SwinTransFuse: Fusing Swin and Multiscale Transformers for Fine-grained Image Recognition and Retrieval

Xu Ouyang[1], Tao Zhou[2], Rene Vidal[2], and Arnab Dhua[2]
[1]Illinois Institute of Technology
[2]Visual Search & AR, Amazon
xouyang3@hawk.iit.edu, {taozho,rvidal,adhua}@amazon.com

Abstract

Fine-grained recognition and retrieval are complex tasks in computer vision due to the high level of similarity between images of different subclasses. Recent work on fine-grained image recognition achieved significant improvements by using the attention mechanisms of the Vision or Swin Transformers to find discriminative image regions at coarse or fine scales, respectively. Here, we propose SwinTransFuse, a novel architecture for fine-grained recognition that fuses a Swin transformer and a Multiscale Vision transformer to capture both local and global features. We also propose Swin Transformer and SwinTransFuse siamese networks for fine-grained image retrieval. Our methods reach state-of-the-art performance on the CUB-200-2011 and Stanford Online Products fine-grained datasets.

1. Introduction

The goal of fine-grained image retrieval is to find images in an extensive database that belong to the same subclass as a given query image. This task finds numerous applications including face retrieval, bird recognition, and product search. Existing approaches to image retrieval [1, 15, 13, 10] are based on global descriptors such as VLAD [9] or Fisher vectors [14] obtained by aggregating features extracted from convolutional neural networks such as VGG19 [16] and ResNet [8]. However, global descriptors are insufficient for fine-grained retrieval due to the high similarity between images of different subclasses, which requires identifying regions in the image that are discriminative of the subclass.

Related work. Architectures based on transformers [18], such as the Vision Transformer (ViT) [3], TransFG [6], the Image Retrieval Transformer (IRT) [4], the Multiscale Transformer (MViT) [5], and the Swin Transformer [12], have recently achieved significant improvements in coarse- and fine-grained image recognition and retrieval. The ViT captures long-range interactions among image regions by dividing the image into patch tokens, computing several layers of self-attention between such tokens and a learned class token, and then classifying the class token. The TransFG extends the ViT to fine-grained recognition by using a part selection module to identify discriminative regions and remove redundant information. The selected part tokens, along with the class token, are fed to a network trained with a contrastive loss to further increase the distance between feature representations of samples from different subclasses and decrease that of samples from the same subclass. The IRT extends the ViT to image retrieval by using a siamese ViT network trained with a contrastive loss and differential entropy regularization. However, the discriminative power of the IRT is limited by its inability to localize discriminative regions and capture multiscale level information. The MViT addresses these issues by using using a multiscale pyramid where early layers operate at high resolutions and deep layers operate at coarse resolutions. However, the ViT, TransFG, IRT and MViT require large-scale training datasets to perform well and the complexity of computing self-attention among all token pairs is quadratic on the number of tokens. The Swin Transformer captures information at multiple spatial scales while achieving linear computational complexity by using a hierarchical design in which the computation of self-attention is reduced to tokens within local shifted windows. However, its use of local windows sacrifices the ability to capture global image features.

Paper contributions. This paper proposes SwinTransFuse, a combined architecture for fine-grained image recognition that leverages the complementary advantages of the Swin Transformer, which focuses more on local feature information, and a Multiscale Vision Transformer (MSViT), which captures global feature information at multiple scales. The SwinTransFuse architecture consists of multiple Swin and MSViT blocks connected in parallel, whose outputs are fused before passing to the next block. We also propose Swin Transformer and SwinTransFuse siamese networks

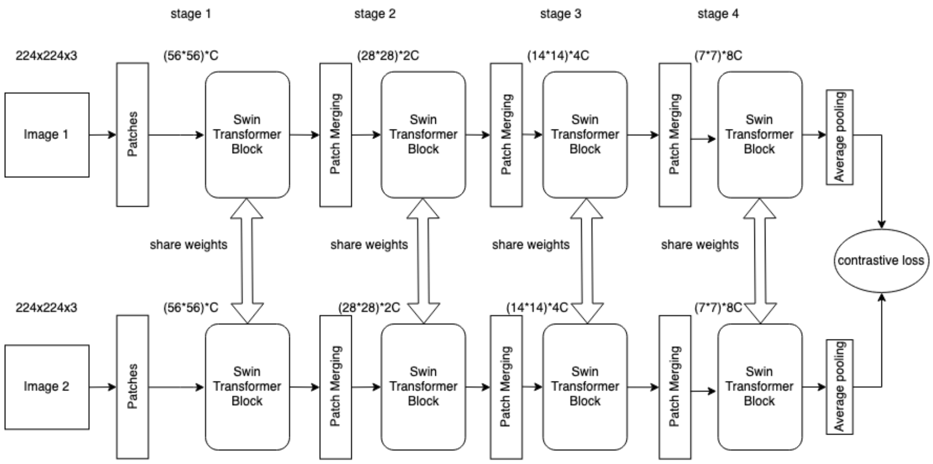

Figure 1. Swin Transformer Siamese Network

for fine-grained image retrieval. Both architectures are trained using a contrastive loss that compares a mini-batch to a cross batch memory (XBM) [20] queue at each iteration of training. Our experiments demonstrate that the proposed models achieve state-of-the-art performance in several datasets for fine-grained image retrieval tasks.

2. Methods

In this section, we introduce the proposed Swin Transformer Siamese Network and SwinTransFuse Siamese Network for fine-grained image recognition and retrieval.

2.1. Swin Transformer Siamese Network

The top row of Figure 1 shows the original architecture of the Swin Transformer [12]. In this architecture, a 224x224x3 image is fed to a convolutional layer whose output is a feature of size 56x56x128. This feature is fed to a network with four stages, each one consisting of a Swin Transformer block and a Patch Merging layer, which halve the resolution of the features and double the number of channels. When used for coarse-grained image recognition, the output of the last stage of this architecture is fed to a global average pooling layer followed by an MLP classification layer. The resulting network is trained by minimizing the categorical cross entropy loss $L_{cross\_entropy}$.

While the original Swin Transformer can also be used for fine-grained recognition by treating each subclass as a separate category, this strategy does not work well in practice because different subclasses can be very similar. To minimize the similarity of features corresponding to different subsclasses and maximize the similarity of features corresponding to the same subclass, we use a siamese network architecture trained with a contrastive loss. Specifically, we use the Swin Transformer Siamese network shown in Figure 1, which consists of two parallel Swin Transformers that share the same weights. We feed two different images, I_i and I_j, to the two Swin Transformers and compare their outputs, f_i and f_j, using the contrastive loss

$$L_{contrastive\_batch} = \frac{1}{N^2}\sum_{i}^{N}\Big[\sum_{j:y_i=y_j}^{N}(1-\cos\text{sim}(f_i,f_j))+ \sum_{j:y_i\neq y_j}^{N}\max((\cos\text{sim}(f_i,f_j)-\alpha),0)\Big], \tag{1}$$

where $\cos\text{sim}(f_i, f_j)$ is the cosine similarity of f_i and f_j, α is a margin on the cosine similarity of negative pairs that prevents the loss from being dominated by easy negatives, and N is the batch size. The final loss for fine-grained recognition is the sum of the cross-entropy and contrastive losses

$$L_{total} = L_{cross\_entropy} + \lambda * L_{contrastive\_batch}, \tag{2}$$

where λ is a parameter with default value of 1.

The loss for fine-grained image retrieval includes an additional contrastive loss with a cross-batch memory (XBM) module [20]. The reason is that the performance of image retrieval methods is related to their ability to mine informative negative pairs, which is limited by the size of the batch used in each training iteration. The XBM module maintains and updates as a queue that can store a large number of embeddings and thus help mine informative negative pairs. Specifically, at each iteration, the XBM enqueues the embeddings and labels of the current mini-batch, and dequeues the entities of the earliest mini-batch. This allows us to compute a contrastive loss $L_{contrastive\_xbm}$ between the current batch and the entire XBM by evaluating (1) with f_i being the output of the Swin Transformer for an image in the current batch, and f_j being one of the features in the

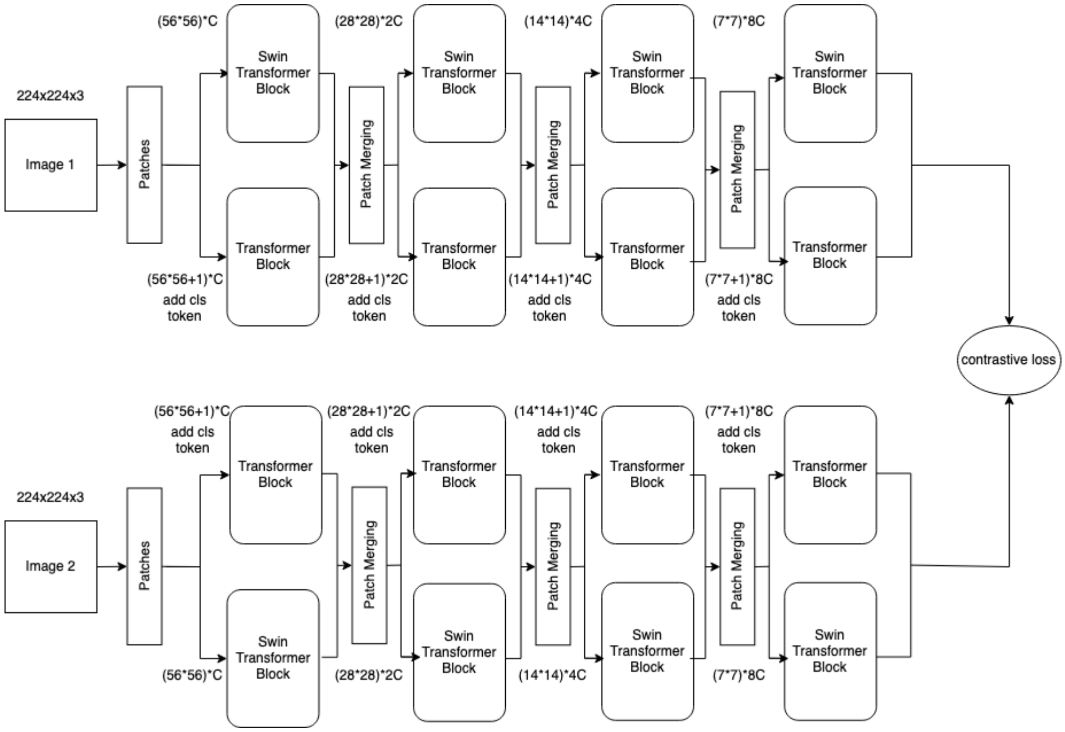

Figure 2. SwinTransFuse Siamese Network

XBM. The final loss for fine-grained retrieval is given by

$$L_{total} = L_{contrastive\_batch} + k * L_{contrastive\_xbm}, \qquad (3)$$

where k is a parameter whose default value is 1.

2.2. SwinTransFuse Siamese Network

First, we present the proposed SwinTransFuse network for fine-grained image recognition, which combines a Swin Transformer and a Multiscale Vision Transformer (MSViT) as shown in the top row of Figure 2. As described in the previous section, the 224x224x3 input image is first fed to a convolutional layer that generates a feature of size 56x56x128. This new feature is then fed to a network with four stages, each one consisting of a Transformer block, a Swin Transformer block, and a Patch Merging layer. The Swin Transformer block and the Patch Merging layer are identical to those in the Swin Transformer described before. The main difference is that, at each stage, the input feature is fed in parallel to a Transformer block and a Swin Transformer block, whose outputs are then added to produce the output of that stage. Note that different Transformer blocks have different spatial scales, hence their concatenation defines a Multiscale Vision Transformer. However, the so obtained MSViT is different from the MViT in [5] in that the downsampling happens outside the block to parallel the architecture of the Swin Transformer. Finally, the output of the SwinTransFuse network is fed to a pooling layer and an MLP classifier, and the overall network is trained using the sum of the cross entropy and contrastive losses, as in (2).

Next, we present the SwinTransFuse Siamese network for fine-grained image retrieval, which consists of two parallel SwinTransFuse networks that share the same weights, as shown in Figure 2. We separately feed two different images into these two parallel SwinTransFuse networks and compute two contrastive losses between their outputs. The first one is the contrastive loss between the two outputs of the SwinTransFuse Siamese network, and the second one is the contrastive loss between the output of one SwinTransFuse network and the XBM of the output of the other SwinTransFuse network, as in (3).

3. Experiments

3.1. Datasets

We evaluate the proposed Swin Transformer Siamese and SwinTransFuse Siamese networks on the CUB-200-2011 [2], Stanford Online Products [17], and Cars196 [11] fine-grained datasets, whose details are shown in Table 1.

Table 1. Datasets for fine-grained tasks

Dataset	Category	Training	Testing
CUB-200-2011	200	5994	5794
Cars196	196	8144	8041
Stanford Online Products	22634	59551	60502

3.2. Implementation Details

During training we resize the images to 256x256 and crop them to 224x224. During evaluation we resize the images to 256x256. The batch size is set to 8 and the XBM size is set to 8,192. We load the pretrained Swin Transformer model on ImageNet21K. We use the SGD optimizer with a momentum of 0.9. The learning rate is initialized as 0.03 and we use cosine annealing as the optimizer scheduler. We train 40K steps for all experiments. For fine-grained image retrieval, we first train the Swin Transformer and the SwinTransFuse networks on the image recognition task, then train the Swin Transformer Siamese network and the SwinTransFuse Siamese network on the image retrieval task with the pretrained models from the image recognition task.

3.3. Results

We compare our SwinTransFuse with the Swin Transformer on the fine-grained image recognition task on the Stanford Online Products, CUB-200-2011 and Cars196 datasets. As we can see in Table 2, our method achieves higher accuracy than the Swin Transformer on all datasets[1].

We then compare our SwinTransFuse Siamese network and the Swin Transformer Siamese network with CGD [10] and IRT [4] on the fine-grained image retrieval task. The Recall@K scores for the three datasets are shown in Table 3, Table 4 and Table 5. As we can see, the Swin Transformer Siamese network achieves state-of-the-art performance for all Recall@K scores on the SOP and CUB datasets, while the SwinTransFuse Siamese Network gets better Recall@K scores than the Swin Transformer Siamese network on the Cars196 dataset. However, the CGD method still achieves the best performance on the Cars196 dataset. We conjecture this is due to our use of the standard Swin network, thus we plan to explore using the Swin large version in the future.

4. Conclusion

This paper proposed various novel architecture for fine-grained image recognition and retrieval. For fine-grained recognition, we proposed SwinTransFuse, which combines the complementary advantages of a Swin Transformer and a Multiscale Vision Transformer to capture both local and global information at different scales. Our experiments showed that SwinTransFuse achieves state-of-the-art performance in fine-grained image recognition on three different datasets. We also proposed two siamese architectures for fine-grained retrieval, Swin Tranformer Siamese Network and SwinTransFuse Siamese Network, and trained them with a contrastive loss with cross batch memory. Our experiments showed that the former achieves state-of-the-art performance on the SOP and CUB-200-2011 datasets.

[1] The input image size is 224*224, while others [7, 19] use 448*448.

Table 2. Fine-grained image recognition task on SOP, CUB-200-2011, Cars196 dataset

Model	Feature size	SOP accuracy	CUB-200-2011 accuracy	Cars196 accuracy
Swin Transformer	1024	89.64%	87.03%	91.55%
SwinTransFuse Network	1024	**90.10%**	**87.24%**	**91.81%**

Table 3. Fine-grained image retrieval task on Stanford Online Products (SOP) dataset

Model	Dataset	Feature size	recall@1	recall@10	recall@100	recall@1000
CGD	SOP	1536	84.2	93.9	97.4	99.2
IRT	SOP	384	84.2	93.7	97.3	99.1
Swin Transformer Siamese Network	SOP	1024	**88.4**	**95.9**	**98.2**	**99.3**
SwinTransFuse Siamese Network	SOP	1024	88.1	95.7	98.1	**99.3**

Table 4. Fine-grained image retrieval task on CUB-200-2011 bird dataset

Model	Dataset	Feature size	recall@1	recall@2	recall@4	recall@8
CGD	CUB-200-2011	1536	79.2	86.6	92.0	**95.1**
IRT	CUB-200-2011	384	76.6	85.0	91.1	94.3
Swin Transformer Siamese Network	CUB-200-2011	1024	**86.1**	**90.4**	**92.8**	94.8
SwinTransFuse Siamese Network	CUB-200-2011	1024	86.0	90.2	**92.8**	94.6

Table 5. Fine-grained image retrieval task on Cars196 dataset

Model	Dataset	Feature size	recall@1	recall@2	recall@4	recall@8
CGD	Cars196	1536	**94.8**	**97.1**	**98.2**	**98.8**
IRT	Cars196	384	-	-	-	-
Swin Transformer Siamese Network	Cars196	1024	92.2	95.4	97.2	98.3
SwinTransFuse Siamese Network	Cars196	1024	92.4	95.6	97.3	98.3

References

[1] Artem Babenko and Victor Lempitsky. Aggregating deep convolutional features for image retrieval. *arXiv preprint arXiv:1510.07493*, 2015. 1

[2] P. Welinder P. Perona C. Wah, S. Branson and S. Belongie. The caltech-ucsd birds-200-2011 dataset. *Technical Report*, CNS-TR-2011-001, 2011. 3

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 1

[4] Alaaeldin El-Nouby, Natalia Neverova, Ivan Laptev, and Hervé Jégou. Training vision transformers for image retrieval. *CoRR*, abs/2102.05644, 2021. 1, 4

[5] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6824–6835, 2021. 1, 3

[6] Ju He, Jieneng Chen, Shuai Liu, Adam Kortylewski, Cheng Yang, Yutong Bai, Changhu Wang, and Alan L. Yuille. Transfg: A transformer architecture for fine-grained recognition. *CoRR*, abs/2103.07976, 2021. 1

[7] Ju He, Jie-Neng Chen, Shuai Liu, Adam Kortylewski, Cheng Yang, Yutong Bai, Changhu Wang, and Alan Yuille. Transfg: A transformer architecture for fine-grained recognition. *arXiv preprint arXiv:2103.07976*, 2021. 4

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. 1

[9] Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez. Aggregating local descriptors into a compact image representation. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3304–3311. IEEE, 2010. 1

[10] HeeJae Jun, ByungSoo Ko, Youngjoon Kim, Insik Kim, and Jongtack Kim. Combination of multiple global descriptors for image retrieval. *CoRR*, abs/1903.10663, 2019. 1, 4

[11] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013. 3

[12] 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. *CoRR*, abs/2103.14030, 2021. 1, 2

[13] Eng-Jon Ong, Sameed Husain, and Miroslaw Bober. Siamese network of deep fisher-vector descriptors for image retrieval. *CoRR*, abs/1702.00338, 2017. 1

[14] Florent Perronnin, Yan Liu, Jorge Sánchez, and Hervé Poirier. Large-scale image retrieval with compressed fisher vectors. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3384–3391. IEEE, 2010. 1

[15] Ali S Razavian, Josephine Sullivan, Stefan Carlsson, and Atsuto Maki. Visual instance retrieval with deep convolutional networks. *ITE Transactions on Media Technology and Applications*, 4(3):251–258, 2016. 1

[16] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 1

[17] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 4004–4012. IEEE Computer Society, 2016. 3

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. 1

[19] Jun Wang, Xiaohan Yu, and Yongsheng Gao. Feature fusion vision transformer for fine-grained visual categorization. *arXiv preprint arXiv:2107.02341*, 2021. 4

[20] Xun Wang, Haozhi Zhang, Weilin Huang, and Matthew R. Scott. Cross-batch memory for embedding learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 6387–6396. Computer Vision Foundation / IEEE, 2020. 2