Context-aware Difference Distilling for Multi-change Captioning

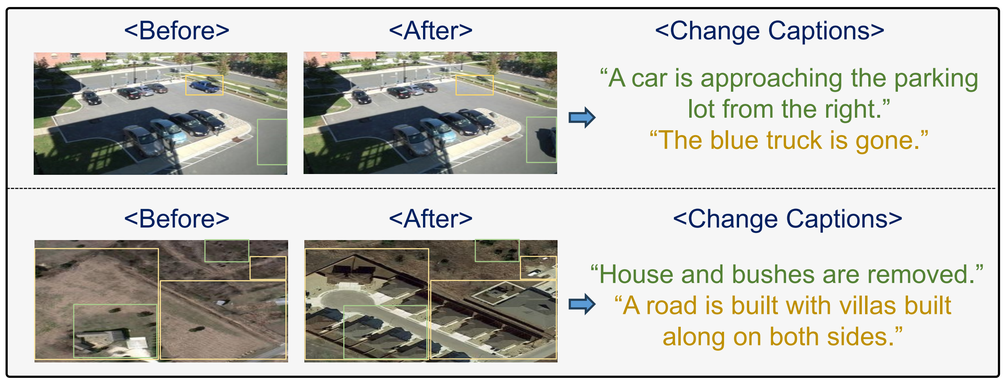
Yunbin Tu[1], Liang Li[2], Li Su[1], Zheng-Jun Zha[3], Chenggang Yan[4] and Qingming Huang[1]

[1]University of Chinese Academy of Sciences, [2]Institute of Computing Technology, CAS,
[3]University of Science and Technology of China, [4]Hangzhou Dianzi University

ACL 2024 Bangkok, Thailand

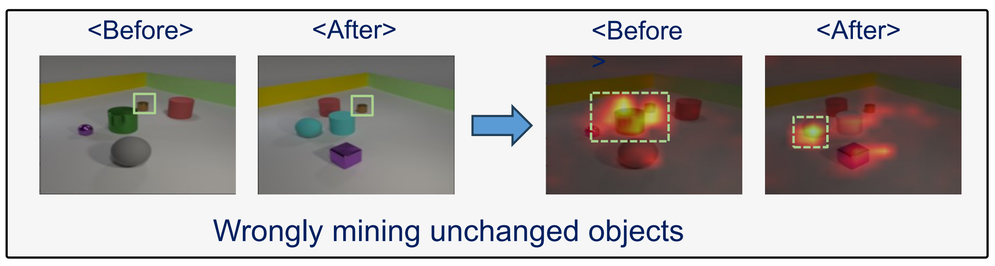
Problem Definition and Contribution

Goal: Multi-change captioning aims to describe complex changes within an image pair in natural language.

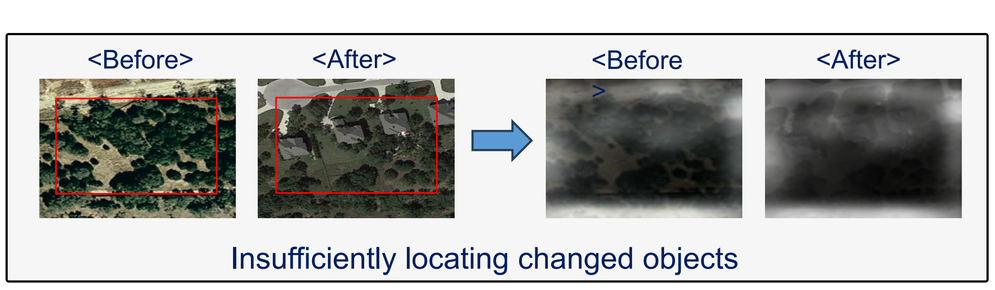


Motivations:

- Existing methods directly match patch features of image pair, wrongly mining unchanged objects.



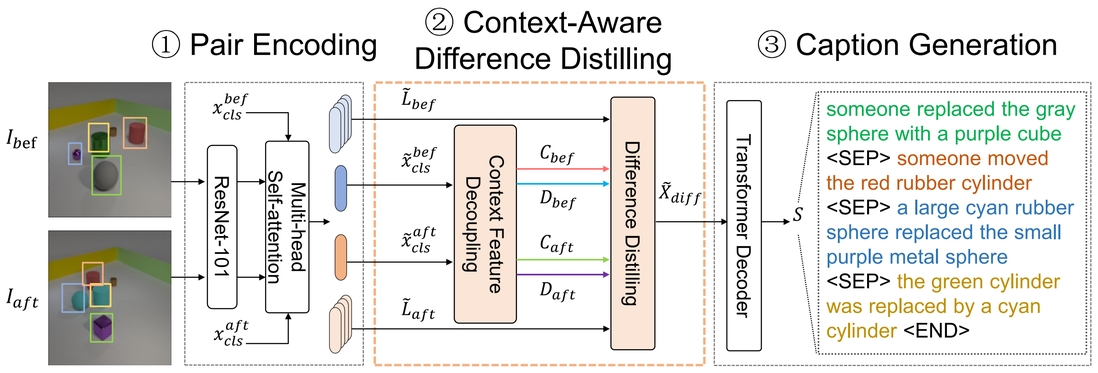
- Existing methods directly model differences by patch features, insufficiently locating changed objects.



Contributions:

- Modeling commonality / difference context features, before learning locally common / difference features.
- Proposing CARD to first decouple the context features; use them to help capture all changes.
- Customizing consistency / independence constraints to guarantee alignment / discrepancy of commonality / difference context features.

Approach Overview



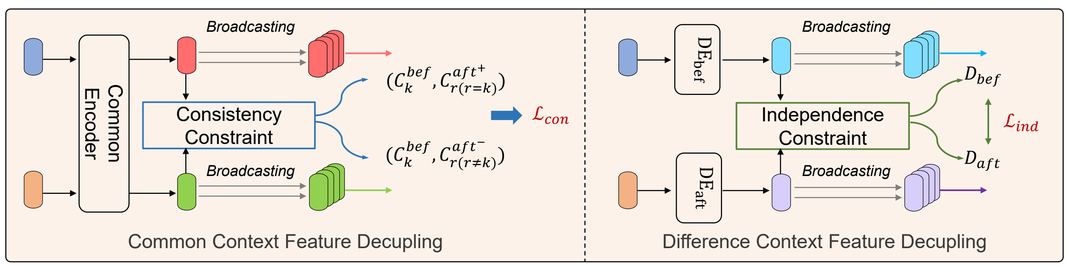
① Extracting n patch features for each image; introducing a [CLS] feature to represent its global content.

② Disentangling common and difference context features. The former helps mine locally common features for deducing locally difference features; the latter augments the locally difference features to distill all changes.

③ Decoding the omni-representation of all changes into natural language sentences by a transformer decoder.

Context-Aware Difference Distilling

Context feature decoupling:



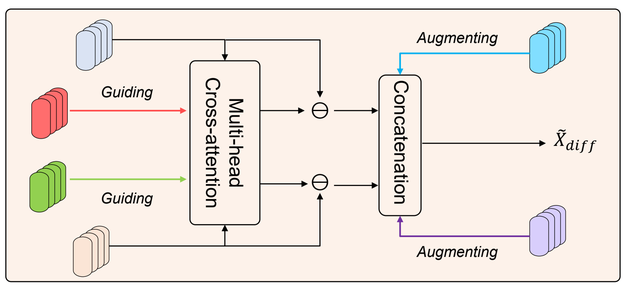
(1) Common and difference context feature predicting:

$$C_{bef(aft)} = \mathcal{CE}(x_{cls}^{bef(aft)}; \theta_c) \qquad D_{bef(aft)} = \mathcal{DE}(x_{cls}^{bef(aft)}; \theta_{bef(aft)})$$

(2) Consistency constraint: $\mathcal{L}_{con} = CA(C_{bef}, C_{aft})$ • (CA: Contrastive Alignment)

(3) Independence constraint: $\mathcal{L}_{ind} = HSIC(D_{bef}, D_{aft})$ • (HSIC: Hilbert-Schmidt Independence Criterion)

Difference distilling:



(1) Integrating $C_{bef(aft)}$ into patch features:

$$\tilde{X}'_{bef} = [C_{bef}, x_1 \ldots, x_n]$$
$$\tilde{X}'_{aft} = [C_{aft}, x_1 \ldots, x_n]$$

(2) $C_{bef(aft)}$ guided locally common feature mining:

$$\tilde{X}_{bef}^c = \mathrm{MHCA}(\tilde{X}'_{bef}, \tilde{X}'_{aft}, \tilde{X}'_{aft})$$
$$\tilde{X}_{aft}^c = \mathrm{MHCA}(\tilde{X}'_{bef}, \tilde{X}'_{aft}, \tilde{X}'_{aft})$$

(3) $D_{bef(aft)}$ augmented locally difference feature disentangling:

$$\tilde{X}_{bef}^d = [\tilde{L}_{bef} - \tilde{X}_{bef}^c; D_{bef}]$$
$$\tilde{X}_{aft}^d = [\tilde{L}_{aft} - \tilde{X}_{aft}^c; D_{aft}]$$
$$\tilde{X}_d = ReLU([\tilde{X}_{bef}^d; \tilde{X}_{aft}^d]W_c + b_c)$$

Experimental Results

Comparison with existing methods:

CLEVR-Multi-Change

Method	B	M	R	S	C
DUDA	41.8	36.2	53.9	64.7	283.5
M-VAM	37.1	34.0	51.5	62.2	242.9
MCCFormers-S	55.9	44.8	56.8	76.6	378.6
MCCFormers-D	56.2	44.8	57.3	76.6	383.2
VARD-Trans	48.1	41.8	55.5	72.1	344.2
SCORER+CBR	56.4	44.9	57.1	76.7	388.0
CARD (Ours)	**56.7**	**45.2**	**57.4**	**76.9**	**391.6**

LEVIR-CC

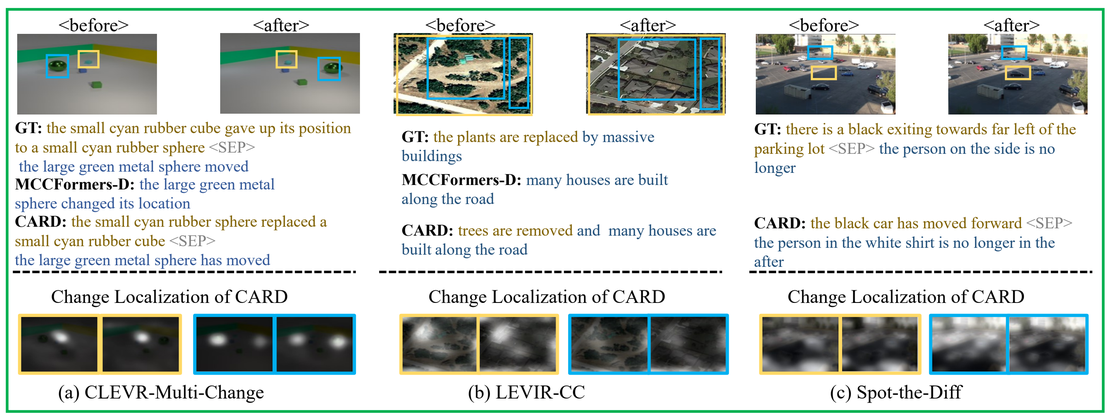
Method	B	M	R	C
DUDA	57.8	37.2	71.0	124.3
MCCFormers-S	56.7	36.2	69.5	120.4
MCCFormers-D	56.4	37.3	70.3	124.4
RSICCFormer	62.8	39.6	74.1	134.1
PSNet	62.1	38.8	73.6	132.6
Prompt-CC (soft)	62.4	38.6	73.4	135.3
Prompt-CC (hard)	63.5	38.8	73.7	136.4
Chg2Cap	64.4	**40.0**	**75.1**	136.6
CARD (Ours)	**65.4**	**40.0**	74.6	**137.9**

Ablation study:

Ablative Variants	CCF	DCF	CLEVR-Multi-Change B	M	R	S	C	LEVIR-CC B	M	R	C
Baseline	×	×	54.7	43.6	56.7	75.6	362.3	60.7	36.3	69.7	120.0
Baseline	✓	×	56.5	45.1	57.1	76.8	385.8	63.5	38.5	72.3	130.4
Baseline	×	✓	56.5	45.0	57.1	**77.0**	385.7	60.6	37.6	71.0	125.9
Baseline	✓	✓	**56.7**	**45.2**	**57.4**	76.9	**391.6**	**65.4**	**40.0**	**74.6**	**137.9**

Ablative Variants	CC	IC	CLEVR-Multi-Change B	M	R	S	C	LEVIR-CC B	M	R	C
CARD	×	×	54.6	44.1	57.2	75.8	363.7	55.9	35.6	72.3	132.2
CARD	✓	×	56.2	44.8	57.1	76.8	384.2	56.2	35.8	72.6	137.6
CARD	×	✓	56.5	45.1	57.2	**77.0**	389.9	60.6	37.7	72.5	133.0
CARD	✓	✓	**56.7**	**45.2**	**57.4**	76.9	**391.6**	**65.4**	**40.0**	**74.6**	**137.9**

Visualization for change locating and captioning:



- Code available at: https://github.com/tuyunbin/CARD