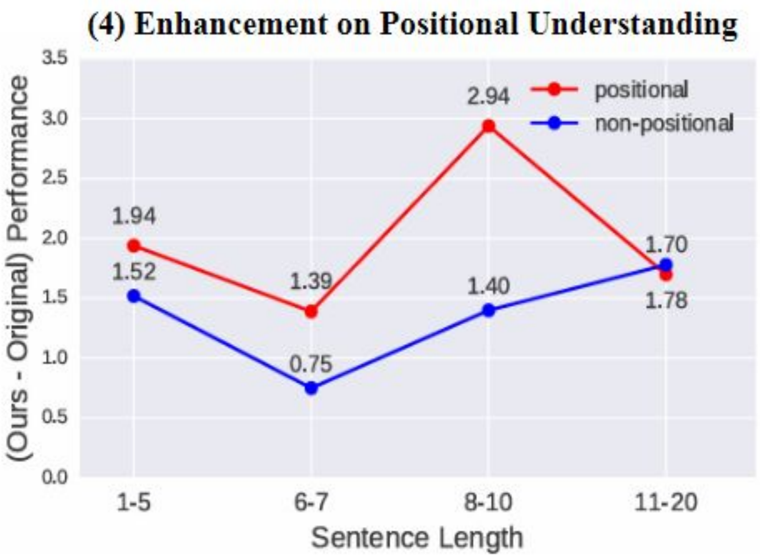
(2) Performance w.r.t Query Complexity

RIS model	NeMo	Length of T 1-5	6-7	8-10	11-20
LAVT [50]	✗	63.95	63.46	63.03	63.00
	✓	**66.50**	**65.39**	**64.40**	**64.72**
CRIS [45]	✗	58.91	56.41	55.29	57.33
	✓	**60.77**	**57.17**	**57.05**	**58.35**
ReLA [28]	✗	**66.67**	64.95	**63.82**	65.95
	✓	66.63	**65.00**	63.75	**67.26**
CGFormer [43]	✗	65.85	65.12	**64.33**	63.87
	✓	**66.30**	**65.44**	63.98	**64.98**
VPD [58]	✗	**67.53**	66.12	65.49	67.44
	✓	66.30	**66.86**	**67.33**	**68.12**

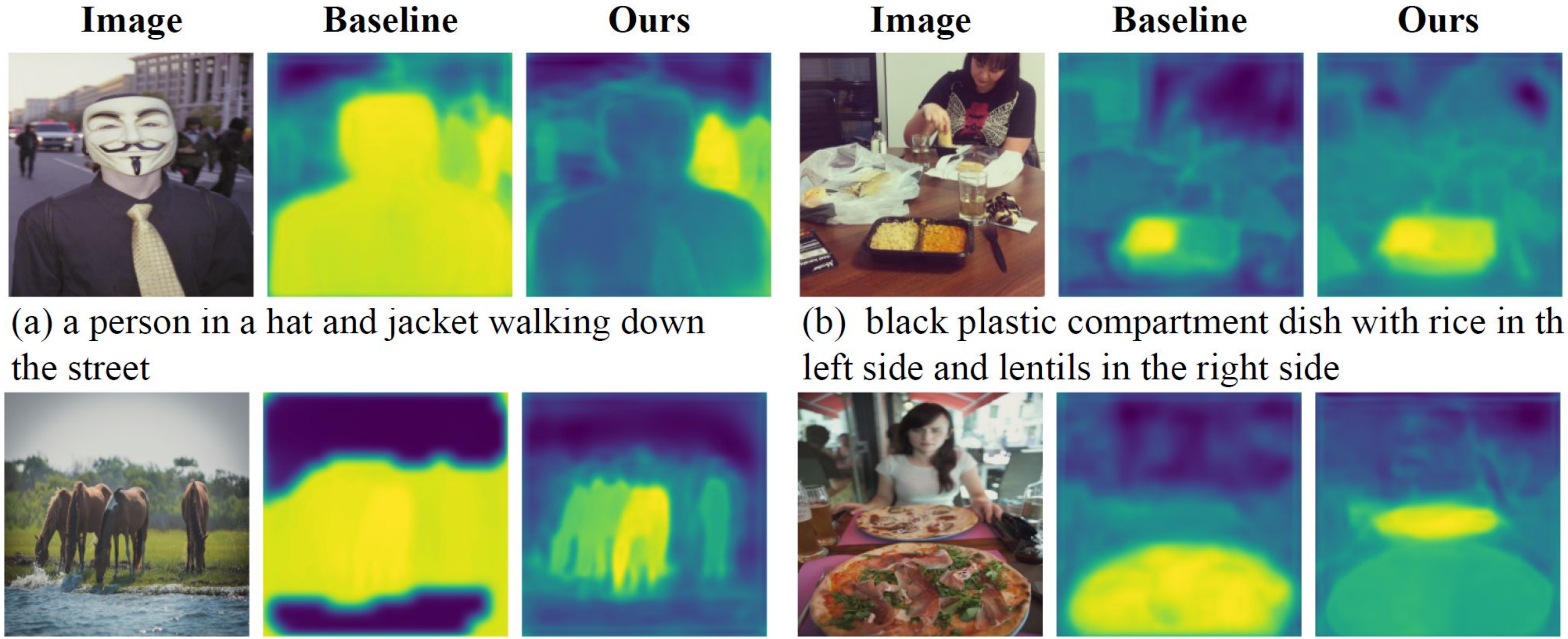
Detailed Analysis (2)

- **Robustness on Object Scale**
  - better in most object sizes, especially for smaller objects.
- **Enhancement on Positional Understanding**
  - better at positional keywords, even in long and complex queries.





Qualitative Analysis (1)



(a) a person in a hat and jacket walking down the street

(b) black plastic compartment dish with rice in th left side and lentils in the right side

(c) the second horse from the right

(d) a pizza on a plate in front of a woman

Fig. 9: Visualization of activation maps with and without NeMo on CRIS

Qualitative Analysis (2)

"a lady in a black dress cuts a wedding cake with her new husband"

"a slice of cheese cake at the top of the fork"

Figure 7. Visualization of results after augmentation on CRIS [45].

Summary

- We introduce NeMo, Negative-mined Mosaic Augmentation, a simple but powerful labor-free data augmentation method for Referring Image Segmentation.

- NeMo involves a systematic way to tune the dataset difficulty by generating training examples at a properly controlled difficulty.

- NeMo brings consistent IoU improvement over various state-of-the-art RIS models on multiple datasets.

- NeMo enhances both visual and textual understanding capabilities for segmenting the right target.