ReCon: Enhancing True Correspondence Discrimination through Relation Consistency for Robust Noisy Correspondence Learning

Supplementary Material

1. Introduction

In this supplemental material, we provide additional information regarding the implementation and evaluation of ReCon. Specifically, we first give the implementation details of ReCon across different extended baselines on three benchmark datasets, enhancing its reproducibility. Additionally, we present extensive experimental results and analysis to further validate the effectiveness and superiority of ReCon. Finally, we include visual examples of image and text retrieval on Flick30K, as well as the detected noisy examples from CC152K, showcasing the practical outstanding performance and robustness of ReCon.

2. Implementation and Training Details

2.1. Model Settings

In this section, we elaborate on the model settings and the implementation details of ReCon. To thoroyghly validate the effectiveness of ReCon, we integrate it with SGR [4], SAF [4], and SGRAF [4], aiming to assess its robustness against noisy correspondence (NC). Similar to DECL [11] and CRCL [12], ReCon is directly performed on the similarity outputs of these models without modifying their models. For all experiments, we maintain consistency by using the same image region features and text backbone. Specifically, we utilize the Faster R-CNN [13] detection model to extract local-level BUTD features, selecting the top-36 salient regions for each image based on confidence scores. There features are encoded into 2,048-dimensional vectors and subsequently projected into 1,024-dimensional image representations in the shared semantic space. Text representations are generated using a Bi-directional GRU [1], which encodes word tokens into the same 1,024-dimensional space as the image features. Additionally, following GPO [3], we employ size augmentation on the training data to enhance model performance. All experiments are conducted under identical conditions to ensure fairness, and the code for ReCon will be made publicly available on GitHub.

2.2. Parameter Settings

In this section, we elaborate the parameter settings used in our experiments, summarized in Table. 3, to facilitate reproducibility across three benchmark datasets: Flickr30K, MS-COCO, and CC152K. The parameters are categorized into two groups. The first group pertains to training without noise, while the second focuses on training under simulated or real-world noise. Note that the results of ReCon reported in our main paper and supplement material are obtained by ensembling ReCon-SAF and ReCon-SGR. Following [4, 8, 12], the ensemble strategy involves averaging the similarities computed by the two models before conducting cross-modal retrieval. Next, we will describe these main parameters. Specifically, τ denotes the temperature coefficient and β is the momentum coefficient. ω_1 and ω_2 represent the division thresholds, and α is the penalization factor. Moreover, γ and ξ, respectively, denote the fixed margin of triplet loss and balance factor.

Table 1. Analysis on different batch sizes on Flickr30K with 40% noise rate. The best results are marked by **bold**.

Methods	Image to Text			Text to Image			
	R@1	R@5	R@10	R@1	R@5	R@10	rSum
32	76.6	92.9	96.7	57.0	82.8	89.1	495.2
64	76.8	**94.4**	97.5	58.6	83.7	89.9	500.8
128	**79.4**	94.3	**97.6**	59.9	**83.9**	**90.1**	**505.2**
256	78.8	93.8	97.5	**60.0**	83.6	89.5	503.1

Table 2. Comparisons with well-annotated NCs on MS-COCO 5K. The **Best** results are marked in each column.

Methods	Image to Text			Text to Image			
	R@1	R@5	R@10	R@1	R@5	R@10	rSum
SCAN	44.7	75.9	86.6	33.3	63.5	75.4	379.4
IMRAM	53.7	83.2	91.0	39.7	69.1	79.8	416.5
SGRAF	58.8	84.8	92.1	41.6	70.9	81.5	429.7
CHAN	60.2	85.9	92.4	41.7	71.5	81.7	433.4
HREM	60.6	86.4	92.5	41.3	71.9	82.4	435.1
NCR	58.2	84.2	91.5	41.7	71.0	81.3	427.9
DECL	59.2	84.5	91.5	41.7	70.6	81.1	428.6
BiCro	59.0	84.4	91.7	42.4	71.2	81.7	430.4
CRCL	61.3	85.8	**92.7**	43.5	72.6	82.7	438.6
CREAM	58.8	85.0	92.1	42.5	71.7	81.9	432.0
UGNCL	60.8	85.4	92.3	43.4	72.1	82.3	436.2
L2RM	60.1	86.1	92.6	43.8	72.1	82.3	437.0
ReCon	**61.6**	**86.7**	**92.7**	**44.4**	**73.1**	**83.1**	**441.6**

3. More Experiments

3.1. More Results under Simulated NCs

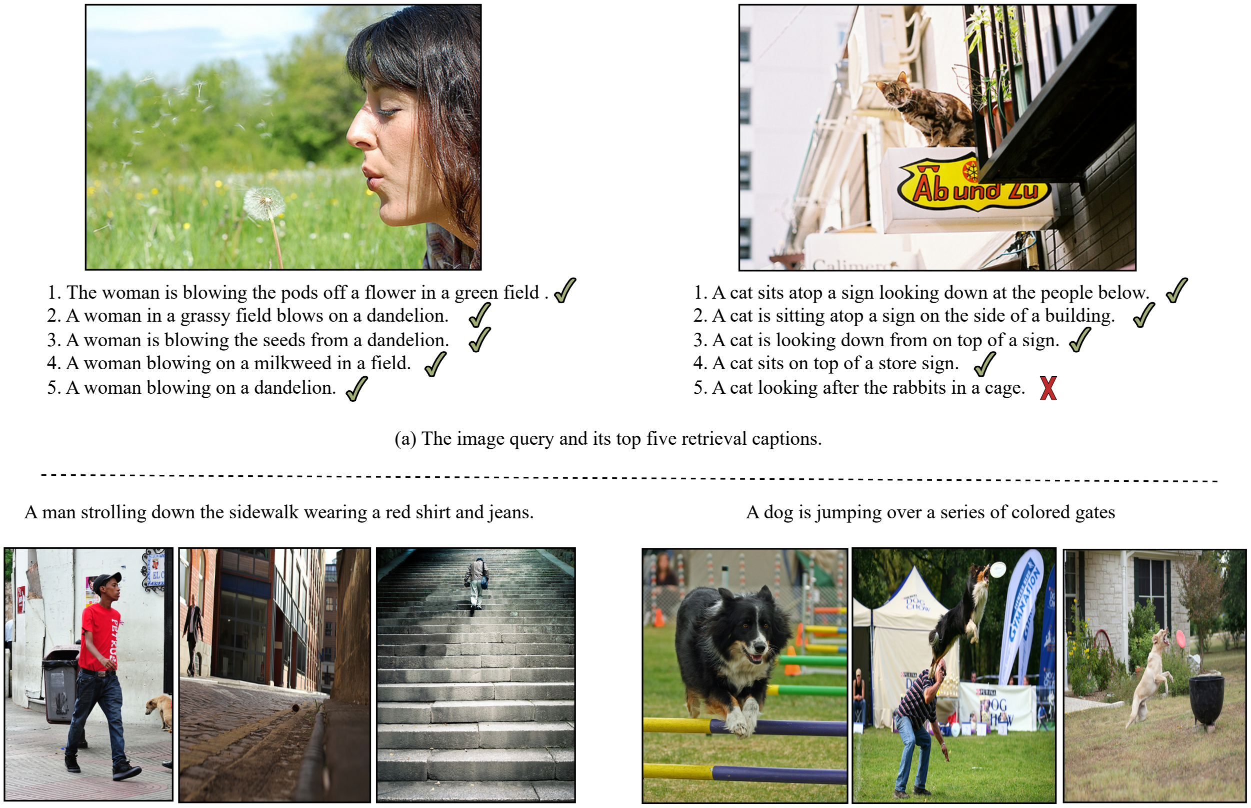
To fully demonstrate the superiority and generalization of the proposed ReCon, we provide additional comparison re-

Table 3. The settings of some key parameters for training on three datasets.

Noise Ratio	Datasets	Training parameters				Model parameters						
		Warmup Epochs	Epochs	lr_update	batch size	τ	ω_1	ω_2	α	β	γ	ξ
0%	MS-COCO	5	20	10	128	0.1	0.5	0.5	0.1	0.6	0.2	5
	CC152K	5	40	20	128	0.1	0.5	0.5	0.1	0.6	0.2	5
20%,40%,60%	Flickr30K	5	40	20	128	0.1	0.5	0.5	0.1	0.6	0.2	5
	MS-COCO	5	20	10	128	0.1	0.5	0.5	0.1	0.6	0.2	5

Table 4. Cross-modal retrieval performance comparison on Flickr30K and MS-COCO 1K. The highest scores are marked in **bold**.

Noise Ratio	Methods	Flickr30K							MS-COCO 1K						
		Image to Text			Text to Image				Image to Text			Text to Image			
		R@1	R@5	R@10	R@1	R@5	R@10	rSum	R@1	R@5	R@10	R@1	R@5	R@10	rSum
20%	IMRAM	59.1	85.4	91.9	44.5	71.4	79.4	431.7	69.9	93.6	97.4	55.9	84.4	89.6	490.8
	SAF	51.8	79.5	88.3	38.1	66.8	76.6	401.1	41.0	78.4	89.4	38.2	74.0	85.5	406.5
	SGR	61.2	84.3	91.5	44.5	72.1	80.2	433.8	49.1	83.8	92.7	42.5	77.7	88.2	434.0
	DECL-SAF	73.1	93.0	96.2	57.0	82.0	88.4	489.7	77.2	95.9	98.4	61.6	89.0	95.3	517.4
	DECL-SGR	75.4	93.2	96.2	56.8	81.7	88.4	491.7	76.9	95.3	98.2	61.3	89.0	95.1	515.8
	BiCro-SAF	77.0	93.3	97.5	57.2	82.3	89.1	496.4	74.5	95.0	98.2	60.7	89.0	95.0	512.4
	BiCro-SGR	76.5	93.1	97.4	58.1	82.4	88.5	496.0	75.7	95.1	98.1	60.5	88.6	94.7	512.7
	RCL-SAF	72.0	91.7	95.8	53.6	79.9	86.7	479.7	77.1	95.5	98.2	61.0	88.8	94.6	515.2
	RCL-SGR	74.2	91.8	96.9	55.6	81.2	87.5	487.2	77.0	95.5	98.1	61.3	88.8	94.8	515.5
	L2RM-SAF	73.7	94.3	**97.7**	56.8	81.8	88.1	492.4	77.9	96.0	98.3	62.1	89.2	94.9	518.4
	L2RM-SGR	76.5	93.7	97.3	55.5	81.5	88.0	492.5	78.4	95.7	98.3	62.1	89.1	94.9	518.5
	ReCon-SAF	77.8	**94.9**	96.9	**59.3**	83.7	**90.2**	502.8	78.5	**96.1**	98.6	63.4	90.3	**95.7**	522.6
	ReCon-SGR	**79.1**	94.7	97.3	59.1	**83.9**	**90.2**	**504.3**	**79.8**	96.0	**98.7**	**63.9**	**90.4**	**95.7**	**524.6**
40%	IMRAM	44.9	73.2	82.6	31.6	56.3	65.6	354.2	51.8	82.4	90.9	38.4	70.3	78.9	412.7
	SAF	7.4	19.6	26.7	4.4	12.0	17.0	87.1	13.5	43.8	48.2	16.0	39.0	50.8	211.3
	SGR	4.1	16.6	24.1	4.1	13.2	19.7	81.8	1.3	3.7	6.3	0.5	2.5	4.1	18.4
	DECL-SAF	72.2	91.4	95.6	54.0	79.4	86.4	479.0	75.8	95.0	98.1	60.3	88.7	94.9	512.8
	DECL-SGR	72.4	92.2	96.5	54.5	80.1	87.1	482.8	75.9	95.3	98.2	60.2	88.3	94.8	512.7
	BiCro-SAF	72.5	91.7	95.3	53.6	79.0	86.4	478.5	75.2	95.0	97.9	59.4	87.9	94.3	509.7
	BiCro-SGR	72.8	91.5	94.6	54.7	79.0	86.3	478.9	74.6	94.8	97.7	59.4	87.5	94.0	508.0
	RCL-SAF	68.8	89.8	95.0	51.0	76.7	84.8	466.1	74.8	94.8	97.8	59.0	87.1	93.9	507.4
	RCL-SGR	71.3	91.1	95.3	51.4	78.0	85.2	472.3	73.9	94.9	97.9	59.0	87.4	93.9	507.0
	L2RM-SAF	72.1	92.1	96.1	52.7	78.8	85.9	477.7	74.4	94.7	98.3	59.2	87.9	94.4	508.9
	L2RM-SGR	73.1	92.4	96.3	52.3	79.4	86.3	479.8	75.2	94.8	98.1	59.4	87.8	94.1	509.4
	ReCon-SAF	**76.9**	**94.2**	**97.4**	**57.3**	82.4	88.6	**496.8**	78.0	95.8	98.4	62.4	89.7	95.4	519.7
	ReCon-SGR	76.5	92.8	97.1	**57.3**	**82.5**	**89.0**	495.1	**78.2**	**96.1**	**98.7**	**62.5**	**89.9**	**95.5**	**520.8**
60%	IMRAM	16.4	38.2	50.9	7.5	19.2	25.3	157.5	18.2	51.6	68.0	17.9	43.6	54.6	253.9
	SAF	0.1	1.5	2.8	0.4	1.2	2.3	8.3	0.1	0.5	0.7	0.8	3.5	6.3	11.9
	SGR	1.5	6.6	9.6	0.3	2.3	4.2	24.5	0.1	0.6	1.0	0.1	0.5	1.1	3.4
	DECL-SAF	66.4	88.1	93.6	49.8	76.1	84.4	458.4	71.1	93.6	97.3	57.9	86.8	93.8	500.5
	DECL-SGR	68.5	89.9	94.8	50.3	76.7	84.1	464.3	73.2	94.4	97.9	58.2	86.8	93.9	504.4
	BiCro-SAF	67.1	88.3	93.8	48.8	75.2	83.8	457.0	72.5	94.3	97.9	57.7	86.9	93.8	503.1
	BiCro-SGR	68.5	89.1	93.1	48.2	74.7	82.7	456.3	73.4	94.0	97.5	58.0	86.8	93.6	503.3
	RCL-SAF	63.9	84.8	91.7	43.0	71.2	79.4	434.0	70.1	93.1	96.8	54.5	84.4	91.9	490.8
	RCL-SGR	62.3	86.3	92.9	45.1	71.3	80.2	438.1	71.4	93.2	97.1	55.4	84.7	92.3	494.1
	L2RM-SAF	66.1	88.8	93.8	47.8	74.2	82.2	452.9	71.2	93.4	97.5	56.5	85.9	93.0	497.5
	L2RM-SGR	65.1	87.8	93.6	47.0	73.5	81.5	448.5	72.7	93.9	97.5	56.9	86.2	93.3	500.5
	ReCon-SAF	71.8	91.4	96.3	53.2	79.3	86.4	478.4	**75.5**	95.1	98.2	**60.3**	88.2	94.7	**512.0**
	ReCon-SGR	**71.9**	**92.5**	**96.5**	**53.5**	**79.9**	**86.7**	**480.9**	75.1	**95.2**	**98.3**	**60.3**	**88.3**	**94.8**	511.9



1. The woman is blowing the pods off a flower in a green field . ✓
2. A woman in a grassy field blows on a dandelion. ✓
3. A woman is blowing the seeds from a dandelion. ✓
4. A woman blowing on a milkweed in a field. ✓
5. A woman blowing on a dandelion. ✓

1. A cat sits atop a sign looking down at the people below. ✓
2. A cat is sitting atop a sign on the side of a building. ✓
3. A cat is looking down from on top of a sign. ✓
4. A cat sits on top of a store sign. ✓
5. A cat looking after the rabbits in a cage. ✗

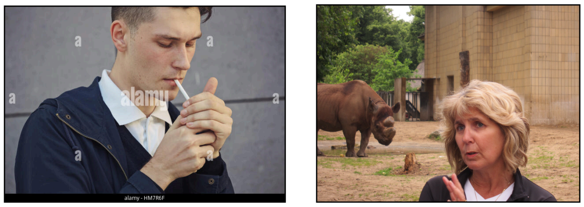
(a) The image query and its top five retrieval captions.

A man strolling down the sidewalk wearing a red shirt and jeans.

A dog is jumping over a series of colored gates

(b) The text query and its top three retrieval images.

Figure 1. Some retrieved examples of ReCon on Flickr30K under 40% noise.



hipster man walking in the streets.

making a serious point during press conference.

Figure 2. Some detected noisy examples by ReCon on CC152K.

sults under various robust frameworks, including IMRAM[1] [2], SAF, SGR[2], DECL[3] [11], BiCro [14], RCL [7], and L2RM [6]. From the results shown in Table.4, both ReCon-SAF and ReCon-SGR demonstrate substantial improvements, significantly surpassing the robustness and effectiveness of existing state-of-the-art methods. These results further confirm the effectiveness and superiority of ReCon across diverse robust frameworks, highlighting its potential as a strong solution for robust cross-modal retrieval.

3.2. Results under Well-annotated MS-COCO 5K

In this section, we supplement two state-of-the-art image-text matching methods, including CHAN (CVPR'23) [10] and HREM (CVPR'23) [5], as well as two robust frameworks, i.e., CREAM (TIP'24) [9] and UGNCL (SIGIR'24) [15], to comprehensively and faithfully evaluate the robustness and effectiveness of our ReCon under MS-COCO 5K. As presented in Table.2, ReCon consistently achieves best performance across all metrics, demonstrating its strong ability to handle well-annotated scenarios and its superiority over existing SOTA methods. This result proves the capability of ReCon to facilitate reliable cross-modal retrieval.

3.3. Analysis of Batch Size

In this section, we explore the model performance under different batch sizes during training. As shown in Table. 1, performance improves with increasing batch size but eventually reaches a plateau. This trend indicates that while a larger batch size enhances model learning by capturing more diverse representations, surpassing a certain thresh-

[1] https://github.com/HuiChen24/IMRAM

[2] https://github.com/Paranioar/SGRAF

[3] https://github.com/QinYang79/DECL

old leads to diminishing performance gains. These results highlight the importance of selecting an optimal batch size to strike a balance between computational efficiency and retrieval effectiveness.

3.4. Visualization of Retrieval Results

We showcase some visualization examples from Flickr30K in Fig.1 to demonstrate the effectiveness and reliability of our ReCon. Moreover, we also present some detected noisy examples by ReCon from CC152K dataset in Fig.2. These examples illustrate that ReCon successfully retrieves the most relevant items and validate its capability to effectively handle noisy correspondence.

References

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[2] Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, Ji Liu, and Jungong Han. Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12652–12660, 2020.

[3] Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. Learning the best pooling strategy for visual semantic embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15789–15798, 2021.

[4] Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. Similarity reasoning and filtration for image-text matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1218–1226, 2021.

[5] Zheren Fu, Zhendong Mao, Yan Song, and Yongdong Zhang. Learning semantic relationship among instances for image-text matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15159–15168, 2023.

[6] Haochen Han, Qinghua Zheng, Guang Dai, Minnan Luo, and Jingdong Wang. Learning to rematch mismatched pairs for robust cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26679–26688, 2024.

[7] Peng Hu, Zhenyu Huang, Dezhong Peng, Xu Wang, and Xi Peng. Cross-modal retrieval with partially mismatched pairs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):9595–9610, 2023.

[8] Zhenyu Huang, Guocheng Niu, Xiao Liu, Wenbiao Ding, Xinyan Xiao, Hua Wu, and Xi Peng. Learning with noisy correspondence for cross-modal matching. In *Proceedings of Advances in Neural Information Processing Systems*, pages 29406–29419, 2021.

[9] Xinran Ma, Mouxing Yang, Yunfan Li, Peng Hu, Jiancheng Lv, and Xi Peng. Cross-modal retrieval with noisy correspondence via consistency refining and mining. *IEEE Transactions on Image Processing*, 33:2587–2598, 2024.

[10] Zhengxin Pan, Fangyu Wu, and Bailing Zhang. Fine-grained image-text matching by cross-modal hard aligning network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19275–19284, 2023.

[11] Yang Qin, Dezhong Peng, Xi Peng, Xu Wang, and Peng Hu. Deep evidential learning with noisy correspondence for cross-modal retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4948–4956, 2022.

[12] Yang Qin, Yuan Sun, Dezhong Peng, Joey Tianyi Zhou, Xi Peng, and Peng Hu. Cross-modal active complementary learning with self-refining correspondence. *Proceedings of International Conference on Neural Information Processing Systems*, pages 24829–24840, 2024.

[13] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017.

[14] Shuo Yang, Zhao Pan Xu, Kai Wang, You Yang, Hongxun Yao, Tongliang Liu, and Min Xu. Bicro: Noisy correspondence rectification for multi-modality data via bi-directional cross-modal similarity consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19883–19892, 2023.

[15] Quanxing Zha, Xin Liu, Yiu-ming Cheung, Xing Xu, Nannan Wang, and Jianjia Cao. Ugncl: Uncertainty-guided noisy correspondence learning for efficient cross-modal matching. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 852–861. Association for Computing Machinery, 2024.