Supplemental Material for BiCro: Noisy Correspondence Rectification for Multi-modality Data via Bi-directional Cross-modal Similarity Consistency

1. Introduction

In this supplemental material, we provide additional information. Specifically, we give a further explanation of the implementation method and report the parameter settings for each dataset in the experiments in Section 2. In Section 3.1, we report experimental results on clean Flickr30K and MS-COCO without injecting manual noise. In Section 3.2, we conduct more experiments for BiCro based on SAF and SGR. In Section 4, we visually show some estimated soft correspondence labels and qualitative results of text and image retrieval on CC152K.

2. Implementation and Training Details

2.1. Implementation Details

Our BiCro is built of top of SGRAF [2], for details of the specific network structure please refer to [2] and [3].

2.2. Parameter Settings

We give the training parameters of BiCro in three datasets in Table 1. Besides, we select the checkpoint with the best performance on the validation set for testing.

3. More Experiments

We conduct more experiments to prove the effectiveness of our BiCro. Note that data pairs with estimated soft labels under a threshold (mismatch threshold β) are treated as mismatched data, and their correspondence labels are set as zero (denoted as BiCro* in Table 2, Table 3 and Table 4).

3.1. Experimental Results without Noise

We conduct comparison experiments in terms of cross-modal retrieval on two datasets to evaluate the performance of our BiCro without simulated noise. The baselines are SCAN [4], VSRN [5], IMRAM [1], SGRAF, SGR, SAF [2], NCR [3], and DECL [6] respectively. The results are shown in Table 2. BiCro achieves competitive performance. Specifically, BiCro is 8.1% and 1% higher than the best baseline in terms of sum in retrieval on Flickr30K and MS-COCO, respectively.

3.2. Experimental Results of BiCro-SAF and BiCro-SGR

Due to the space limitation of the main text, we supplement in this section the performance of BiCro-SAF and BiCro-SGR which apply BiCro to SAF and SGR, respectively. Table 3 reports the experimental results on the 1K test images of Flickr30K and over 5 folds of 1K test images of MS-COCO dataset. The result of CC152K is reported in Table 4. The baselines are DECL-SAF and DECL-SGR [6], whose performance is state-of-the-art for robust learning methods against noisy correspondence. As shown in Table 3 and Table 4, our BiCro improves the ability of SAF and SGR to resist noisy data and achieve state-of-the-art performance. Comparison with DECL-SAF [6] on Flickr30K and MS-COCO dataset, BiCro-SAF improves sum by 7.3%, 11.9%, 15.7%, 42.6%, 2.1%, 0.4%, 2.8%, and 10.0% for retrieving texts and image under different noise rates, respectively. On the other hand, BiCro-SGR improves sum by 9.4%, 10.9%, 21.1%, 8.4%, 0.2%, 1.6%, 1.6%, and 7.2% . On the real-world noise dataset, CC152K, our BiCro is 2.2% and 0.9% higher than the best baseline in terms of sum based on SAF and SGR, respectively. Moreover, the large performance gap between BiCro and BiCro* shows that the filtering of data pairs according to soft correspondence labels can further reduce the impact of data mismatch issue on performance.

4. Visualization Experiments

In this section, we first show the estimated soft labels of BiCro with the visualization results in Fig. 1. Then, We show qualitative results of BiCro for image-to-text retrieval in Fig. 2 and text-to-image retrieval in Fig. 3. We conduct experiments on the real-world noise dataset, CC152K, to show the estimation of the oft labels of our BiCro for mismatched and weakly-matched data pairs. The image-to-text retrieval results on the CC152K dataset. In each panel, the left column shows the image-text pair in the CC152K dataset and our estimated soft correspondence label. The right column shows the most similar texts retrieved by the image.

Table 1. The settings of some key parameters for training on three datasets. Warmup Epochs means the epochs for warmuping model and BiCro decays the leaning rate (lr) by 0.1 in lr_update epoch. α and β are warmup selection ratio and mismatch threshold of BiCro.

		Training parameters				Model parameters	
Noise	Dataset	Warmup Epochs	Epochs	lr_update	batch size	α	β
0%	Flickr30K	10	40	20	128	0.5	0.5
	MS-COCO	10	20	10	128	0.5	0.5
	CC-152K	10	40	20	128	0.5	0.5
20%,40%,60%	Flickr30K	10	40	20	128	0.3	0.5
	MS-COCO	10	20	10	128	0.5	0.6

Table 2. Performance comparison without simulated noisy correspondence (0% noise) on Flickr30K and MS-COCO 1K.

		Flickr30K							MS-COCO						
		Image⟶Text			Text⟶Image				Image⟶Text			Text⟶Image			
Noise	Methods	R@1	R@5	R@10	R@1	R@5	R@10	Sum	R@1	R@5	R@10	R@1	R@5	R@10	Sum
0%	SCAN	67.4	90.3	95.8	48.6	77.7	85.2	465.0	69.2	93.6	97.6	56.0	86.5	93.5	496.4
	VSRN	71.3	90.6	96.0	54.7	81.8	88.2	482.6	76.2	94.8	98.2	62.8	89.7	95.1	516.8
	IMRAM	74.1	93.0	96.6	53.9	79.4	87.2	484.2	76.7	95.6	98.5	61.7	89.1	95.0	516.6
	SAF	73.7	93.3	96.3	56.1	81.5	88.0	488.9	76.1	95.4	98.3	61.8	89.4	95.3	516.3
	SGR	75.2	93.3	96.6	56.2	81.0	86.5	488.9	78.0	95.8	98.2	61.4	89.3	95.4	518.1
	SGRAF	77.8	94.1	97.4	58.5	83.0	88.8	499.6	**79.6**	96.2	98.5	63.2	**90.7**	**96.1**	524.3
	NCR	77.3	94.0	97.5	59.6	84.4	89.9	502.7	78.7	95.8	98.5	63.3	90.4	95.8	522.5
	DECL-SGRAF	79.8	94.9	97.4	59.5	83.9	89.5	505.0	79.1	96.3	**98.7**	63.3	90.1	95.6	523.1
	BiCro-SGRAF	81.2	**96.1**	98.0	61.3	85.6	**90.9**	513.1	79.3	96.3	**98.7**	**63.8**	90.1	95.9	524.1
	BiCro-SGRAF*	**81.7**	95.3	**98.4**	**61.6**	**85.6**	90.8	**513.4**	79.1	**96.4**	98.6	**63.8**	90.4	96.0	**524.5**

Table 3. Image-Text Retrieval on Flickr30K and MS-COCO 1K.

		Flickr30K							MS-COCO						
		Image⟶Text			Text⟶Image				Image⟶Text			Text⟶Image			
Noise	Methods	R@1	R@5	R@10	R@1	R@5	R@10	Sum	R@1	R@5	R@10	R@1	R@5	R@10	Sum
0%	DECL-SAF	77.0	93.9	96.9	56.8	81.7	88.0	494.3	77.8	95.8	98.4	61.4	89.2	95.2	517.8
	DECL-SGR	77.1	93.6	96.7	57.3	82.1	88.4	495.2	76.9	95.8	**98.6**	61.6	89.4	95.2	517.5
	BiCro-SAF	76.8	94.3	97.3	58.9	84.3	89.9	501.6	77.5	**96.3**	**98.6**	62.2	89.8	95.5	519.9
	BiCro-SGR	78.8	**94.8**	97.4	59.5	84.3	89.8	504.6	77.7	95.7	98.1	61.5	89.3	95.4	517.7
	BiCro-SAF*	79.3	**94.8**	**97.7**	**60.0**	**84.5**	**90.3**	**506.6**	76.9	95.7	**98.6**	62.4	89.7	95.4	518.7
	BiCro-SGR*	**80.7**	94.3	97.6	59.8	83.8	89.7	505.8	**78.3**	95.8	98.5	**62.7**	**90.0**	**95.7**	**521.0**
20%	DECL-SAF	73.4	92.0	96.4	53.6	79.7	86.4	481.5	74.4	**95.3**	98.2	59.8	88.3	94.8	510.8
	DECL-SGR	74.5	92.9	97.1	53.6	79.5	86.8	484.4	75.6	95.1	**98.3**	59.9	88.3	94.7	511.9
	BiCro-SAF	75.9	93.7	96.8	56.7	81.7	88.7	493.4	74.0	94.9	98.2	60.1	88.8	95.2	511.2
	BiCro-SGR	76.8	**93.8**	96.5	57.8	82.3	88.2	495.3	**76.1**	95.2	98.1	60.6	88.6	94.9	**513.5**
	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	495.9	75.7	95.1	98.1	60.5	88.6	94.7	512.7
40%	DECL-SAF	70.1	90.6	94.4	49.7	76.6	84.1	465.5	73.3	94.6	**98.1**	57.9	87.2	94.1	505.2
	DECL-SGR	69.0	90.2	94.8	50.7	76.3	84.1	465.1	73.6	94.6	97.9	59.5	86.9	93.9	504.7
	BiCro-SAF	71.9	92.0	95.7	54.7	79.8	87.1	481.2	73.6	94.5	97.9	**59.9**	**88.0**	**94.5**	508.0
	BiCro-SGR	**74.0**	**92.8**	**96.4**	**55.1**	**80.6**	87.3	**486.2**	73.1	94.4	97.8	59.0	87.7	94.3	506.3
	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
60%	DECL-SAF	56.6	82.5	89.7	40.4	66.6	**76.6**	412.4	68.6	92.9	97.4	54.1	84.9	92.7	490.6
	DECL-SGR	64.5	85.8	92.6	44.0	71.6	80.6	439.1	69.7	93.4	97.5	54.5	85.2	92.6	492.9
	BiCro-SAF	65.6	87.8	93.2	**49.3**	75.6	83.5	455.0	72.0	93.7	97.6	56.9	86.6	93.8	500.6
	BiCro-SGR	64.9	87.8	93.2	46.3	73.4	81.9	447.5	72.5	93.8	97.4	56.9	86.2	93.3	500.1
	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**

References

[1] 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*

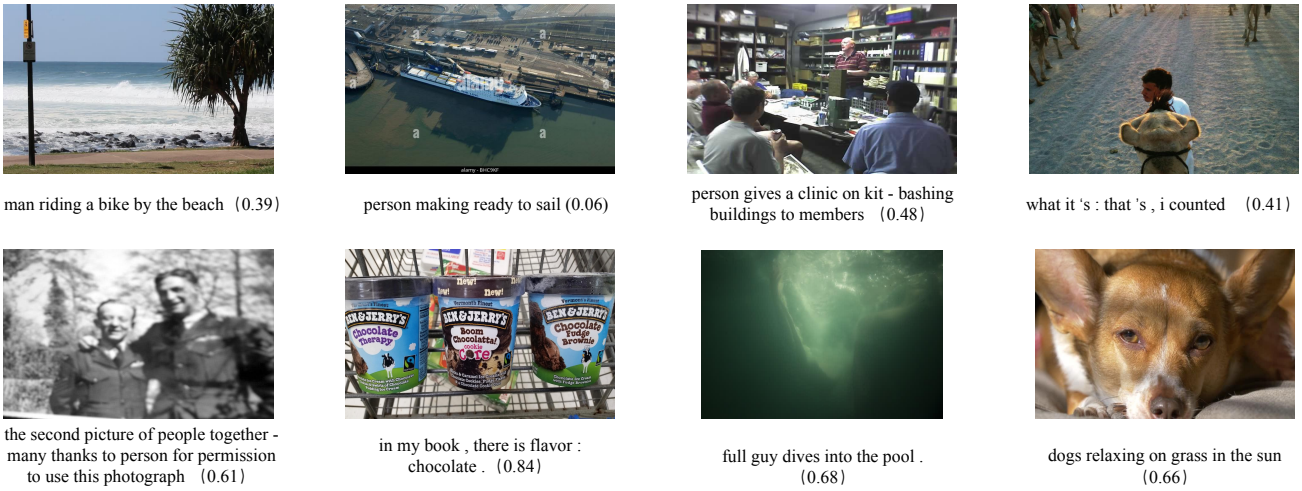

Figure 1. Visualization of our estimated soft correspondence labels on CC152K dataset.

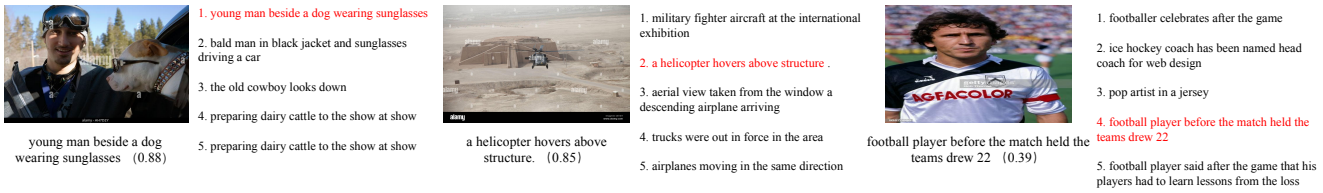

Figure 2. The image-to-text retrieval results on the CC152K dataset. In each panel, the left column shows the image-text pair in the CC152K dataset and our estimated soft correspondence label. The right column shows the most similar texts retrieved by the image.

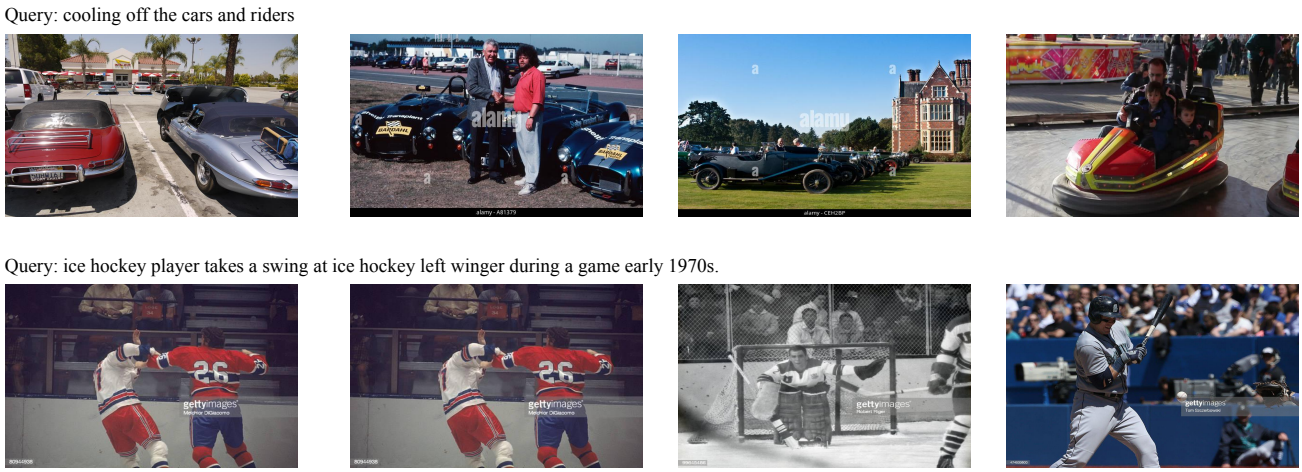

Figure 3. The text to image retrieval results on CC152K dataset.

pattern recognition, pages 12655–12663, 2020. 1

[2] Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. Similarity reasoning and filtration for image-text matching. In *AAAI*, pages 1218–1226, 2021. 1

[3] Zhenyu Huang, Guocheng Niu, Xiao Liu, Wenbiao Ding, Xinyan Xiao, Hua Wu, and Xi Peng. Learning with noisy correspondence for cross-modal matching. *Advances in Neural Information Processing Systems*, 34:29406–29419, 2021. 1

[4] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *ECCV*, 2018. 1

[5] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *ICCV*,

Table 4. Image-Text Retrieval on CC152K.

	Image⟶Text			Text⟶Image			
Methods	R@1	R@5	R@10	R@1	R@5	R@10	Sum
DECL-SAF	36.6	63.0	73.3	38.5	63.2	73.5	348.1
DECL-SGR	36.2	63.6	73.2	37.1	63.6	73.7	347.4
BiCro-SAF	38.2	64.1	71.6	38.2	64.3	73.9	350.3
BiCro-SGR	38.5	62.9	71.0	38.0	64.3	73.6	348.3
BiCro-SAF*	**39.5**	**66.0**	**74.9**	**38.7**	**65.7**	**76.4**	**361.2**
BiCro-SGR*	36.9	63.4	72.3	36.8	64.9	74.5	348.8

pages 4653–4661, 2019. 1

[6] Yang Qin, Dezhong Peng, Xi Peng, Xu Wang, and Peng Hu. Deep evidential learning with noisy correspondence for cross-modal retrieval. In *ACMMM, pages = 4948–4956, year = 2022*. 1