NEURAL INFORMATION PROCESSING SYSTEMS

Queen Mary University of London

UNIVERSITY OF OXFORD

Benchmarking Robustness of Text-Image Composed Retrieval

NeurIPS Ro-FoMo Workshop

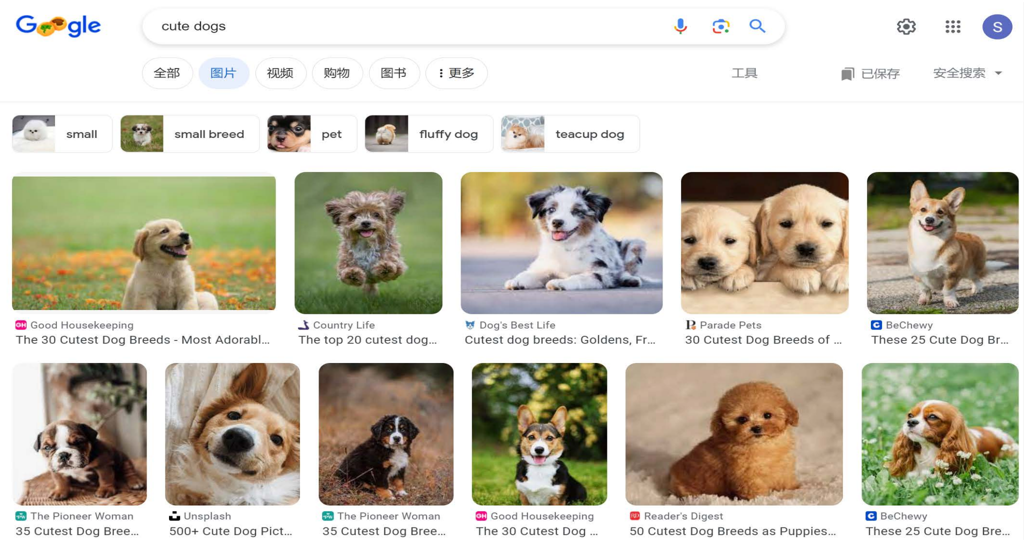


Shitong Sun
QMUL

Jindong Gu
Oxford

Shaogang Gong
QMUL

Image retrieval



text-image retrieval

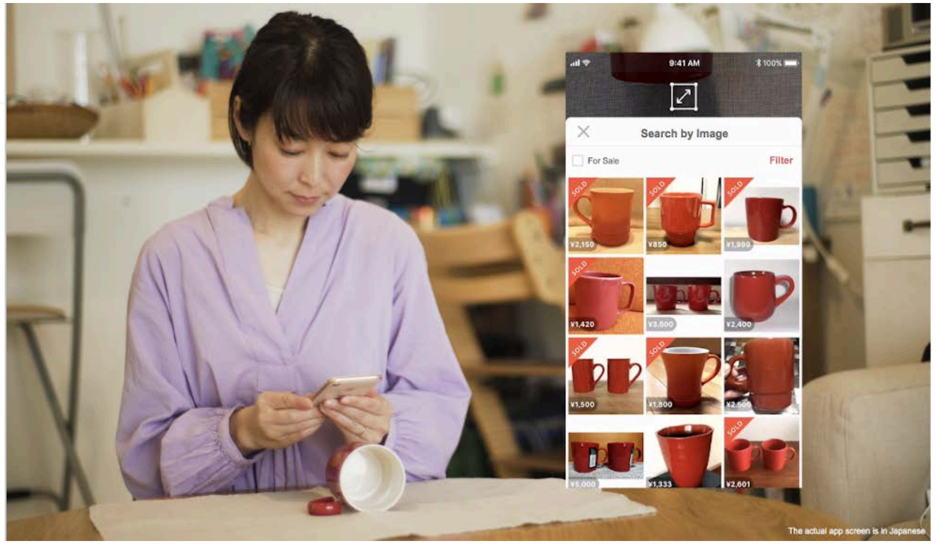
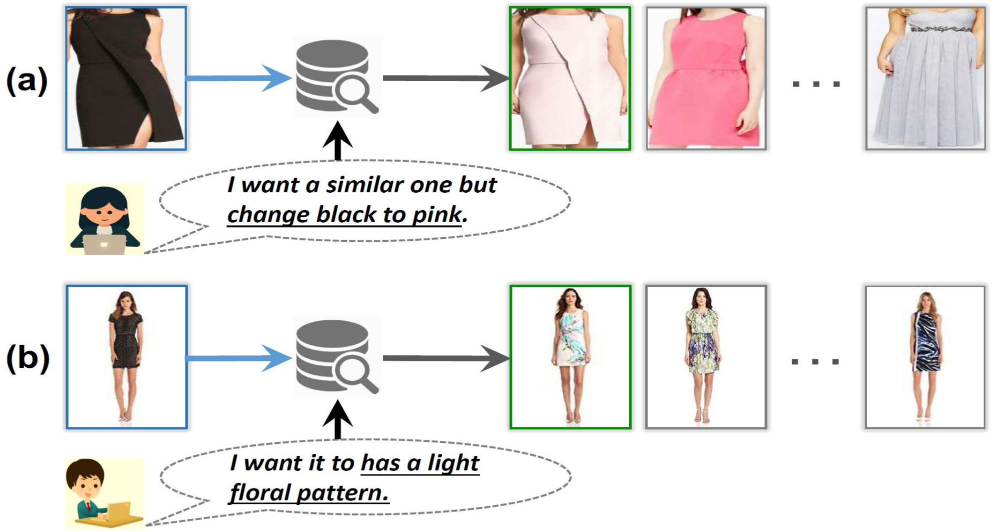


image-image retrieval

person reidentification



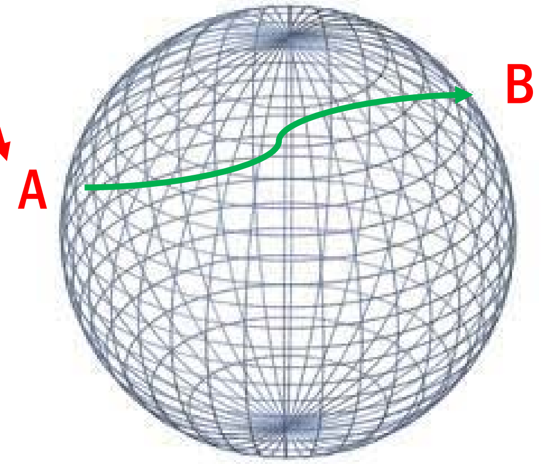
Text-image composed retrieval/composed image retrieval

Foundation of Text-Image Composed Retrieval

A picture is worth a thousand words

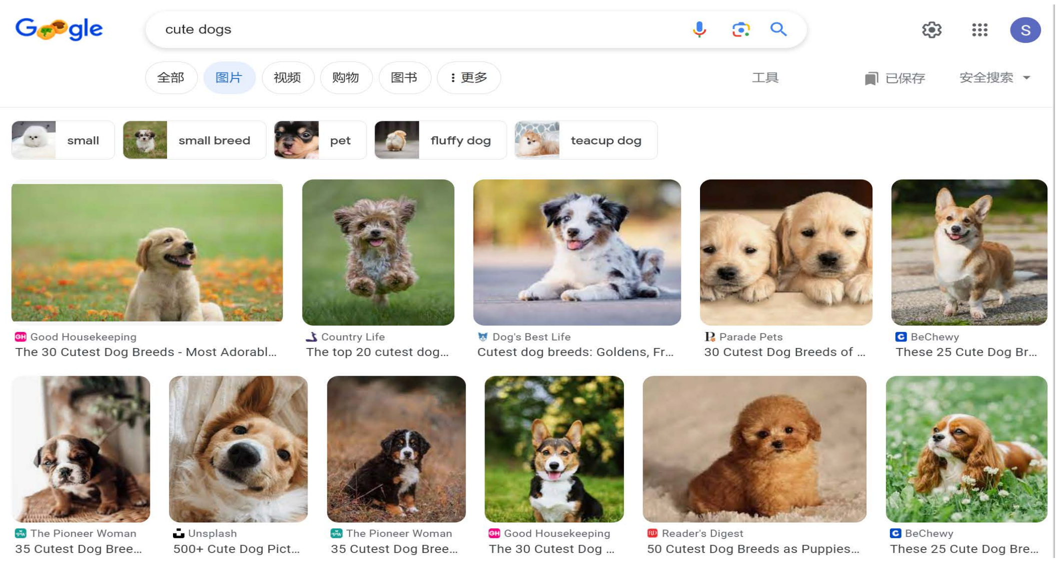
Image space is dense and continues

[Dalle2 example, OpenAI, 2022]



Text to gap the subtle or significant difference

Text space is sparse and discrete



- Image representation to supply a precise anchor in the dense continuous visual space.
- Text representation to supply subtle or significant differences between visual contents
- Generalize sparse modified text attributes to dense reference images

Motivation & Definition

Text-image composed retrieval:

- Real-world application: fashion domain e-commerce; open domain internet search
- Robustness of real-world application is **crucial**

Definition of robustness in text-image composed retrieval:

- Robustness against natural corruption including both visual and textual
- Robustness against text understanding

"Is black with lawyer written on it."

+

Swap *"**Ptu** the parrot in the basket with toys."*

Qwerty *"Put the parrot in **5he basktd** with toys."*

Repetition *"Put the parrot in the basket **with with** toys."*

Homophones *"Put the parrot **inn** the basket with toys."*

"Change to a bathroom with white vanity and two mirrors."

Evaluation Metrics

Evaluation metrics:

- Robustness against natural corruption including both visual and textual: relative robustness

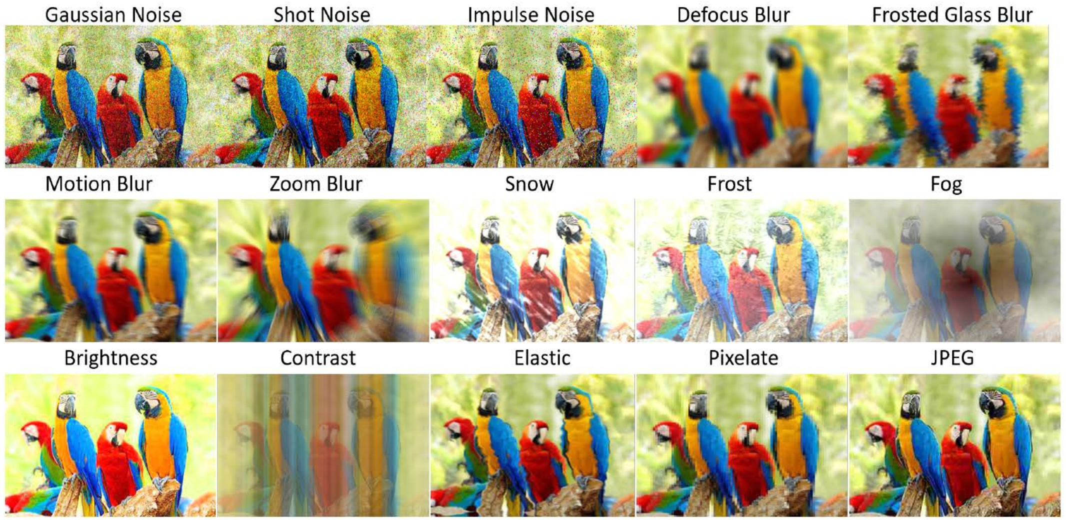
$$\gamma = 1 - \left(R_c - R_p\right) / R_c$$

- Robustness against text understanding: Recall@5

Three evaluation datasets:

- FashionIQ-C: fashion domain with 15 visual corruptions and 7 textual corruptions
- CIRR-C: open domain with 15 visual corruptions and 7 textual corruptions
- CIRR-D: Diagnostic dataset with text variations including number variations, attribute variations (color, shape, size), object removal, and background variations and fine-grained variations.

Visual corruptions (Noise, blur, weather and digital):



(a) Sample visualization with 15 standard image corruptions.

Textual corruptions(character level and word level):

Original text:
There were two adult dogs on the road - there was one grown puppy in the yard

character_filter
'There were two adutl dogs on teh road - there wsa oen grown puppy in the yard.'

qwerty_filter
'There were two adult dogs on the road - there was one grow5 puppy in the yard.'

RemoveChar_filter
'Thre were two adult dogs on the road - tere ws ne grown puppy in the yard.'

remove_space_filter
'There were two adult dogs onthe road - there was one grown puppy in the yard.'

misspelling_filter
'There were were two adult dogs on the road - there was one grown puppy in the yard.'

repetition_filter
'There were two adult adult dogs on the road - there was one grown grown puppy in the yard.'

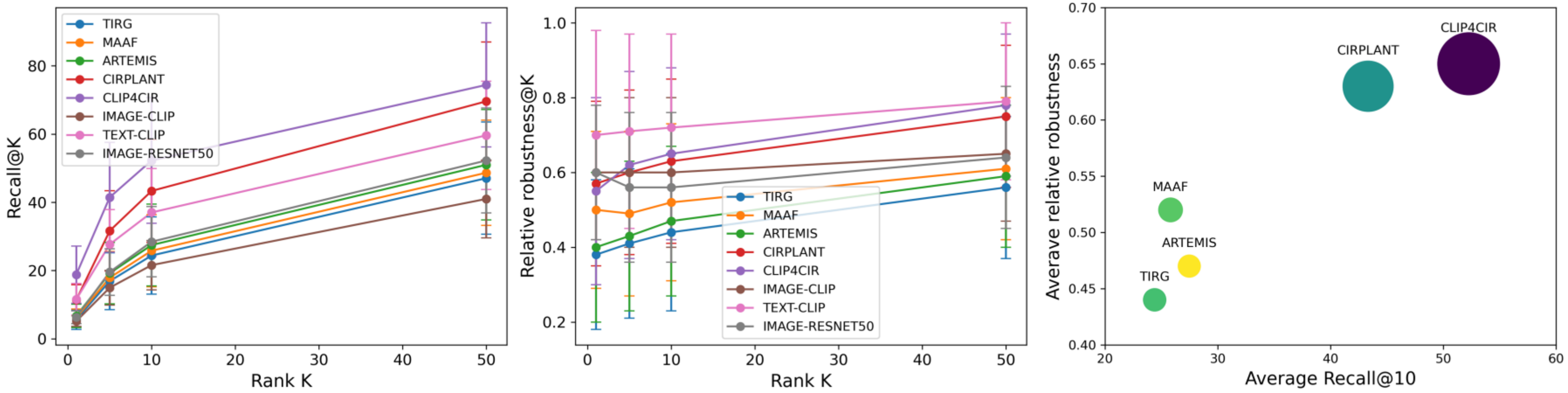
Result Analysis of Natural Corruptions

	Noise				Blur				Weather				Digital			
CIRR-C	Clean	Gauss.	Shot	Impluse	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Bright	Contrast	Elastic	Pixel	JPEG
Image-only(RN50)	50.4	0.57	0.55	0.58	0.68	0.28	0.82	0.45	0.38	0.34	0.64	0.86	0.20	0.48	0.76	0.88
Image-only(CLIP)	36.2	0.56	0.55	0.58	0.66	0.32	0.83	0.49	0.52	0.45	0.77	0.91	0.24	0.41	0.78	0.91
Text-only(CLIP)	51.2	0.79	0.76	0.81	0.85	0.29	1.0	0.55	0.65	0.70	0.89	1.0	0.19	0.40	0.96	1.0
TIRG [40]	55.1	0.34	0.36	0.34	0.48	0.21	0.70	0.43	0.31	0.22	0.40	0.70	0.12	0.47	0.74	0.84
MAAF [8]	49.9	0.50	0.49	0.50	0.62	0.26	0.80	0.41	0.36	0.31	0.50	0.74	0.11	0.48	0.83	0.87
ARTEMIS [7]	59.0	0.39	0.42	0.38	0.51	0.25	0.70	0.44	0.31	0.26	0.45	0.71	0.10	0.47	0.75	0.86
CIRPLANT [25]	68.8	0.70	0.69	0.71	0.77	0.28	0.89	0.51	0.44	0.43	0.66	0.88	0.17	0.56	0.85	0.92
CLIP4CIR [2]	**80.3**	0.68	0.68	0.69	**0.77**	**0.28**	**0.90**	**0.52**	**0.55**	**0.60**	**0.80**	**0.91**	0.16	0.39	**0.91**	**0.92**

	Noise				Blur				Weather				Digital			
FashionIQ-C	Clean	Gauss.	Shot	Impluse	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Bright	Contrast	Elastic	Pixel	JPEG
TIRG [40]	23.8	0.28	0.26	0.23	0.34	0.22	0.61	0.57	0.32	0.27	0.37	0.61	0.12	0.64	0.85	0.85
MAAF [8]	23.4	0.31	0.27	0.25	0.44	0.21	0.67	0.53	0.29	0.24	0.31	0.54	0.13	0.54	0.83	0.83
ARTEMIS [7]	24.9	0.24	0.24	0.20	0.38	0.26	0.65	0.60	0.36	0.25	0.38	0.55	0.14	0.63	0.86	0.87
FashionViL [13]	23.4	0.26	0.28	0.25	0.40	**0.31**	**0.82**	**0.67**	0.33	0.31	0.34	**0.70**	0.15	**0.86**	**1.09**	**1.06**
CLIP4CIR [2]	**35.9**	**0.44**	**0.42**	**0.44**	**0.54**	0.21	0.72	0.50	**0.46**	**0.43**	**0.60**	**0.70**	**0.22**	0.37	0.74	0.83

		Character				Word		
CIRR-C	Clean	Swap	QWERTY	RemoveChar	RemoveSpace	Misspelling	Repetition	Homophone
Text-only	51.2	0.75	0.74	0.78	1.0	0.99	0.98	0.92
TIRG [40]	55.1	0.77	0.76	0.80	1.0	0.98	1.0	0.89
MAAF [8]	49.9	**0.95**	**0.97**	**0.96**	**1.0**	**1.0**	**1.0**	**0.97**
ARTEMIS [7]	59.0	0.61	0.58	0.65	**1.0**	0.98	0.98	0.82
CIRPLANT [25]	68.8	0.92	0.93	0.93	**1.0**	**1.0**	**1.0**	**0.97**
CLIP4CIR [2]	**80.3**	0.89	0.89	0.90	**1.0**	**1.0**	0.99	**0.97**

		Character				Word		
FashionIQ-C	Clean	Swap	QWERTY	RemoveChar	RemoveSpace	Misspelling	Repetition	Homophone
TIRG [40]	23.8	0.26	0.20	0.29	0.66	0.63	0.61	0.52
MAAF [8]	23.4	0.40	0.39	0.39	0.70	0.68	0.68	0.62
ARTEMIS [7]	24.9	0.25	0.20	0.31	0.70	0.67	0.67	0.55
FashionViL [13]	23.4	**0.55**	**0.59**	**0.60**	**0.86**	**0.84**	**0.85**	**0.76**
CLIP4CIR [2]	**35.9**	0.52	0.51	0.54	0.71	0.70	0.69	0.67



Observation & results: + larger models also have better robustness + multi-task training may boost robustness
+ text features from aligned space can help boost the robustness, while independent space will damage the robustness

Robustness against text understanding

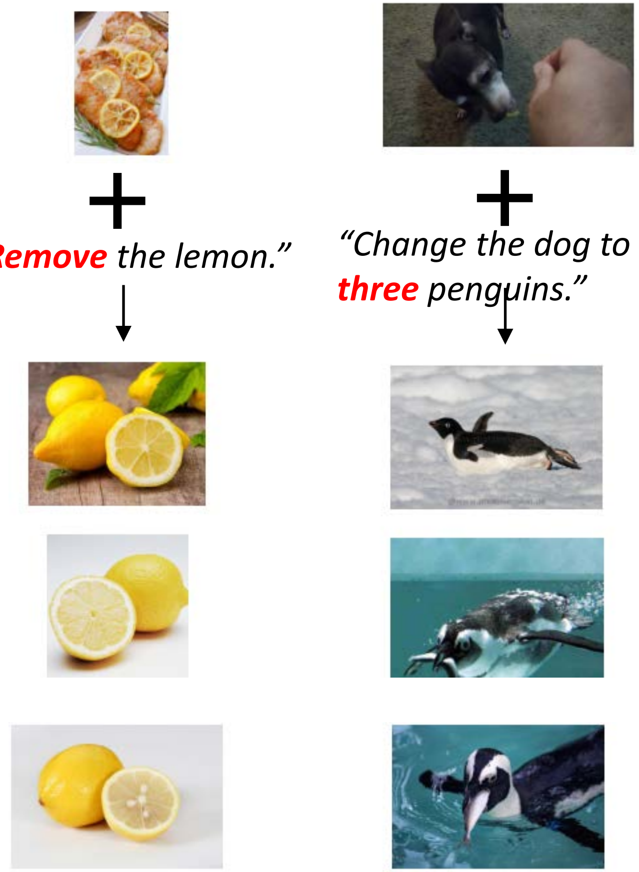
Reference image:

Modified text

Ranked-1

Ranked-2

Ranked-3



Our reproduced result [CLIP4Cir, CVPR2022].

Current models have difficulty of **text understanding.**

We build a benchmark to detect:

+ numerical variations
+ attribute variations
+ object removal
+ background variations
+fine-grained vatiations

Based on CIRR, we generate text variations and corresponding images

	Images	Numerical	Attribute	Removal	Background	Fine-grained
Val.	2297	820	1397	233	358	4181
Extend caption	-	-	-	505	812	-
Synthetic	1245	305	700	140	-	-
Total	3542	1125	2097	878	1170	4181

CIRR-D with generated images

Reference Image

Target Images

1. *"Change to **two** blue and red Pepsi Colas are on the bus."*
2. *"Change to **three** blue and red Pepsi Colas are on the bus."*
3. *"Change to **four** blue and red Pepsi Colas are on the bus."*

4. *"Change to **two** dogs with pink backgrounds are lounging on the couch."*
5. *"Change to **three** dogs with pink backgrounds are lounging on the couch."*
6. *"Change to **four** dogs with pink backgrounds are lounging on the couch."*

Numerical

Reference Target

1. *"Change to **remove** the lemon."*

Reference Target

2. *"Change to **remove** the lemon."*

Reference Target

3. *"Change to remove the hand."*

Reference Target

4. *"Change to remove the horse drawn carriage."*

Object removal

Reference Image

Target Images

*1~5. "Change to a bathroom with a **grey** vanity with **fewer drawers**./ **white** vanity/**dark brown** vanity/ **light brown** vanity/ **small brown** vanity and two mirrors."*

*6~10. "Change to a **spotted red and black / striped / blue / an orange and yellow / purple** stingfish in the sand."*

Color/shape/Size

CIRR-D uses extend caption of and sub-category of CIRR

1. "Add green grass, trees and humans on the background."

2. "Add a beach in the background."

3. "Change to white background."

4. "Snow on the background."

Background variations

Fine-grained variations

"A seal laying down on the sand and touch its mouth."

Results analysis of text understanding

Table 5: Recall of CIRR-D dataset. The red and green arrows indicate the performance increase of decrease compared with CIRR queries. **Bold** and underline are the largest decrease and increase.

	R@5					Rsub@1
	CIRR	Numerical	Attribute	Removal	Background	Fine grained
Image-only(ResNet50)	31.55	31.47 ↓ (0.08)	32.57 ↑ (1.02)	35.99 ↑ (4.44)	39.15 ↑ (7.60)	20.25
Image-only(CLIP)	22.51	24.80 ↑ (2.29)	29.09 ↑ (6.58)	27.90 ↑ (5.39)	25.64 ↑ (3.13)	20.02
Text-only	39.02	42.84 ↑ (3.82)	49.45 ↑ (10.43)	11.62 ↓ (27.4)	11.62 ↓ (**27.4**)	53.73
TIRG [36]	36.35	39.64 ↑ (3.29)	37.77 ↑ (1.42)	30.41 ↓ (5.94)	32.82 ↓ (3.53)	35.90
MAAF [7]	32.19	32.53 ↑ (0.34)	35.57 ↑ (3.38)	31.09 ↓ (1.10)	34.27 ↑ (2.08)	28.63
ARTEMIS [36]	40.05	39.56 ↓ (0.49)	42.68 ↑ (2.63)	33.26 ↓ (6.79)	35.56 ↓ (4.49)	40.80
CIRPLANT [23]	48.82	45.07 ↓ (**3.75**)	47.73 ↓ (**1.09**)	41.12 ↓ (7.70)	45.98 ↓ (2.84)	38.19
CLIP4CIR [2]	62.94	64.18 ↑ (1.24)	69.15 ↑ (6.21)	31.66 ↓ (**31.28**)	41.88 ↓ (21.06)	62.66

Observation & results:

+ models gain stronger discriminative ability for attribute, instead of object removal and background
+ text guidance expands the possibility of the targets, which over guidance the model decision.
+ a modified text offers accurate information while minimizing the number of feasible targets can enhance the model's discriminative ability

Summary

- ❑ Model pre-trained on large datasets will lead to better robustness
- ❑ Multi-task training may boost performance
- ❑ Text features help boost the robustness when its
    - from aligned space of image feature
    - Minimize the number of feasible targets