Supplementary material for

A Self-Supervised Descriptor for Image Copy Detection

We provide more details about the ablations (Appendix A) and Copydays results (Appendix B). We also report a few additional details about the embedding distribution (Appendix C) and implementation details (Appendix D). The last appendix F shows additional example matches.

A. Additional ablations

Table 6 shows how copy detection accuracy is affected by several hyper-parameters.

Descriptor dimensionality. The descriptor dimension is a tradeoff between accuracy and the efficiency of the retrieval step. When constraining the descriptor to 256 dimensions for retrieval, we see highest accuracy for descriptors trained at that size.

Batch size. The training objective learns to match pairs within the global batch (across all GPUs). A larger batch size makes the training task more challenging, improving the final accuracy. Large batch sizes require training with more machines, and incur synchronization overhead due in part to synchronized batch normalization.

Training schedule. We compare accuracy as we vary the number of training epochs, and find no benefit to longer training schedules.

Variance between initializations. We train using the same setting, initializing the model with five random seeds, and find a standard deviation of 0.2% μAP and 0.1% μAP_{SN}.

Similarity normalization settings. We show score normalized accuracy given several similarity normalization settings in Table 7. Several score normalization settings work similarly well. When using a single neighbor to normalize similarity, using the 2nd nearest neighbor works best ($n = 2$). When using an average similarity across multiple neighbors, averaging the first 2, 3 or 4 neighbors work similarly well. We find that $\beta = 1$ is a good normalization weight. Our similarity normalized results use $n = 1$, $n_{end} = 3$, $\beta = 1$, a setting that we found to work well across many descriptors.

$\beta = 1, n = n_{\text{end}}$		$\beta = 1, n = 1$		$n = 1, n_{\text{end}} = 3$	
n	μAP	n_{end}	μAP	β	μAP
1	69.5	1	69.5	0.50	68.4
2	**71.1**	2	71.0	0.75	70.4
3	70.8	3	**71.1**	1.00	**71.1**
4	70.3	4	**71.1**	1.25	**71.1**
5	69.7	5	71.0	1.50	70.6

Table 7. DISC2021 μAP with different score normalization settings for a SSCD trained on DISC2021 with advanced augmentations.

Trunk and projected features. We compare SSCD trunk and projected features in Table 8. Using the linear projection at inference time improves accuracy, despite a significantly more compact code.

descriptor	dims	μAP	μAP_{SN}
trunk	2048	57.2	71.9
projected	512	**61.5**	**72.5**

Table 8. DISC2021 accuracy of SSCD trunk and projected trained on DISC2021 with advanced + mixup augmentations.

B. Full Copydays results

We provide additional Copydays results in Table 9, evaluating SSCD and $\text{SSCD}_{\text{large}}$ using preprocessing settings from prior published results. In each case, we evaluate our

batch size	μAP	μAP_{SN}
2048	54.4	67.7
4096	56.6	69.2
8192	58.2	70.0
16384	**59.4**	**70.2**

epochs	μAP	μAP_{SN}
25	54.4	67.4
50	56.2	68.9
100	**56.6**	**69.2**
200	56.3	68.9
400	55.7	68.1

dimensions	μAP	μAP_{SN}	μAP_{SN} 256d
128	49.4	59.4	59.4
256	53.9	65.6	**65.6**
512	56.6	69.2	64.0
1024	**57.3**	**70.9**	62.8
2048	56.8	70.8	62.9

Table 6. Impact of three training parameters on the accuracy: batch size, number of epochs and dimensionality. We report μAP performance on DISC21 for SSCD including advanced augmentations and $\lambda = 15$, with and without score normalization. For the dimensionality experiment we additionally report the accuracy after reduction to 256 dimensions.

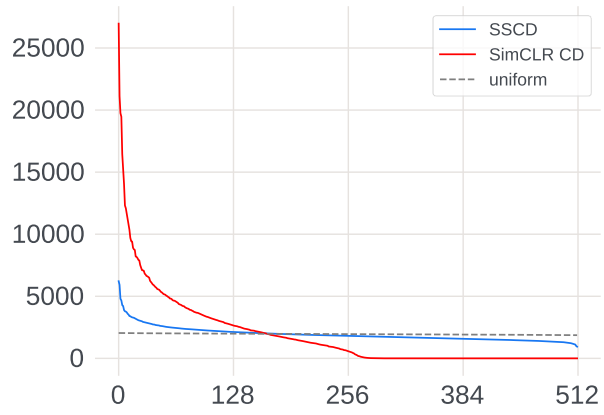

Figure 5. Descriptor principal values on the DISC2021 reference set: SSCD ($\lambda = 30$) and SimCLR$_{CD}$ ($\lambda = 0$), compared to a reference uniform distribution.

method with no tuning, *e.g.* we don't adjust the GeM p as proposed in [7].

model	trunk	dims	size	mAP	μAP
Multigrain [7]	ResNet50	1500	long 800	82.3	77.3
DINO [9]	ViT-B/16	1536	224^2	82.8	92.3
DINO [9]	ViT-B/8	1536	320^2	86.1	88.4
SSCD	ResNet50	512	224^2	84.9	98.3
SSCD	ResNet50	512	320^2	87.4	98.3
SSCD	ResNet50	512	short 288	86.6	98.1
SSCD	ResNet50	512	long 800	90.0	93.9
SSCD$_{\text{large}}$	ResNeXt101	1024	224^2	87.3	98.6
SSCD$_{\text{large}}$	ResNeXt101	1024	320^2	90.6	98.6
SSCD$_{\text{large}}$	ResNeXt101	1024	short 288	91.8	**98.7**
SSCD$_{\text{large}}$	ResNeXt101	1024	long 800	**93.6**	97.1

Table 9. Full Copydays (CD10K) results: accuracy measured in mAP on the "strong" subset, and μAP on the full dataset.

We note that at 224^2 inference size, ResNet50 has approximately $4\times$ the throughput as ResNeXt101 or ViT-B/16, and $20\times$ that of ViT-B/8. [9]

C. Embedding distribution

We plot principal values for SSCD ($\lambda = 30$) compared to SimCLR$_{CD}$ ($\lambda = 0$), and a uniform distribution in Figure 5. We see that the $\lambda = 0$ model fails to make full use of the descriptor space, as observed in [29, 60]. With entropy regularization, all components have similar energy, spanning less than an order of magnitude (the maximum is $6.6\times$ the minimum).

D. Implementation details

Mixup and Cutmix. Mixup and Cutmix augmentations both combine content from two source images. The amount of content used from each image is determined by a mixing parameter γ, sampled from a β distribution: $\gamma \sim \beta(\alpha, \alpha)$. We set $\alpha = 2$ to reduce the prevalence of "trivial" mixed images that draw nearly all content from one of the inputs.

DINO baseline details. We follow the copy detection method presented in [9] for the DINO baseline. We use the concatenation of the CLS token and GeM pooled ($p = 4$) patch token features as the descriptor.

Our DINO DISC evaluation uses the ViT-B/16 trunk. We resize inputs to 224×224 without center cropping. This outperformed other preprocessing for this model, including our default aspect-ratio preserving resize, and resizing inputs to a larger fixed size (288×288). We suspect that ViT models may be less adaptable to rectangular inputs than fully convolutional networks.

E. Visualizing matches

To view which parts of an image A match strongly to another image B, we experiment by keeping the activation map on A at full resolution by removing the GeM pooling operation. This results into one descriptor per activation map pixel, that can be compared with a global SSCD descriptor. We can thus build a spatial heatmap with the strongest activations. Figure 6 shows image pairs and the corresponding heatmaps. The areas on the left image that match with the image on the right are clearly identified.

F. Retrieved matches

We compare the first result retrieved by SSCD and SimCLR on the DISC2021 dataset. Both models are trained on ImageNet and evaluated with whitening. We use trunk features for SimCLR, which are more accurate for this model. We do not use score normalization, since it has no effect on top-1 accuracy.

SSCD	SimCLR	queries
✓	✓	38.9 %
✓	✗	39.0 %
✗	✓	0.3 %
✗	✗	21.8 %

Table 10. Percentage of DISC2021 query first result accuracy by model for SSCD and SimCLR trained on ImageNet.

Table 10 shows quantitative results from this exercise. SSCD correctly identifies the copy as the first result $2\times$ as often as SimCLR. Correct SSCD matches are nearly a superset of SimCLR matches: very rarely does SimCLR have a correct first result that SSCD misses.

Figure 7 shows additional queries and retrieved results for examples that only SSCD correctly identifies. One pattern we observe is that SimCLR often matches images with

Figure 6. Left and right columns: Pairs of matching images from the DISC2021 dataset. The central column shows which areas of the left image match best with the image on the right: yellow is strong match, blue is neutral or negative.

similar types of distortion together. Images with text at an angle, or strong diagonal features, may be incorrectly matched with images with similar features. Images with a blurry, or grainy, quality are matched to other images with a similar quality. This is surprising given that SimCLR trains with a blur augmentation, albeit weaker, and should be somewhat blur invariant.

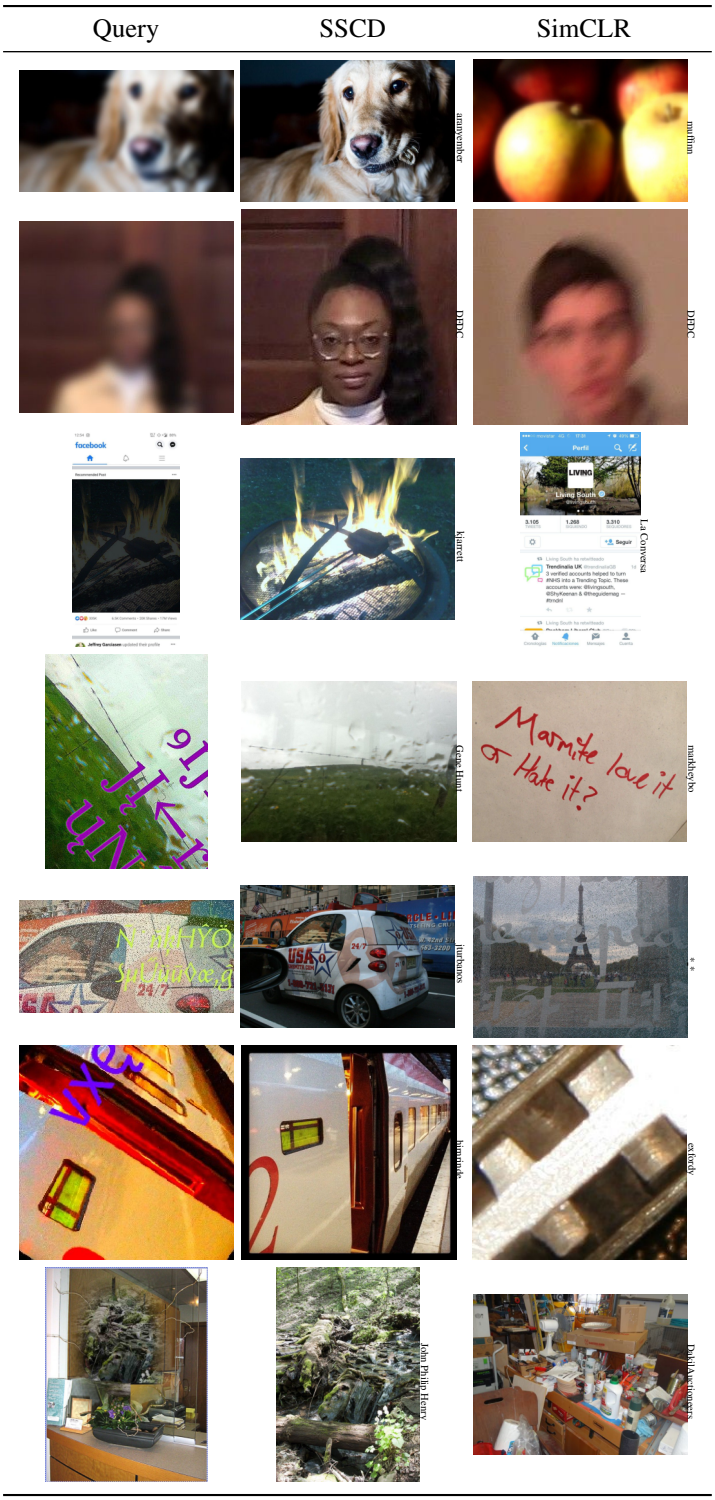

Figure 7. Example retrieval results from the DISC2021 dataset. For each row, we show the query image, the top retrieval result for SSCD, the top retrieval result for SimCLR.