Supplementary Material: Beyond the CLS Token: Image Reranking using Pretrained Vision Transformers

Chao Zhang[1]
chao.zhang@crl.toshiba.co.uk
Stephan Liwicki[1]
stephan.liwicki@crl.toshiba.co.uk
Roberto Cipolla[1,2]
rc10001@cam.ac.uk

[1] Cambridge Research Lab
Toshiba Europe Ltd
Cambridge, UK

[2] Department of Engineering
University of Cambridge
Cambridge, UK

1 Datasets

To evaluate the performance of the proposed method, we conduct experiments on three benchmarks used widely in the image retrieval task: CUB200-2011 [8], Cars196 [4], and Stanford Online Products [7]. We evaluate the method following a zero-shot image retrieval setting, where the training and test set contain image classes with no overlapping. Training/test split is the same as previous work [11]. Details of fine-grained image retrieval datasets are:

- CUB200-2011 contains 11,788 images of birds from 200 species. The first 100 classes (5,864 images) are used for training, while the other 100 classes (5,924 images) are used for testing.
- Cars196 contains 16,185 images of cars from 196 classes. We use the first 98 classes (8,054 images) for training and leave the remaining 98 classes (8,131 images) for testing.
- Stanford Online Products contains 120,053 images of 22,634 classes collected from ebay.com. We use the first 11,318 classes (59,551 images) for training and other 11,316 classes (60,502 images) for testing.

We also evaluate on MSLS [9] dataset, a benchmark for visual place recognition. It includes 1.68M images collected from 30 major cities across the globe of various season, time of day, date and viewpoint.

1.1 Data Augmentation

We apply color normalization using ImageNet precomputed mean: [0.485, 0.456, 0.406] and standard variation: [0.229, 0.224, 0.225]. During training, random resize crop is used to crop input to 224x224. Random horizontal flip with 0.5.

© 2022. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

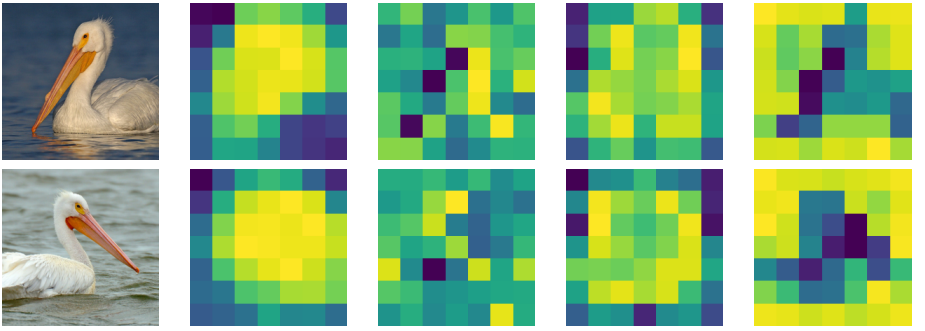


Figure 1: Visualization of CLS-to-Patch similarity. The value $M^{cls}_{i,j}$ indicates the correlation between the CLS token and patch token at location (i, j).

2 Analyzing Patch Similarity of Transformers

Vision transformers have been successfully applied in ImageNet [1] image classification, COCO [5] object detection and ADE20K [12] segmentation tasks. On one hand, for image classification, it is common to use CLS token as the global image representation. On the other hand, spatial feature maps of different scales are fused for semantic segmentation, similarly to CNNs. However, less attention is drawn towards the interaction between the CLS token and patch tokens. Beyond the difference of convolution and self-attention, CNNs and vision transformer are different in the way how the global representation is obtained. CNNs simply average/pool local feature maps, while transformer learn it via the CLS token. Hence, vision transformer may present some unique properties that CNNs do not possess.

For transformers, there are two types of tokens: CLS token for global embedding and patch tokens for local embeddings. We consider two types of similarity: CLS-Patch Similarity and Patch-Patch Similarity. We denote H and W as the height and width of patch representation.

CLS-Patch Similarity In Fig.1, we show the $H \times W$ similarity map M^{c2p}. The value $M^{c2p}_{i,j}$ denotes the similarity between CLS token and the patch at location (i, j). We note that cross-correlation between class token and patch tokens carries different information due to the self-attention layers. Specifically, the CLS token of CvT is less similar to patches belonging to the object, than that from background. We believe this is due to the random initialization of the attention mechanism of the CLS token, which is more similar to background pixels. In essence, background pixel store that they belong to an image that show an object, while foreground pixels need to retain the specific information of the object and thus differ to each other and the CLS token.

Patch-Patch Similarity We further demonstrate patch-to-patch similarity visualization in Fig.2 to show the difference in patch representation learning of several models. The patch with red box is moving from left to right, and the corresponding similarities to all other patches are computed. Specifically, we compare the results of ResNet-50 [3], DeiT-Small [7], Swin-Tiny [6] and CvT-13 [11]. We use ResNet-50 as the representative CNNs model, for its impact and widely usage in computer vision. DeiT-S is the vanilla ViT with similar

size of parameters. Swin Transformer is one hybird model fusing CNNs hierarchical and ViT self-attention design, but lacking CLS token. Finally, CvT-13 uses convolutional embedding and attention projection with CLS token in the output. A detailed comparison can be found in Tab.1

We now analyze the results in Fig.2. ResNet shows the most smooth results, and is able to segment background from foreground. Due to the increased receptive field, local patch has non-local understanding of the image. This is supported by the plot in col.2. DeiT produces less blurry segmentation, and shows its advantage of global receptive field via self-attention. Compared to ResNet, DeiT performs better in encoding more semantic information. Swin-T functions similarly to ResNet. This could be explained by that both Swin-T borrows the architecture design from ResNet, at the same time replacing Convolution with Attention. Basically, it gives rise to a better foreground and background segmentation. CvT gives the sharpest and best segmentation results among all, showing nice segmentation of the beaks in col.3. The patch features are discriminative, and demonstrate global understanding of the whole image.

Table 1: Comparison of architectures

Method	#params	Pos.Embed	CLS Token	Hierarchical	ImageNet top-1(%)
ResNet-50	23M	✗	✗	✓	76.2
DeiT-S	22M	✓	✓	✗	79.8
Swin-T	29M	✓	✗	✓	81.3
CvT-13	20M	✗	✓	✓	81.6

3 Results with Training

For both CNNs and Transformers, we have demonstrated the effectiveness of leveraging structural similarity for trained models. In this experiment, we show whether fusing global and structural similarity in training stage is helpful for image retrieval. Using the modified Margin loss proposed in [11], we train ResNet-50 and CvT-13 using corresponding baseline models on all datasets. In Tab.2, we find that training leads to performance gains on CUB200. While the results on Cars196 and SOP are not improved for both ResNet and CvT. On one hand, this indicates that training with global and structural similarity is not trivia, as we simply use equal weights following [11]. On the other hand, this proves the effectiveness of reranking through structural similarity. We also note that the implementation of DIML freezes batchnorm layers of ResNet-50 to avoid overfitting. We do not freeze any layers and simply finetune the CvT for each dataset.

Table 2: Effects of using structural similarity for training. In test mode, $K = 100$ is used as truncation number, and in train mode, $K = 1$ is used. Our method uses CvT-13 as backbone.

Method	Training	CUB200-2011			Cars196			SOP		
		P@1	RP	M@R	P@1	RP	M@R	P@1	RP	M@R
DIML	✗	64.97	35.28	24.45	**83.17**	**35.10**	**25.60**	**78.86**	**46.22**	**43.00**
	✓	**68.07**	**37.50**	**26.74**	78.02	33.90	23.52	78.66	45.30	42.21
Ours	✗	73.72	**42.68**	**32.12**	**83.66**	**35.63**	**25.82**	**77.15**	**44.45**	**41.36**
	✓	**75.00**	41.68	31.43	81.14	32.61	22.48	76.52	43.13	40.02

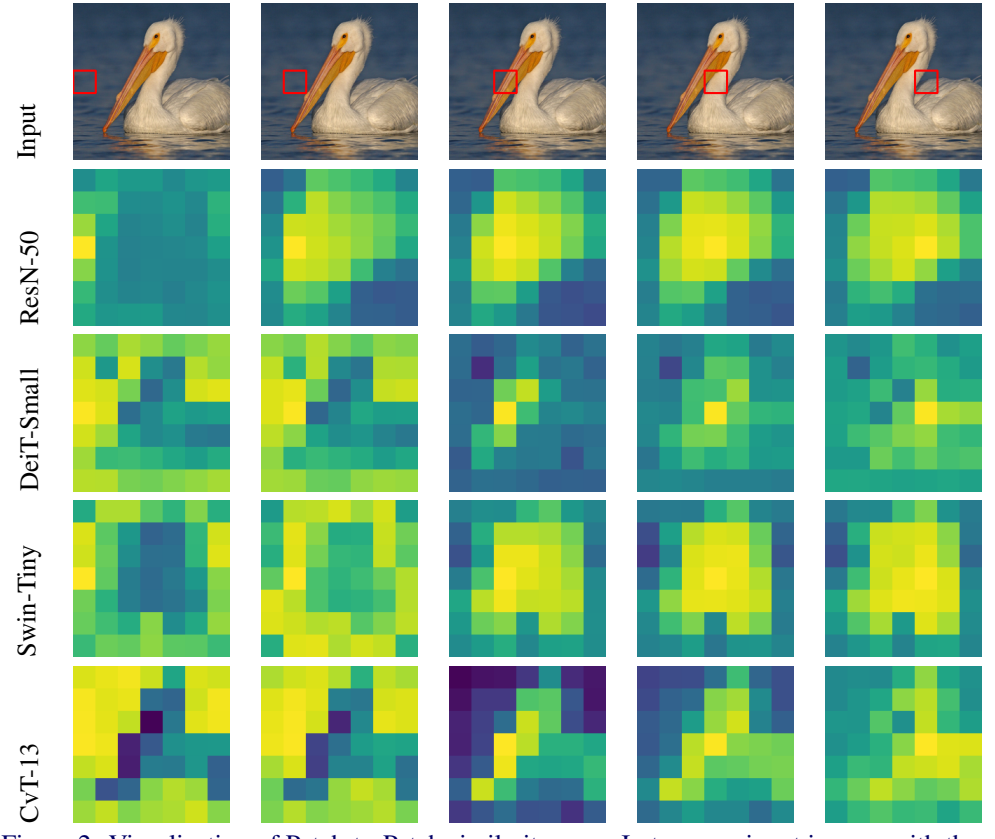


Figure 2: Visualization of Patch-to-Patch similarity map. In top row, input images with the the query patch are shown. The similarity between the highlighted patch and all patches are shown. ResNet, DeiT, Swin-T and CvT results are shown from row 2 to row 5.

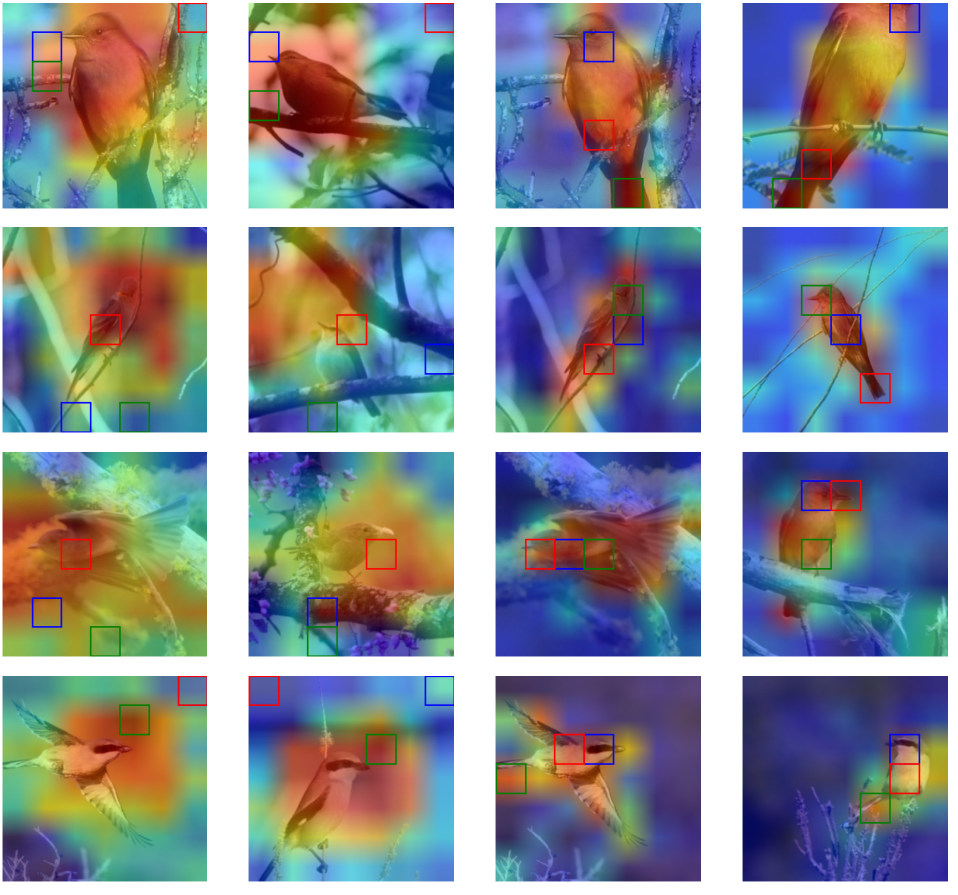
4 Additional Qualitative Results

4.1 CUB200 Visualization

In this section, we visualize some example images from CUB200 in Fig.3. For all samples, correlation-based spatial weighting is used for both DIML and CvT. Cross-correlation of ResNet is blurry and is not accurately located at the object. On the contrary, Transformer generates tighter importance map through cross-correlation. In addition, top matchings show that the optimal matching flow is accurate and robust to pose changes and non-rigid deformation.

4.2 Cars196 Visualizations

In Fig.4, for each query image, the top-5 retrieval results are presented for the two methods: ResNet-50 and CvT-13. A red bounding box indicates the wrong image label, while a green bounding box denotes the correct one. As illustrated by the top-1 results, ResNet shows it is picking images with similar layout, which can be understood as image level similarities. While Transformers method finds the matches with different poses and layouts. In the last



ResNet query ResNet match CvT query CvT match

Figure 3: Qualitative examples from CUB200-2011. Four query images are shown. The heatmap indicates the cross-correlation and inverse cross-correlation for Res50 and CvT-13. Top-3 matching are highlighted.

example, it shows that ResNet performs better than CvT. It is possibly that ResNet is using the car logo as important feature even it is small in scale. However, CvT is not sensitive to the this logo feature, possibly due to that the patch feature is either less locally discriminative, or is not considered important than other features, such as car plate features.

4.3 SOP Visualizations

Similarly, the top-5 retrieval results are presented for the two methods. As illustrated in Fig.5, CvT retrieves images with a single *fan* instance with similar structure, while ResNet sees multiple ones or the one with different color as positive. In the *bike* example, images with full bike layout are preferred by CvT, while parts images are chosen by ResNet. Similarly, both methods are challenged by query images with partial information, as shown in the *bike part* sample. In the *sofa chair* example, among negative retrieval results, we observe that CvT considers images containing sofa more similar, while ResNet picks images with chair

object.

5 Ablation Study using More Losses

We run additional experiments using Triplet Loss and Multi-Similarity Loss on CUB200. Results in Tab.3 show that our proposed method performs consistently well across different metric learning methods.

Table 3: Comparison of DIML and our method on CUB200-2011. **G** denotes global similarity, and **G+S** means both global and structural similarity are used. ↑ denotes the improvement of adding structural similarity. Our method outperforms DIML consistently and improves global similarity by a large margin.

Method	Type	Margin		Triplet		MS	
		P@1	RP	P@1	RP	P@1	RP
DIML	G	62.12	34.50	61.49	34.10	62.25	33.11
	G+S	64.97	35.28	63.63	34.57	65.68	34.02
	↑	**2.85**	0.78	**2.14**	0.47	**3.43**	0.91
Ours	G	71.75	41.94	72.24	42.04	72.78	41.62
	G+S	73.85	43.15	74.57	43.12	74.79	42.61
	↑	**2.10**	1.21	**2.33**	1.08	**2.01**	0.99

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[4] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013.

[5] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[6] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[7] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4004–4012, 2016.

[8] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[9] Frederik Warburg, Soren Hauberg, Manuel Lopez-Antequera, Pau Gargallo, Yubin Kuang, and Javier Civera. Mapillary street-level sequences: A dataset for lifelong place recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2626–2635, 2020.

[10] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22–31, 2021.

[11] Wenliang Zhao, Yongming Rao, Ziyi Wang, Jiwen Lu, and Jie Zhou. Towards interpretable deep metric learning with structural matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9887–9896, 2021.

[12] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017.

Figure 4: Visualization of retrieval results on Cars196. We compare the top-5 retrieval results using ResNet-50 and CvT-13. Both results are obtained using structural similarity. Cross-correlation and Inverse Cross-correlation are used for ResNet and CvT, respectively.

Figure 5: Visualization of retrieval results on SOP. We compare the top-5 retrieval results using ResNet-50 and CvT-13. Both results are obtained using structural similarity. Cross-correlation and Inverse Cross-correlation are used for ResNet and CvT, respectively.