

PC²: Pseudo-Classification Based Pseudo-Captioning for Noisy Correspondence Learning in Cross-Modal Matching - Supplementary Materials -

Anonymous Authors

1 MORE DETAILS OF NOW

Following NCR [6], for all images, we use the detection model to extract the top 36 region proposals whose features are used for training. Considering that the image data in NoW consists of web pages, we utilize the detection model APT [3], which is specialized in Mobile User Interface (MUI), to obtain region proposals. Following the original APT paper [3], we fine-tune APT with a ResNet-50 backbone [5] pre-trained by CLIP [9] on MUI-zh dataset [3]. MUI-zh is a dataset for high-quality detection of MUI elements. It consists of screenshots collected from mobile tinyapps. The training set contains 4,769 images and 41k elements, while the validation set consists of 1,000 images and 9,000 elements across 18 categories.. For specific training details, please refer to Sec. 5.1 in [3]. Finally, we select the top 36 region proposals for each image to construct the final NoW dataset. In order to better showcase the content of NoW, we provide additional examples in Fig. 1.

2 IMPLEMENTATION DETAILS

We show the complete hyper-parameters of PC² in Tab. 1.

2.1 Component Details

Co-Dividing. Given \mathcal{D} , we compute the per-sample loss by

$$l_i = \sum_{\hat{T}_i} [\alpha - S(I_i, T_i) + S(I_i, \hat{T}_i)]_+ + \sum_{\hat{I}_i} [\alpha - S(I_i, T_i) + S(\hat{I}_i, T_i)]_+, \quad (1)$$

where \hat{T}_i and \hat{I}_i are negative text and negative image, respectively. Then, we use the two-component Gaussian Mixture Model (GMM) to fit the losses $\{l_i\}_{i=1}^B$. Taking advantage of the memorization effect of DNNs, we classify the component with a lower mean value (*i.e.*, lower loss) as the clean set, while considering the other component as the noisy set. Thus, with the GMM optimized by Expectation-Maximization algorithm, we compute the posterior probability of each pair as clean probability w_i . Please refer to Sec. 3.1 in [6] for a complete description of the co-dividing module.

Co-Teaching. For co-teaching, we train two networks, denoted as $A = \{f_A, g_A, S_A\}$ and $B = \{f_B, g_B, S_B\}$, simultaneously. These networks have the same architecture but are trained on different data sequences and initialized differently. In each epoch, either network A or network B models its per-sample loss distribution using co-dividing. This allows the dataset to be divided into clean and noisy subsets, which are then utilized for training the other network. During the test phase, we calculate the average of the similarities predicted by A and B for the retrieval evaluation.

Prediction Oscillation Based Correspondence Rectification. We attempt to gain insights into the model’s awareness of correspondence strength by measuring the level of prediction oscillation,

Table 1: Complete list of hyper-parameters in PC².

Hyper-parameter	Description	Flickr30K	MS-COCO	NoW
B	Batch size		128	
τ	Clean data threshold		0.5	
K	K -way pseudo-classifier		128	
α	Original margin		0.2	
m	Curve parameter		10	
λ^n	Loss weight of \mathcal{L}^n		1	
λ^{pse}	Loss weight of \mathcal{L}^{pse}		1	
λ^{ent}	Loss weight of \mathcal{L}^{ent}		10	
-	word embedding size		300	
-	joint embedding size		2048	

Table 2: Comparison of image-text retrieval on CC152K.

Methods	Image → Text			Text → Image			Rsum
	R@1	R@5	R@10	R@1	R@5	R@10	
SCAN [7]	30.5	55.3	65.3	26.9	53.0	64.7	295.7
VSRN [8]	32.6	61.3	70.5	32.5	59.4	70.4	326.7
IMRAM [1]	33.1	57.6	68.1	29.0	56.8	67.4	312.0
SAF [2]	31.7	59.3	68.2	31.9	59.0	67.9	318.0
SGR [2]	11.3	29.7	39.6	13.1	30.1	41.6	165.4
NCR [6]	39.5	64.5	73.5	40.3	64.6	73.2	355.6
PC ² (Ours)	39.3	66.4	75.4	39.8	66.4	76.8	364.1

Table 3: Comparison of image-text retrieval on Flickr30K. X^* means the results of X with mismatch threshold.

Noise	Methods	Image → Text			Text → Image			Rsum
		R@1	R@5	R@10	R@1	R@5	R@10	
20%	BiCro* [11]	<u>78.1</u>	<u>94.4</u>	<u>97.5</u>	<u>60.4</u>	<u>84.4</u>	<u>89.9</u>	<u>504.7</u>
	PC* (Ours)	78.4	95.2	97.0	60.5	84.6	90.3	506.0
40%	BiCro* [11]	<u>74.6</u>	<u>92.7</u>	<u>96.2</u>	<u>55.5</u>	<u>81.1</u>	<u>87.4</u>	<u>487.5</u>
	PC* (Ours)	76.3	93.8	97.4	57.6	82.3	88.6	496.0
60%	BiCro* [11]	<u>67.6</u>	90.8	<u>94.4</u>	<u>51.2</u>	<u>77.6</u>	<u>84.7</u>	<u>466.3</u>
	PC* (Ours)	70.9	<u>90.7</u>	95.0	53.0	79.1	86.2	474.9

which provides additional information for correspondence correction. The definition of prediction oscillation in Eq. (7) relies on measuring the variation in the predicted probability distribution of the same sample across different epochs. Therefore, it is natural to choose the KL-divergence, which is commonly used to measure the distance between distributions, to quantify the strength of prediction oscillation. After calculating the evaluation values for all

117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174

175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232

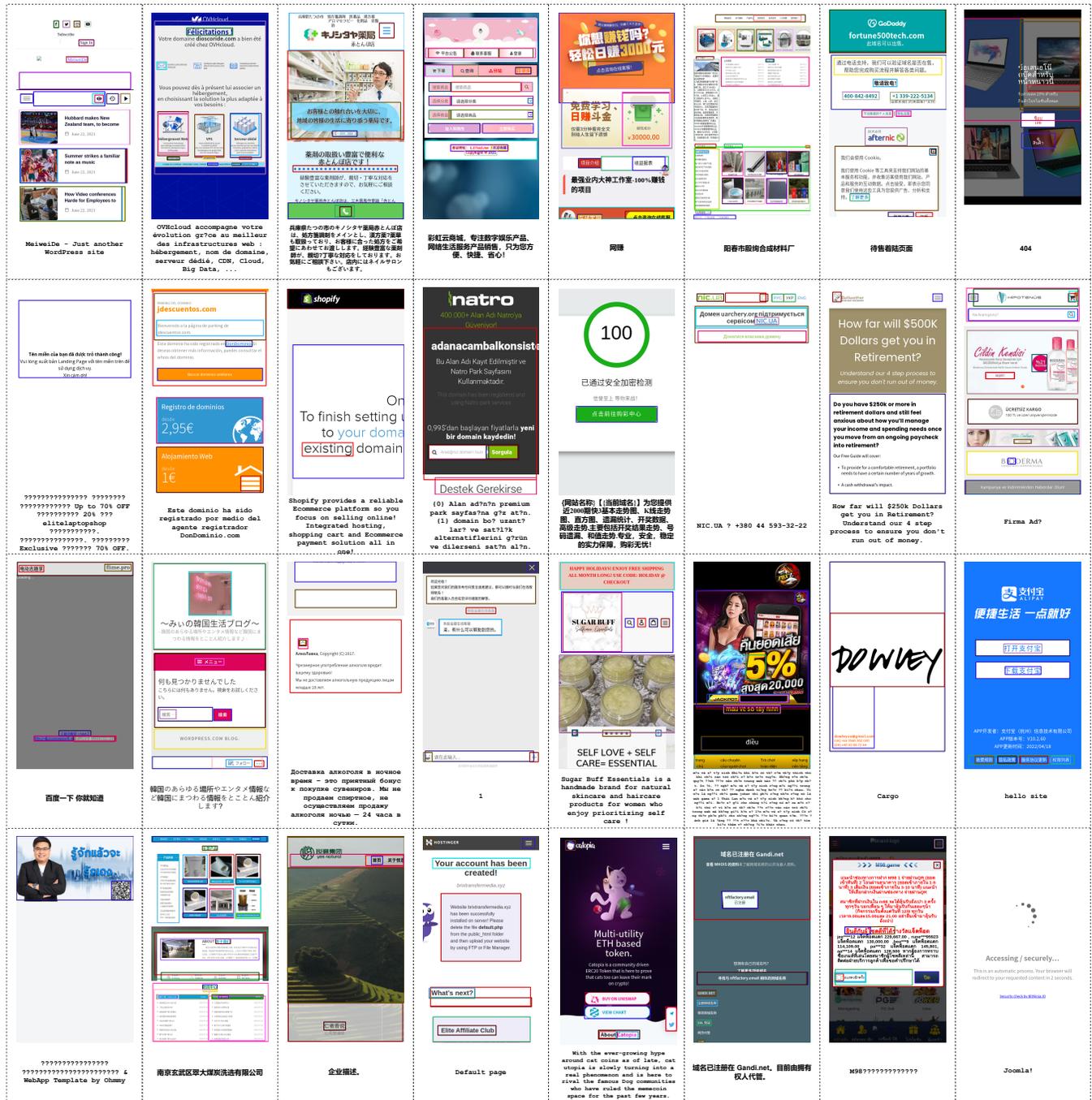


Figure 1: More examples of image-text pairs in NoW. We randomly select 36 pairs for display. The colored boxes is the region proposals provided by our APT model.

prediction oscillations $\{o_i^{(e)}\}_{i=1}^B$, we utilize them as a supplementary for the dataset splitting. Following the co-dividing described in Sec. 2.1, we fit another two-component GMM using $\{o_i^{(e)}\}$. This allows us to obtain the clean probabilities based on the prediction oscillation, which is denoted as $\{w_i^0\}_{i=1}^B$. In Eq. (8), we aim

to adjust the margin based on correspondence strength of clean data by combining the predictions from the fitted loss-GMM and prediction-oscillation-GMM. We primarily rely on the predictions from the loss-based GMM (i.e., the term w_i^c in Eq. (8)). However, if the confidence level of the correspondence based on the loss is

low, we utilize the term $(1 - w_i^c) \mathbb{1}(w_i^o \geq \tau) w_i^o$ in Eq. (8) to assist in determining the correspondence strength. Note that this term only works into effect when prediction-oscillation-based GMM indicates that the image-text pair is clean data (*i.e.*, $\mathbb{1}(w_i^o \geq \tau)$).

3 MORE EXPERIMENTS

3.1 Results on Conceptual Captions

In addition to NoW, a dataset with real noise, we also conducted experiments on another real dataset Conceptual Captions [10]. Following NCR [6], our experiments utilize a subset of Conceptual Captions, specifically CC152K [6]. Within this dataset, 150,000 images are allocated for training the model, while 1,000 images are set aside for model validation, and another 1,000 images are designated for model testing. As shown in Tab. 2, although the noise ratio in CC152K is not high (3%~20%), competitive results are still achieved by our method, *i.e.*, PC² outperforms the most popular NCL method NCR by 8.4% in the term of Rsum.

3.2 Using Mismatch Threshold

Following the idea of filtering mismatched pairs with low correspondence levels in BiCro [11] using a mismatch threshold, we directly set this threshold to exclude these pairs from training, *i.e.*, set $\lambda^n = 0$. Using a threshold of 0.5, we achieve results as shown in Tab. 3 and consistently outperform BiCro with the same mismatch threshold. Since this filtering design introduces a new hyper-parameter, and our method can achieve sufficiently good results without it, we present the results of original PC² in the main text.

3.3 Results Combined with MSCN

Given that the design of MSCN [4] requires ground-truth, which is not available in the real-world NCL setting, directly comparing it with MSCN is unfair to PC². Nevertheless, to demonstrate the superiority of our method's design, we construct PC² based on MSCN for evaluating NCL performance. Specifically, we insert the meta similarity correction network (MSCN) described in [4] (see Sec 3.2 of [4] for details) behind $S(\cdot, \cdot)$ in PC². We maintain all original training settings of MSCN and use MSCN to output the final image-text similarity in place of $S(I, T)$ in Eqs. (6) and (9). Note that the training of MSCN requires ground-truth. Thus, we use the original ground-truth data protocol from [4] (details can be found in Sec 4.2 of [4]) and list the results in Tab. 3 of the main text. There, we can observe that PC², constructed based on MSCN, can still achieve stable performance improvements in the NCL setting.

3.4 More Analysis of Results on NoW

To further showcase the capabilities of PC² on NoW, we visualize the performance of pseudo-classifier C . The predictions of C do not represent fixed semantic concepts but affects how C groups images into pseudo-classes of varying granularities. C transforms feature space for efficient learning and assisting in feature matching. We use category labels predicted by C to show the distribution of image features within each pseudo-class in Fig. 2 (does not represent classification performance). We observe that C captures increasingly discriminative patterns and gradually attempts to partition similar images into the same semantic cluster.

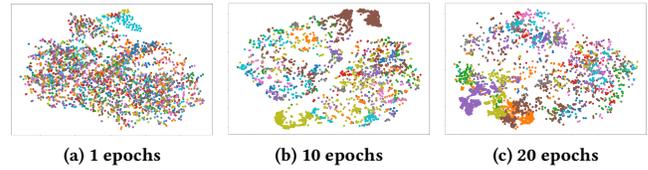


Figure 2: t-SNE visualization of image features during the training on NoW (10 pseudo-classes are randomly selected).

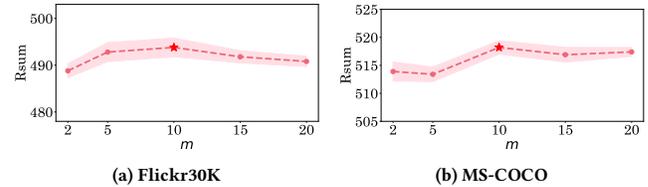


Figure 3: Ablation studies on curve parameter m on both Flickr30K and MS-COCO with 40% noise.

3.5 More Ablation Studies

Curve parameter m is an important hyper-parameter for robust image-text matching method based on triplet ranking loss [4, 6, 11]. Following previous methods [4, 6, 11], we set $m = 10$ for PC². As shown in 3, we indeed verify the suitability of this setting for PC².

REFERENCES

- [1] Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, Ji Liu, and Jungong Han. 2020. Imran: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [2] Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. 2021. Similarity reasoning and filtration for image-text matching. In *AAAI Conference on Artificial Intelligence*.
- [3] Zhangxuan Gu, Zhuoer Xu, Haoxing Chen, Jun Lan, Changhua Meng, and Weiqiang Wang. 2023. Mobile User Interface Element Detection Via Adaptively Prompt Tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [4] Haochen Han, Kaiyao Miao, Qinghua Zheng, and Minnan Luo. 2023. Noisy Correspondence Learning with Meta Similarity Correction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [5] Kaiping He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *IEEE/CVF International Conference on Computer Vision*.
- [6] Zhenyu Huang, Guocheng Niu, Xiao Liu, Wenbiao Ding, Xinyan Xiao, Hua Wu, and Xi Peng. 2021. Learning with noisy correspondence for cross-modal matching. *Advances in Neural Information Processing Systems* (2021).
- [7] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *European Conference on Computer Vision*.
- [8] Kungpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. 2019. Visual semantic reasoning for image-text matching. In *IEEE/CVF International Conference on Computer Vision*.
- [9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.
- [10] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Annual Meeting of the Association for Computational Linguistics*.
- [11] Shuo Yang, Zhaopan Xu, Kai Wang, Yang You, Hongxun Yao, Tongliang Liu, and Min Xu. 2023. BiCro: Noisy Correspondence Rectification for Multi-modality Data via Bi-directional Cross-modal Similarity Consistency. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.