Robust Noisy Correspondence Learning with Equivariant Similarity Consistency

Supplementary Material

1. Hyperparameters Analysis

In this section, we add some comparative experiments to analyze our choice of hyperparameters. Following the setting of [2], the division margin α_1 and matching margin α_2 are chosen from $\{0, 0.04, 0.1\}$ and the weighted factor β is selected from $\{0.2, 0.5, 1.0\}$. All experiments are conducted on Flickr30K with a noise ratio of 40%.

α_1	Image $\longrightarrow$ Text			Text $\longrightarrow$ Image			
	R@1	R@5	R@10	R@1	R@5	R@10	Sum
0	**76.1**	**93.1**	**96.4**	**56.0**	**80.8**	**87.2**	**489.5**
0.04	**76.1**	91.6	96.1	54.3	78.6	86.2	482.9
0.1	72.2	91.8	96.1	54.2	79.0	85.8	479.1

Table 1. Recall rates with different division margin α_1 in l_{ESC}.

β	Image $\longrightarrow$ Text			Text $\longrightarrow$ Image			
	R@1	R@5	R@10	R@1	R@5	R@10	Sum
0	72.5	91.7	95.3	53.6	79.0	86.4	478.5
0.2	73.0	92.3	95.8	55.4	**81.0**	**87.4**	484.9
0.5	**76.1**	**93.1**	**96.4**	**56.0**	80.8	87.2	**489.5**
1	74.4	92.5	95.5	55.6	80.0	86.7	484.7

Table 2. Recall rates with different weighted factor β in l_{div}.

The choice of division margin α_1. In Tab. 1, the matching model exhibits optimal retrieval performance when the division margin α_1 is equal to 0. In the three experiments above, we evaluate the impact of hyperparameter α_1 on performance by setting β to 0.5 and α_2 to 0. As the division margin α_1 increases, the consistency of equivariant similarity is no longer strictly constrained, resulting in an increased probability of misclassification within the α_1 range and thus a decrease in retrieval performance.

The choice of weighted factor β. The different strengths of regularization have a significant impact on the matching model's ability to filter out noisy samples effectively. Applying a "coarse-to-fine" strategy, we combine the triplet loss with our ESC regularization to divide the training data. As shown in Tab. 2, we conduct experiments separately with β values of 0, 0.2, 0.5, and 1, while simultaneously setting the division margin α_1 and the matching margin α_2 to 0. With the increases of β, our division ESC improves the accuracy compared to not using this regularization ($\beta = 0$). In addition, further increasing the β leads to a decline in performance since the strength of regularization becomes too large, resulting in the generalization ability drop. When β is set to 0.5, the performance of most recall rates is the highest, where R@5 and R@10 in text-to-image retrieval are a bit lower than when β is set to 0.2. However, given the overall retrieval performance, we still apply $\beta = 0.5$ in all experiments.

α_2	Image $\longrightarrow$ Text			Text $\longrightarrow$ Image			
	R@1	R@5	R@10	R@1	R@5	R@10	Sum
0	**76.1**	**93.1**	**96.4**	**56.0**	**80.8**	87.2	**489.5**
0.04	73.1	92.6	96.1	55.6	80.6	**87.4**	485.4
0.1	73.0	91.9	96.0	54.9	80.3	86.9	483.0

Table 3. Recall rates with different matching margin α_2 in $\mathcal{L}_{ESC}$.

The choice of matching margin α_2. Similar to division margin α_1, the matching margin α_2 also represents the zero-loss threshold in the regularization term, where no loss is incurred if the difference in cross-instance similarity between two samples is less than this threshold. Observing from Tab. 3, the retrieval accuracy is the highest when α_2 is set to 0.

2. Training Pipeline

Our method is trained in a co-teaching manner [1]. The detailed training pipeline is shown in the Fig. 1 and illustrated in the Algorithm 1.

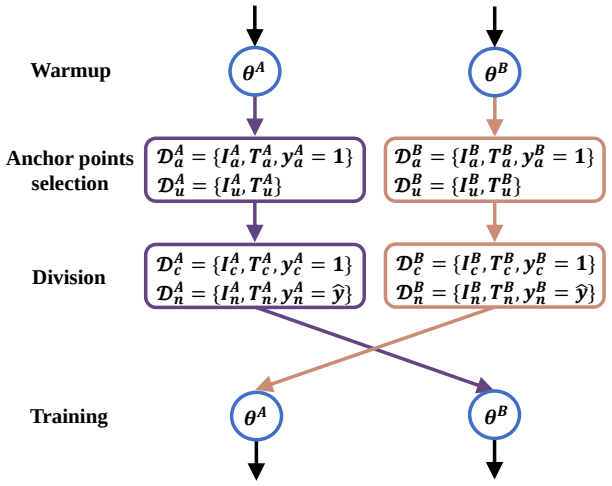

Figure 1. Training pipeline of ESC.

Algorithm 1 The training pipeline of our ESC method

Input: Given training data $\mathcal{D}$, matching models $\theta^A = \{f^A, g^A, S^A\}$ and $\theta^B = \{f^B, g^B, S^B\}$

Warmup the models (θ^A, θ^B) using l_{hard} in Eq. (1)
for i = 1 : num_epochs **do**
  Compute the division triplet loss l_{hard} using Eq. (1)
  Select anchor points in $\mathcal{D}$ and calculate the division regularization l_{ESC} using Eq. (8)
  Combine l_{hard} and l_{ESC} to acquire total division loss l_{div} using Eq. (9) and normalize the l_{div} as l_i
  $\mathcal{P}^A = \{p_i^A | p_i^A = p(k=0|l_i)\}_{i=1}^N$ $\longleftarrow$BMM$(\mathcal{D}, B)$
  $\mathcal{P}^B = \{p_i^B | p_i^B = p(k=0|l_i)\}_{i=1}^N$ $\longleftarrow$BMM$(\mathcal{D}, A)$
  for $r \in \{A, B\}$ **do**
    for k = num_steps **do**
      Sample a mini-batch $\mathcal{B}^k$ from $\mathcal{D}^r$ for θ^A and θ^B, respectively
      Select the anchor point (I_a, T_a) in this mini-batch $\mathcal{B}^k$ and obtain the undivided mini-batch $\mathcal{B}_u^k = \mathcal{B}^k / (I_a, T_a)$
      $\mathcal{B}_c^k = (I_a, T_a) \cup \{(I_i, T_i, y_i) | p_i^k > \delta, \forall (I_i, T_i) \in \mathcal{B}_u^k\}$
      $\mathcal{B}_n^k = \{(I_i, T_i, y_i) | p_i^r \leq \delta, \forall (I_i, T_i) \in \mathcal{B}_u^k\}$
      Refine the labels $y_i = \hat{y}_i$ of $\{\mathcal{B}_c^k, \mathcal{B}_n^k\}$ by [3]
      Train the network θ^r on $\{\hat{\mathcal{B}}_c^k, \hat{\mathcal{B}}_n^k\}$ by optimizing $\mathcal{L}_{soft}$ in Eq. (13) and $\mathcal{L}_{ESC}$ in Eq. (14)

Output: Matching models (θ^A, θ^B)

3. Retrieval Results

We show some retrieval results of ESC for image-to-text retrieval in Fig. 2, and text-to-image retrieval in Fig. 3.

Top1: A Boston Terrier is running on lush green grass in front of a white fence.

Top2: A black and white dog is running in a grassy garden surrounded by a white fence .

Top1: Five people wearing winter clothing , helmets , and ski goggles stand outside in the snow .

Top2: Five people wearing winter jackets and helmets stand in the snow , with snowmobiles in the background .

Top1: A baby boy in a blue and white striped shirt is sitting on his mother 's shoulders .

Top2: Asian looking lady holding a baby while sitting and looking at it .

Figure 2. Image-to-text retrieval.

Text: A photographer takes a picture of a group of one girl in a pink dress and 10 boys in suits and hats .

Ground Truth

Top1

Text: A young female student performing a downward kick to break a board held by her Karate instructor .

Ground Truth

Top1

Figure 3. Text-to-image retrieval.

References

[1] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *NeurIPS*, 31, 2018. 1

[2] Tan Wang, Kevin Lin, Linjie Li, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Equivariant similarity for vision-language foundation models. In *ICCV*, pages 11998–12008, 2023. 1

[3] Shuo Yang, Zhaopan Xu, Kai Wang, Yang You, Hongxun Yao, Tongliang Liu, and Min Xu. Bicro: Noisy correspondence rectification for multi-modality data via bi-directional cross-modal similarity consistency. In *CVPR*, pages 19883–19892, 2023. 2