Dual Cross-Attention Learning for Fine-Grained Visual Categorization and Object Re-Identification

Haowei Zhu*, Wenjing Ke*, Dong Li, Ji Liu, Lu Tian, Yi Shan
Advanced Micro Devices, Inc., Beijing, China
{haowei.zhu, wenjing.ke, d.li, lu.tian, yi.shan}@amd.com

Abstract

Recently, self-attention mechanisms have shown impressive performance in various NLP and CV tasks, which can help capture sequential characteristics and derive global information. In this work, we explore how to extend self-attention modules to better learn subtle feature embeddings for recognizing fine-grained objects, e.g., different bird species or person identities. To this end, we propose a dual cross-attention learning (DCAL) algorithm to coordinate with self-attention learning. First, we propose global-local cross-attention (GLCA) to enhance the interactions between global images and local high-response regions, which can help reinforce the spatial-wise discriminative clues for recognition. Second, we propose pair-wise cross-attention (PWCA) to establish the interactions between image pairs. PWCA can regularize the attention learning of an image by treating another image as distractor and will be removed during inference. We observe that DCAL can reduce misleading attentions and diffuse the attention response to discover more complementary parts for recognition. We conduct extensive evaluations on fine-grained visual categorization and object re-identification. Experiments demonstrate that DCAL performs on par with state-of-the-art methods and consistently improves multiple self-attention baselines, e.g., surpassing DeiT-Tiny and ViT-Base by 2.8% and 2.4% mAP on MSMT17, respectively.

1. Introduction

Self-attention is an attention mechanism that can relate different positions of a single sequence and draw global dependencies. It is originally applied in natural language processing (NLP) tasks [10, 46] and exhibits the outstanding performance. Recently, Transformer with self-attention learning has also been explored for various vision tasks (e.g., image classification [5, 12, 19, 37, 45, 51] and object detection [2, 68]) as an alternative of convolutional neural network (CNN). For general image classification, self-attention has been proved to work well for recognizing 2D images by viewing image patches as words and flattening them as sequences [12, 45].

In this work, we investigate how to extend self-attention modules to better learn subtle feature embeddings for recognizing fine-grained objects, e.g., different bird species or person identities. Fine-grained recognition is more challenging than general image classification owing to the subtle visual variations among different sub-classes. Most of existing approaches build upon CNN to predict class probabilities or measure feature distances. To address the subtle appearance variations, local characteristics are often captured by learning spatial attention [15, 34, 40, 60] or explicitly localizing semantic objects / parts [11, 56, 58, 61]. We adopt a different way to incorporate local information based on vision Transformer. To this end, we propose global-local cross-attention (GLCA) to enhance the interactions between global images and local high-response regions. Specifically, we compute the cross-attention between a selected subset of query vectors and the entire set of key-value vectors. By coordinating with self-attention learning, GLCA can help reinforce the spatial-wise discriminative clues to recognize fine-grained objects.

Apart from incorporating local information, another solution to distinguish the sutble visual differences is pair-wise learning. The intuition is that one can identify the subtle variations by comparing image pairs. Exiting CNN-based methods design dedicated network architectures to enable pair-wise feature interaction [16, 69]. A contrastive loss [16] or score ranking loss [69] is used for feature learning. Motivated by this, we also employ a pair-wise learning scheme to establish the interactions between image pairs. Different from optimizing the feature distance, we propose pair-wise cross-attention (PWCA) to regularize the attention learning of an image by treating another image as distractor. Specifically, we compute the cross-attention between query of an image and combined key-value from both images. By introducing confusion in key and value vectors, the attention scores are diffused to another image so that

*Equal contribution.

the difficulty of the attention learning of the current image increases. Such regularization allows the network to discover more discriminative regions and alleviate overfitting to sample-specific features. It is noted that PWCA is only used for training and thus does not introduce extra computation cost during inference.

The proposed two types of cross-attention are easy-to-implement and compatible with self-attention learning. We conduct extensive evaluations on both fine-grained visual categorization (FGVC) and object re-identification (Re-ID). Experiments demonstrate that DCAL performs on par with state-of-the-art methods and consistently improves multiple self-attention baselines. Particularly, for FGVC, DCAL improves DeiT-Tiny by 2.5% and reaches 92.0% top-1 accuracy with the larger R50-ViT-Base backbone on CUB-200-2011. For Re-ID, DCAL improves DeiT-Tiny and ViT-Base by 2.8% and 2.4% mAP on MSMT17, respectively.

Our main contributions can be summarized as follows. (1) We propose global-local cross-attention to enhance the interactions between global images and local high-response regions for reinforcing the spatial-wise discriminative clues. (2) We propose pair-wise cross-attention to establish the interactions between image pairs by regularizing the attention learning. (3) The proposed dual cross-attention learning can complement the self-attention learning and achieves consistent performance improvements over multiple vision Transformer baselines on various FGVC and Re-ID benchmarks.

2. Related Work

2.1. Self-Attention Mechanism

The self-attention mechanism is originally proposed to relate distinct positions in a sequence and draw global dependencies. Transformer carrying forward this mechanism has dominated in various sequence-to-sequence NLP tasks [10, 46]. Transformer usually consists of multiple encoder and decoder modules. Each encoder / decoder includes a multi-head self-attention (MSA) layer and a feed-forward network (FFN) layer. A decoder also has an extra MSA layer to handle the output of encoder. Besides, layer normalization (LN) and residual connection are used in each MSA or FFN layer. Recent work has applied Transformers to various vision tasks (e.g., image classification [5, 12, 19, 37, 45, 51], object detection [2, 41, 44, 68], semantic segmentation [23, 39, 50, 51, 63] and low-level tasks [4]) and shown competitive performance compared to the state-of-the-art CNNs. For general image classification, iGPT [5] first uses auto-regressive and BERT [10] objectives for self-supervised pre-training and then fine-tunes for classification tasks. ViT [12] reshapes an image into a sequence of flattened fixed-size patches for training Transformer encoders only. Attempts have also been made to improve ViT by knowledge distillation [45] and progressive tokenization [57]. Fine-grained recognition is more challenging than general image classification owing to the sutble visual variations among different sub-classes. In this work, we extend self-attention to better recognize fine-grained objects with two types of cross-attention modules.

2.2. Fine-Grained Visual Categorization

Fine-grained visual categorization (FGVC) is a special case of image classification, which aims to identify those highly-confused categories with fine differences. Prior CNN-based methods address this task by mining effective information from multi-level features [13, 34, 58], adopting multi-granularity training strategies [13], locating discriminative objects or parts [11, 61] and exploring feature interaction in pair-wise learning [16, 69]. Recently, a few Transformer-based methods address FGVC by feature fusion on multi-level Transformer layers [52] and part selection [17]. Our motivation is similar with [17, 52] in the aspects of aggregating multi-level attention and selecting patch tokens. However, they are based on self-attention only while we design two cross-attention modules for learning.

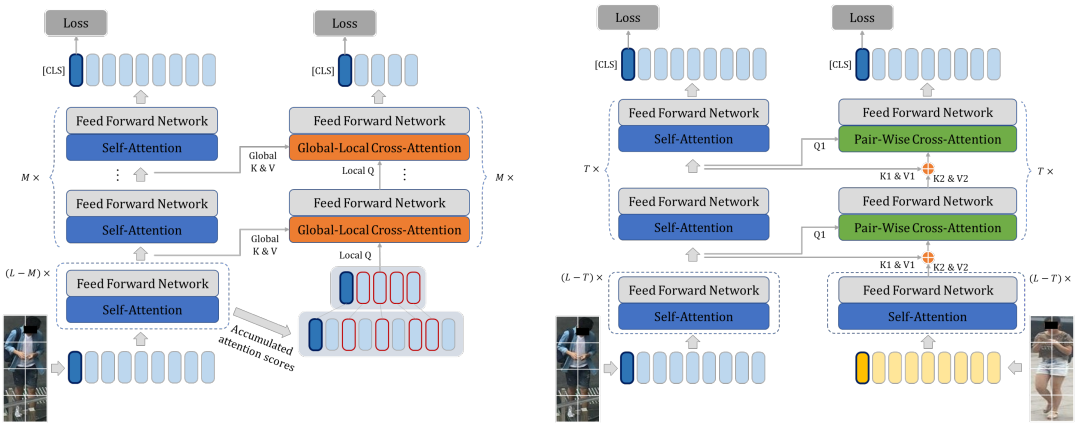
2.3. Object Re-Identification

Similar to FGVC, object re-identification also aims to distinguish different person / vehicle identities with subtle inter-class differences. Mainstream Re-ID methods are based on the CNN structure and metric learning [30, 32]. Local information is crucial for Re-ID and many different approaches have been presented by encoding discriminative part-level features [31, 42, 49]. Transformer with self-attention structure has recently been applied to Re-ID by introducing part tokens [67], shuffling patch embeddings [17], and learning disentangled features [24]. Our work differs from the most related methods [17, 67] in the following aspects. First, we adopt a different way to encode the local information by GLCA, while [17] does not explicitly mine part regions and [67] computes the attention between a part token and its associated subset of patch embeddings by online clustering. Second, [17, 67] uses a single image for training while we employ image pairs for PWCA. Third, [17] requires side information (e.g., camera IDs and viewpoint labels) while our method only takes images as input.

3. Proposed Approach

3.1. Revisit Self-Attention

[46] originally proposes the self-attention mechanism to address NLP tasks by calculating the correlation between each word and all the other words in the sentence. [12] inherits the idea by taking each patch in the image / feature map as a word for general image classification. In gen-



(a) Global-Local Cross-Attention (GLCA) (b) Pair-Wise Cross-Attention (PWCA)

Figure 1. Overview of the proposed two types of cross-attention mechanisms. We stack L self-attention, M global-local cross-attention, T pair-wise cross-attention modules in our network. See Section 3 for details.

eral, a self-attention function can be depicted as mapping a query vector and a set of key and value vectors to an output. The output is computed as a weighted sum of value vectors, where the weight assigned to each value is computed by a scaled inner product of the query with the corresponding key. Specifically, a query $q \in \mathbb{R}^{1\times d}$ is first matched against N key vectors ($K = [k_1; k_2; \cdots; k_N]$, where each $k_i \in \mathbb{R}^{1\times d}$) using inner product. The products are then scaled and normalized by a softmax function to obtain N attention weights. The final output is the weighted sum of N value vectors ($V = [v_1; v_2; \cdots; v_N]$, where each $v_i \in \mathbb{R}^{1\times d}$). By packing N query vector into a matrix $Q = [q_1; q_2; \cdots; q_N]$, the output matrix of self-attention (SA) can be represented as:

$$f_{\text{SA}}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d}})V = SV \qquad (1)$$

where $\frac{1}{\sqrt{d}}$ is a scaling factor. Query, key and value matrices are computed from the same input embedding $X \in \mathbb{R}^{N\times D}$ with different linear transformations: $Q = XW_Q$, $K = XW_K$, $V = XW_V$, respectively. $S \in \mathbb{R}^{N\times N}$ denotes the attention weight matrix.

To jointly attend to information from different representation subspaces at different positions, multi-head self-attention (MSA) is defined by considering multiple attention heads. The process of MSA can be computed as linear transformation on the concatenations of self-attention blocks with subembeddings. To encode positional information, fixed / learnable position embeddings are added to patch embeddings and then fed to the network. To predict the class, an extra class embedding $\hat{\texttt{CLS}} \in \mathbb{R}^{1\times d}$ is prepended to the input embedding X throughout the network, and finally projected with a linear classifer layer for prediction. Thus, the input embeddings as well as query, key and value matrices become $(N + 1) \times d$ and the self-attention function (Eq. 1) allows to spread information between patch and class embeddings.

Based on self-attention, a Transformer encoder block can be constructed by an MSA layer and a feed forward network (FFN). FFN consists of two linear transformation with a GELU activation. Layer normalization (LN) is put prior to each MSA and FFN layer and residual connections are used for both layers.

3.2. Global-Local Cross-Attention

Self-attention treats each query equally to compute global attention scores according to Eq. 1. In other words, each local position of image is interacted with all the positions in the same manner. For recognizing fine-grained objects, we expect to mine discriminative local information to facilitate the learning of subtle features. To this end, we propose global-local cross-attention to emphasize the interaction between global images and local high-response regions. First, we follow attention rollout [1] to calculate the accumulated attention scores for i-th block:

$$\hat{S}_i = \bar{S}_i \otimes \bar{S}_{i-1} \cdots \otimes \bar{S}_1 \qquad (2)$$

where $\bar{S} = 0.5S + 0.5E$ means the re-normalized attention weights using an identity matrix E to consider residual connections, $\otimes$ means the matrix multiplication operation. In this way, we track down the information propagated from the input layer to a higher layer. Then, we use the aggregated attention map to mine the high-response regions. According to Eq. 2, the first row of $\hat{S}_i = [\hat{s}_{i,j}]_{(N+1)\times(N+1)}$ means the accumulated weights of class embedding $\hat{\texttt{CLS}}$. We select top R query vectors from Q_i that correspond to the top R highest responses in the accumulated weights of $\hat{\texttt{CLS}}$ to construct a new query matrix Q^l, representing the most attentive local embeddings. Finally, we compute

the cross attention between the selected local query and the global set of key-value pairs as below.

$$f_{\text{GLCA}}(Q^l, K^g, V^g) = \text{softmax}(\frac{Q^l K^{gT}}{\sqrt{d}})V^g \quad (3)$$

In self-attention (Eq. 1), all the query vectors will be interacted with the key-value vectors. In our GLCA (Eq. 3), only a subset of query vectors will be interacted with the key-value vectors. We observe that GLCA can help reinforce the spatial-wise discriminative clues to promote recognition of fine-grained classes. Another possible choice is to compute the self-attention between local query Q^l and local key-value vectors (K^l, V^l). However, through establishing the interaction between local query and global key-value vectors, we can relate the high-response regions with not only themselves but also with other context outside of them. Figure 1 (a) illustrates the proposed global-local cross-attention and we use $M = 1$ GLCA block in our method.

3.3. Pair-Wise Cross-Attention

The scale of fine-grained recognition datasets is usually not as large as that of general image classification, e.g., ImageNet [9] contains over 1 million images of 1,000 classes while CUB [47] contains only 5,994 images of 200 classes for training. Moreover, smaller visual differences between classes exist in FGVC and Re-ID compared to large-scale classification tasks. Fewer samples per class may lead to network overfitting to sample-specific features for distinguishing visually confusing classes in order to minimize the training error.

To alleviate the problem, we propose pair-wise cross attention to establish the interactions between image pairs. PWCA can be viewed as a novel regularization method to regularize the attention learning. Specifically, we randomly sample two images (I_1, I_2) from the same training set to construct the pair. The query, key and value vectors are separately computed for both images of a pair. For training I_1, we concatenate the key and value matrices of both images, and then compute the attention between the query of the target image and the combined key-value pairs as follows:

$$f_{\text{PWCA}}(Q_1, K_c, V_c) = \text{softmax}(\frac{Q_1 K_c^T}{\sqrt{d}})V_c \quad (4)$$

where $K_c = [K_1; K_2] \in \mathbb{R}^{(2N+2)\times d}$ and $V_c = [V_1; V_2] \in \mathbb{R}^{(2N+2)\times d}$. For a specific query from I_1, we compute $N+1$ self-attention scores within itself and $N+1$ cross-attention scores with I_2 according to Eq. 4. All the $2N+2$ attention scores are normalized by the softmax function together and thereby contaminated attention scores for the target image I_1 are learned. Optimizing this noisy attention output increases the difficulty of network training and reduces the overfitting to sample-specific features. Figure 1 (b) illustrates the proposed pair-wise cross-attention and we use $T = 12$ PWCA blocks in our method. Note that PWCA is only used for training and will be removed for inference without consuming extra computation cost.

4. Experiments

4.1. Experimental Setting

Datasets. We conduct extensive experiments on two fine-grained recognition tasks: fine-grained visual categorization (FGVC) and object re-identification (Re-ID). For FGVC, we use three standard benchmarks for evaluations: CUB-200-2011 [47], Stanford Cars [27], FGVC-Aircraft [35]. For Re-ID, we use four standard benchmarks: Market1501 [62], DukeMTMC-ReID [54], MSMT17 [53] for Person Re-ID and VeRi-776 [64] for Vehicle Re-ID. In all experiments, we use the official train and validation splits for evaluation.

Baselines. We use DeiT and ViT as our self-attention baselines. In detail, ViT backbones are pre-trained on ImageNet-21k [9] and DeiT backbones are pre-trained on ImageNet-1k [9]. We use multiple architectures of DeiT-T/16, DeiT-S/16, DeiT-B/16, ViT-B/16, R50-ViT-B/16 with $L = 12$ SA blocks for evaluation.

Implementation Details. We coordinate the proposed two types of cross-attention with self-attention in the form of multi-task learning. We build $L = 12$ SA blocks, $M = 1$ GLCA blocks and $T = 12$ PWCA blocks as the overall architecture for training. The PWCA branch shares weights with the SA branch while GLCA does not share weights with SA. We follow [59] to adopt dynamic loss weights for collaborative optimization, avoiding exhausting manual hyper-parameter search. The PWCA branch has the same GT target as the SA branch since we treat another image as distractor.

For FGVC, we resize the original image into 550×550 and randomly crop to 448×448 for training. The sequence length of input embeddings for self-attention baseline is $28 \times 28 = 784$. We select input embeddings with top $R = 10\%$ highest attention responses as local queries. We apply stochastic depth [21] and use Adam optimizer with weight decay of 0.05 for training. The learning rate is initialized as $\text{lr}_{scaled} = \frac{5e-4}{512} \times batchsize$ and decayed with a cosine policy. We train the network for 100 epochs with batch size of 16 using the standard cross-entropy loss.

For Re-ID, we resize the image into 256×128 for pedestrian datasets, and 256×256 for vehicle datasets. We select input embeddings with top $R = 30\%$ highest attention responses as local queries. We use SGD optimizer with a momentum of 0.9 and a weight decay of 1e-4. The batch size is set to 64 with 4 images per ID. The learning rate is initialized as 0.008 and decayed with a cosine policy. We train the

Method	Backbone	Accuracy (%)		
		CUB	CAR	AIR
RA-CNN [15]	VGG19	85.3	92.5	88.4
MA-CNN [60]	VGG19	86.5	92.8	89.9
MAMC [40]	ResNet101	86.5	93.0	-
PC [14]	DenseNet161	86.9	92.9	89.2
FDL [29]	DenseNet161	89.1	94.0	-
NTS-Net [56]	ResNet50	87.5	93.9	91.4
Cross-X [34]	ResNet50	87.7	94.6	-
S3N [11]	ResNet50	88.5	94.7	92.8
MGE-CNN [58]	ResNet50	88.5	93.9	-
DCL [8]	ResNet50	87.8	94.5	93.0
TASN [61]	Resnet50	87.9	93.8	-
PMG [13]	ResNet50	89.6	95.1	93.4
CIN [16]	ResNet50	88.1	94.5	92.8
API-Net [69]	DenseNet161	90.0	95.3	93.9
LIO [65]	ResNet50	88.0	94.5	92.7
SPS [22]	ResNet50	88.7	94.9	92.7
CAL [38]	ResNet101	90.6	95.5	94.2
TransFG [17]	ViT-Base	91.7	94.8	-
RAMS-Trans [20]	ViT-Base	91.3	-	-
FFVT [52]	ViT-Base	91.6	-	-
Baseline	DeiT-Tiny	82.1	87.2	84.7
Baseline + DCAL	DeiT-Tiny	84.6	89.4	87.4
Baseline	DeiT-Small	85.8	90.7	88.1
Baseline + DCAL	DeiT-Small	87.6	92.3	90.0
Baseline	DeiT-Base	88.0	92.9	90.3
Baseline + DCAL	DeiT-Base	88.8	93.8	92.6
Baseline	ViT-Base	90.8	92.5	90.0
Baseline + DCAL	ViT-Base	91.4	93.4	91.5
Baseline	R50-ViT-Base	91.3	94.0	92.4
Baseline + DCAL	R50-ViT-Base	92.0	95.3	93.3

Table 1. Performance comparisons in terms of top-1 accuracy on three standard FGVC benchmarks: CUB-200-2011, Stanford Cars and FGVC-Aircraft.

network for 120 epochs using the cross-entropy and triplet losses.

All of our experiments are conducted on PyTorch with Nvidia Tesla V100 GPUs. Our method costs 3.8 hours with DeiT-Tiny backbone for training using 4 GPUs on CUB, and 9.5 hours with ViT-Base for training using 1 GPU on MSMT17. During inference, we remove all the PWCA modules and only use the SA and GLCA modules. We add class probabilities output by classifiers of SA and GLCA for prediction for FGVC, and concat two final class tokens of SA and GLCA for prediction for Re-ID. A single image with the same input size as training is used for test.

4.2. Results on Fine-Grained Visual Categorization

We evaluate our method on three standard FGVC benchmarks and compare with the state-of-the-art approaches in Table 1. Our method achieves competitive performance compared to the prior CNN-based and Transformer-based methods. Particularly, with the R50-ViT-Base backbone, DCAL reaches 92.0%, 95.3% and 93.3% top-1 accuracy on CUB-200-2011, Stanford Cars and FGVC-Aircraft benchmarks, respectively. Table 1 also shows our method can consistently improve different vision Transformer baselines on all the three benchmarks, e.g., surpassing the pure Transformer (DeiT-Tiny) by 2.2% and the hybrid structure of CNN and Transformer (R50-ViT-Base) by 1.3% on Stanford Cars. The results validate the compatibility of our method to different Transformer architectures.

Comparisons to Transformer-based Methods. Our method performs on par with the recent Transformer variants on FGVC: TransFG [17], RAMS-Trans [20], FFVT [52]. These existing methods also select tokens based on aggregated attention responses. Differently, they continue to model the selected tokens by self-attention while we perform cross-attention between local query and global key-value vectors. Compared to self-attention in selected tokens, we can relate the high-response regions with not only themselves but also with other context outside of them. Besides, TransFG [17] uses overlapping patches and will largely increase training time and computation overhead, while we adopt the standard non-overlapping patch split method.

Comparisons to CNN-based Methods. (1) Existing region-based methods can be divided to two categories. Explicit localization methods (e.g, RACNN [15], MA-CNN [60], NTS-Net [56], MGE-CNN [58]) utilize attention / localization sub-network with ranking losses to mine object regions. Implicit localization methods (e.g., S3N [11], TASN [61]) use class activation map and Gaussian sampling to amplify object regions in the original image. Our GLCA adopts a different scheme to incorporate the local information with higher performance, e.g., +3.5% over MGE-CNN on CUB. (2) Pair-wise learning is also applied for FGVC by interacting features (CIN [16], API-Net [69]) or introducing confusion (PC [14], SPS [22]) between image pairs during training. Our motivation of PWCA is similar to [14, 22] but we implement a different regularization method to alleviate overfitting. Our method surpasses these related pair-wise learning methods, e.g., +3.9% over CIN and +5.1% over PC on CUB.

4.3. Results on Object Re-ID

We evaluate our method on four standard Re-ID benchmarks in Table 2 and achieve competitive performance compared to the state-of-the-art methods on both Person Re-ID and Vehicle Re-ID tasks. Particularly, with the ViT-Base backbone, DCAL reaches 80.2%, 64.0%, 87.5%, 80.1% mAP on VeRi-776, MSMT17, Market1501, DukeMTMC, respectively. Similar to FGVC, our method can consistently improve different vision Transformer baselines, e.g., surpassing the light-weight Transformer (DeiT-Tiny) by

Method	VeRi-776		MSMT17		Market1501		DukeMTMC	
	mAP (%)	R1 (%)	mAP (%)	R1 (%)	mAP (%)	R1 (%)	mAP (%)	R1 (%)
SPReID [26]	-	-	-	-	83.4	93.7	73.3	86.0
PCB [43]	-	-	-	-	81.6	93.8	69.2	83.3
MGN [49]	-	-	52.1	76.9	86.9	95.7	78.4	88.7
SAN [25]	72.5	93.3	55.7	79.2	88.0	96.1	75.7	87.9
ABDNet [6]	-	-	60.8	82.3	88.3	95.6	78.6	89.0
HOReID [48]	-	-	-	-	84.9	94.2	75.6	86.9
ISP [66]	-	-	-	-	88.6	95.3	80.0	89.6
STNReID [33]	-	-	-	-	84.9	93.8	-	-
CDNet [28]	-	-	54.7	78.9	86.0	95.1	76.8	88.6
FIDI [55]	77.6	95.7	-	-	86.8	94.5	77.5	88.1
SPAN [7]	68.9	94.0	-	-	-	-	-	-
PVEN [36]	79.5	95.6	-	-	-	-	-	-
CAL (ResNet50) [38]	74.3	95.4	56.2	79.5	87.0	94.5	76.4	87.2
DRL-Net [24]	-	-	55.3	78.4	86.9	94.7	76.6	88.1
AAformer [67]	-	-	63.2	83.6	87.7	95.4	80.0	90.1
TransReID* (ViT-Base) [18]	79.2	96.9	63.6	82.5	-	-	-	-
DeiT-Tiny	71.3	94.3	42.1	63.9	77.9	90.3	69.5	82.9
DeiT-Tiny + DCAL (Ours)	74.1	94.7	44.9	68.2	79.8	91.8	71.7	84.9
DeiT-Small	76.7	95.5	53.3	75.0	84.3	93.7	75.7	87.6
DeiT-Small + DCAL (Ours)	78.1	95.9	55.1	77.3	85.3	94.0	77.4	87.9
DeiT-Base	78.3	95.9	60.5	81.6	86.6	94.4	79.1	88.7
DeiT-Base + DCAL (Ours)	80.0	96.5	62.3	83.1	87.2	94.5	80.2	89.6
ViT-Base	78.1	96.0	61.6	81.4	87.1	94.3	78.9	89.4
ViT-Base + DCAL (Ours)	80.2	96.9	64.0	83.1	87.5	94.7	80.1	89.0

Table 2. Performance comparisons on four Re-ID benchmarks: VeRi-776, MSMT17, Market1501, DukeMTMC. The input size is 256×128 for pedestrian datasets and 256×256 for vehicle datasets. * means results without side information for fair comparison.

2.8% and the larger Transformer (ViT-Base) by 2.4% on MSMT17.

Comparisons to Transformer-based Methods. Our method performs on par with the recent Transformer variants on Re-ID: DRL-Net [24], AAformer [67], TransReID [18]. DRL-Net [24] imposes decorrelation constraints on Transformer decoder to disentangle ID relevant and irrelevant features, while we only employ Transformer encoder and extend self-attention to cross-attention. Both of existing methods (TransReID [18], AAformer [67]) and our methods incorporate local information for recognition but adopt different manners. TransReID [18] designs a jigsaw patch module to shuffle the patch embeddings for learning robust features. AAformer [67] computes the attention between a part token and its associated subset of patch embeddings by online clustering. Differently, we proposes global-local cross-attention to enhance the interactions between global images and local regions.

Comparisons to CNN-based Methods. (1) Many prior approaches have been presented to encode discriminative part-level features for recognition. Typical part-based ReID methods include SPReID [26] and PCB [43]. SPReID [26] utilizes a parsing model to generate human part masks to compute reliable part representations, which consumes extra computation overhead in segmentation part. PCB [43] utilizes a refined part pooling to retrieve the body part information. Our method does not aim to mine precise object parts but establish the interactions between global images and high-response local regions. (2) Image pairs or triplets are widely used in Re-ID for metric learning. Recent Re-ID methods also introduce pair-wise spatial transformer to match the holistic and partial image pairs [33] or design pair-wise loss to learn fine-grained features for recognition [55]. Our pair-wise cross-attention is a new practice in Re-ID in contrast to previous work.

4.4. Ablation Study

Contributions from Algorithmic Components. We examine the contributions from the two types of cross-attention modules using different vision Transformer baselines in Table 3. We use DeiT-Tiny for FGVC and ViT-Base for Re-ID. With either GLCA or PWCA alone, our method can obtain higher performance than the baselines. With both cross-attention modules, we can further improve the results. We note that PWCA will be removed for inference so that it does not introduce extra parameters or FLOPs. We

Method	CUB-200-2011			VeRi-776				MSMT17			
	Params	FLOPs	Acc	Params	FLOPs	mAP	R1	Params	FLOPs	mAP	R1
Baseline	5.5M	8.6G	82.1	81.6M	41.1G	78.1	96.0	81.6M	20.5G	61.6	81.4
+ GLCA	6.0M	8.8G	83.1	88.4M	42.4G	79.5	96.5	88.4M	21.3G	63.7	83.0
+ PWCA	5.5M	8.6G	83.1	81.6M	41.1G	79.2	96.5	81.6M	20.5G	62.8	82.3
Ours	6.0M	8.8G	84.6	88.4M	42.4G	80.2	96.9	88.4M	21.3G	64.0	83.1

Table 3. Effect of the proposed two types of cross-attention learning on CUB-200-2011, VeRi-776 and MSMT17. We use DeiT-Tiny for CUB, ViT-Base for VeRi-776 and MSMT17 as baselines in this ablation experiment.

Method	CUB Acc	MSMT17 mAP
Baseline	82.1	61.6
+ PWCA	83.1	62.8
+ Adding noise in I_1	77.3	56.0
+ Adding noise in label of I_1	81.6	60.8
+ I_2 from noise	82.1	62.1
+ I_2 from COCO	82.5	62.2
+ I_2 from intra-class only	81.7	62.2
+ I_2 from inter-class only	83.0	62.7
+ I_2 from intra- & inter-class (1:1)	83.0	62.5

Table 4. Comparisons of different regularization methods. DeiT-Tiny is used for CUB and ViT-Base is used for MSMT17.

uses one GLCA module in our method, which only requires a small increase of parameters or FLOPs compared to the baseline.

Ablation Study on GLCA. (1) Cross-ViT [3] is a most recent method based on cross-attention for general image classification. It constructs two Transformer branches to handle image tokens of different sizes and uses the class token from one branch to interact with patch tokens from another branch. We implement this idea using the same selected local queries and the same DeiT-Tiny backbone. The cross-token strategy obtains 82.1% accuracy on CUB, which is worse than our GLCA by 1%. (2) Another possible baseline to incorporate local information is computing the self-attention for the high-response local regions (i.e., local query, key and value vectors). This local self-attention baseline obtains 82.6% accuracy on CUB using the DeiT-Tiny backbone, which is also worse than our GLCA (83.1%). (3) We conduct more ablation experiments to examine the effect of GLCA. We obtain 82.6% accuracy on CUB by selecting local query randomly and obtain 82.8% by selecting local query based on the penultimate layer only. Our GLCA outperforms both baselines, validating that mining high-response local query with aggregated attention map is effective for our cross-attention learning.

Ablation Study on PWCA. We compare PWCA with different regularization strategies in Table 4 by taking I_1 as the target image. The results show that adding image noise or label noise without cross-attention causes degraded performance compared to the self-attention learning baseline. As the extra image I_2 used in PWCA can be viewed as distractor, we also test replacing the key and value embeddings of I_2 with Gaussian noise. Such method performs better than adding image / label noise, but still worse than our method. Moreover, sampling I_2 from a different dataset (i.e., COCO), sampling intra-class / inter-class pair only, or sampling intra-class & inter-class pairs with equal probability performs worse than PWCA. We assume that the randomly sampled image pairs from the same dataset (i.e., natural distribution of the dataset) can regularize our cross-attention learning well.

Amount of Cross-Attention Blocks. Figure 2 presents the ablation experiments on the amount of our cross-attention blocks using DeiT-Tiny for CUB and ViT-Base for MSMT17. For GLCA, the results show that $M = 1$ performs best. We analyze that the deeper Transformer encoder can produce more accurate accumulated attention scores as the attention flow is propagated from the input layer to higher layer. Moreover, using one GLCA block only introduces small extra Parameters and FLOPs for inference. For PWCA, the results show that $T = 12$ performs best. It implies that adding I_2 throughout all the encoders can sufficiently regularize the network as our self-attention baseline has $L = 12$ blocks in total. Note that PWCA is only used for training and will be removed for inference without consuming extra computation cost.

4.5. Qualitative Analysis

Figure 3 (a) and Figure 4 (a) visualize the generated attention map using [1] and the selected high-response patches. We observe that self-attention tend to highlight the most discriminative regions in the image. Thanks to GLCA, our method can reduce misleading attention and encourage the network to discover more discriminative clues for recognition.

Figure 3 (b) and Figure 4 (b) visualize the generated attention map using [1] for self-attention and PWCA. We

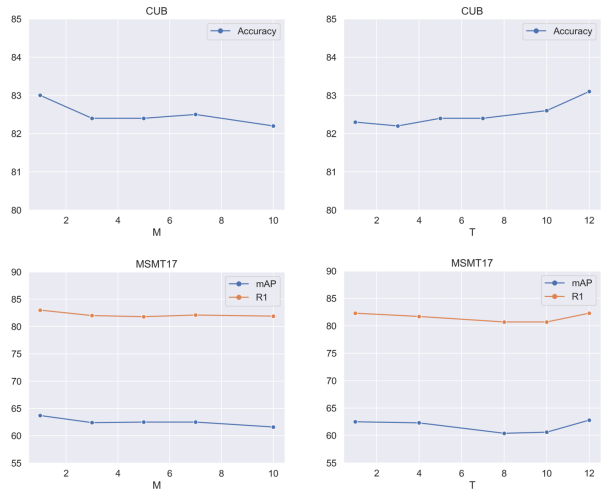
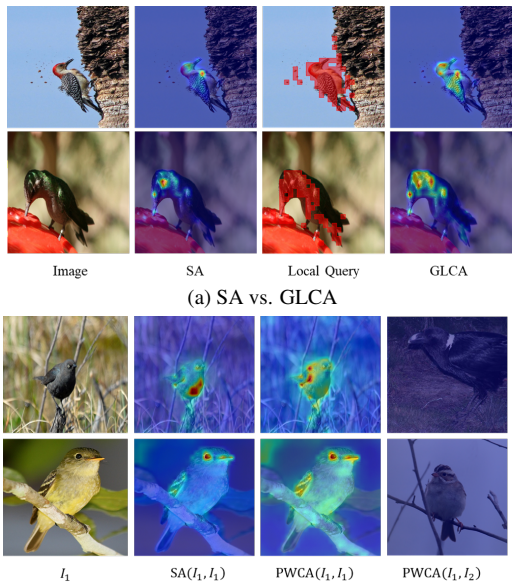


Figure 2. Effect on the amount of cross-attention blocks. DeiT-Tiny is used for CUB and ViT-base ise used for MSMT17. For all the backbones and all the datasets, we build the same $M = 1$ GLCA block and same $T = 12$ PWCA blocks in our method.

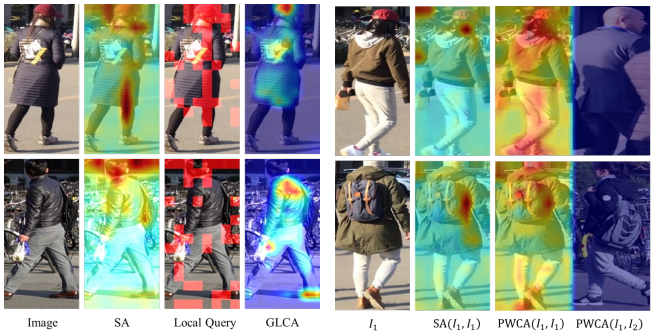


(a) SA vs. GLCA

(b) SA vs. PWCA

Figure 3. Visualization of the generated attention map for self-attention learning and our cross-attention learning on CUB.

observe that PWCA can diffuse the attention responses to explore more complementary parts of objects compared to self-attention. We also visualize the attention map on the distractor image and the blue gauze on it indicates that little attention is derived. It is accordance with our expectation that the attention weights will dominate on the target image as we compute the cross-attention between the query of target image and the combined key-value vectors (Eq. 4).



(a) SA vs. GLCA

(b) SA vs. PWCA

Figure 4. Visualization of the generated attention map for self-attention learning and our cross-attention learning on MSMT17.

4.6. Limitations

Compared to the self-attention learning baseline, our method may take longer time for network convergence as we perform joint training of self-attention and the proposed two types of cross-attention. For example, the self-attention baseline costs 2.1 hours while our method costs 3.8 hours for training on CUB with the same DeiT-backbone and same epochs of 100. However, it is noted that fine-grained recognition datasets are much smaller than the large-scale image classification benchmark and thereby our training time in practice is still acceptable.

Another limitation is that GLCA will increase small computation cost compared to the self-attention baseline. For example, Table 3 shows that GLCA increases 9% Params and 2% FLOPs for DeiT-Tiny on CUB and increases 8% Params and 3% FLOPs for ViT-Base on VeRi-776. We also test removing both GLCA and PWCA blocks for maintaining the same computation cost with the self-attention baseline, and the performance slightly drops, e.g, 84.3% vs. 84.6% (Ours) accuracy on CUB and 80.1% vs. 80.2% (Ours) mAP on VeRi-776.

5. Conclusion

In this work, we introduce two types of cross-attention mechanisms to better learn subtle feature embeddings for recognizing fine-grained objects. GLCA can help reinforce the spatial-wise discriminative clues by modeling the interactions between global images and local regions. PWCA can establish the interactions between image pairs and can be viewed as a regularization strategy to alleviate overfitting. Our cross-attention design is easy-to-implement and compatible to different vision Transformer baselines. Extensive experiments on seven benchmarks have demonstrated the effectiveness of our method on FGVC and Re-ID tasks. We expect that our method can inspire new insights for the self-attention learning regime in Transformer.

References

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*, 2020. 3, 7

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 1, 2

[3] Chun-Fu Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. *arXiv preprint arXiv:2103.14899*, 2021. 7

[4] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *CVPR*, 2021. 2

[5] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *ICML*, 2020. 1, 2

[6] Tianlong Chen, Shaojin Ding, Jingyi Xie, Ye Yuan, Wuyang Chen, Yang Yang, Zhou Ren, and Zhangyang Wang. Abd-net: Attentive but diverse person re-identification. In *ICCV*, pages 8351–8361, 2019. 6

[7] Tsai-Shien Chen, Chih-Ting Liu, Chih-Wei Wu, and Shao-Yi Chien. Orientation-aware vehicle re-identification with semantics-guided part attention network. In *ECCV*, pages 330–346. Springer, 2020. 6

[8] Yue Chen, Yalong Bai, Wei Zhang, and Tao Mei. Destruction and construction learning for fine-grained image recognition. In *CVPR*, 2019. 5

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 4

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1, 2

[11] Yao Ding, Yanzhao Zhou, Yi Zhu, Qixiang Ye, and Jianbin Jiao. Selective sparse sampling for fine-grained image recognition. In *ICCV*, 2019. 1, 2, 5

[12] 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. 1, 2

[13] Ruoyi Du, Dongliang Chang, Ayan Kumar Bhunia, Jiyang Xie, Zhanyu Ma, Yi-Zhe Song, and Jun Guo. Fine-grained visual classification via progressive multi-granularity training of jigsaw patches. In *ECCV*, 2020. 2, 5

[14] Abhimanyu Dubey, Otkrist Gupta, Pei Guo, Ramesh Raskar, Ryan Farrell, and Nikhil Naik. Pairwise confusion for fine-grained visual classification. In *ECCV*, 2018. 5

[15] Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *CVPR*, 2017. 1, 5

[16] Yu Gao, Xintong Han, Xun Wang, Weilin Huang, and Matthew Scott. Channel interaction networks for fine-grained image categorization. In *AAAI*, 2020. 1, 2, 5

[17] 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. 2, 5

[18] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. *arXiv preprint arXiv:2102.04378*, 2021. 6

[19] Han Hu, Zheng Zhang, Zhenda Xie, and Stephen Lin. Local relation networks for image recognition. In *ICCV*, 2019. 1, 2

[20] Yunqing Hu, Xuan Jin, Yin Zhang, Haiwen Hong, Jingfeng Zhang, Yuan He, and Hui Xue. Rams-trans: Recurrent attention multi-scale transformer for fine-grained image recognition. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4239–4248, 2021. 5

[21] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *European conference on computer vision*, pages 646–661. Springer, 2016. 4

[22] Shaoli Huang, Xinchao Wang, and Dacheng Tao. Stochastic partial swap: Enhanced model generalization and interpretability for fine-grained recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 620–629, 2021. 5

[23] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *ICCV*, 2019. 2

[24] Mengxi Jia, Xinhua Cheng, Shijian Lu, and Jian Zhang. Learning disentangled representation implicitly via transformer for occluded person re-identification. *arXiv preprint arXiv:2107.02380*, 2021. 2, 6

[25] Xin Jin, Cuiling Lan, Wenjun Zeng, Guoqiang Wei, and Zhibo Chen. Semantics-aligned representation learning for person re-identification. In *AAAI*, volume 34, pages 11173–11180, 2020. 6

[26] Mahdi M Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In *CVPR*, 2018. 6

[27] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV workshop*, 2013. 4

[28] Hanjun Li, Gaojie Wu, and Wei-Shi Zheng. Combined depth space based architecture search for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6729–6738, 2021. 6

[29] Chuanbin Liu, Hongtao Xie, Zheng-Jun Zha, Lingfeng Ma, Lingyun Yu, and Yongdong Zhang. Filtration and distillation: Enhancing region attention for fine-grained visual categorization. In *AAAI*, 2020. 5

[30] Hao Liu, Jiashi Feng, Meibin Qi, Jianguo Jiang, and Shuicheng Yan. End-to-end comparative attention networks for person re-identification. *TIP*, 26(7):3492–3506, 2017. 2

[31] Xinchen Liu, Wu Liu, Jinkai Zheng, Chenggang Yan, and Tao Mei. Beyond the parts: Learning multi-view cross-part

correlation for vehicle re-identification. In *ACM MM*, pages 907–915, 2020. 2

[32] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *CVPR Workshop*, 2019. 2

[33] Hao Luo, Wei Jiang, Xing Fan, and Chi Zhang. Stnreid: Deep convolutional networks with pairwise spatial transformer networks for partial person re-identification. *TMM*, 22(11):2905–2913, 2020. 6

[34] Wei Luo, Xitong Yang, Xianjie Mo, Yuheng Lu, Larry S Davis, Jun Li, Jian Yang, and Ser-Nam Lim. Cross-x learning for fine-grained visual categorization. In *ICCV*, 2019. 1, 2, 5

[35] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 4

[36] Dechao Meng, Liang Li, Xuejing Liu, Yadong Li, Shijie Yang, Zheng-Jun Zha, Xingyu Gao, Shuhui Wang, and Qingming Huang. Parsing-based view-aware embedding network for vehicle re-identification. In *CVPR*, pages 7103–7112, 2020. 6

[37] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jonathon Shlens. Stand-alone self-attention in vision models. *arXiv preprint arXiv:1906.05909*, 2019. 1, 2

[38] Yongming Rao, Guangyi Chen, Jiwen Lu, and Jie Zhou. Counterfactual attention learning for fine-grained visual categorization and re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1025–1034, 2021. 5, 6

[39] Aravind Srinivas, Tsung-Yi Lin, Niki Parmar, Jonathon Shlens, Pieter Abbeel, and Ashish Vaswani. Bottleneck transformers for visual recognition. In *CVPR*, 2021. 2

[40] Ming Sun, Yuchen Yuan, Feng Zhou, and Errui Ding. Multi-attention multi-class constraint for fine-grained image recognition. In *ECCV*, 2018. 1, 5

[41] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, et al. Sparse r-cnn: End-to-end object detection with learnable proposals. In *CVPR*, 2021. 2

[42] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, 2018. 2

[43] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, pages 480–496, 2018. 6

[44] Zhiqing Sun, Shengcao Cao, Yiming Yang, and Kris Kitani. Rethinking transformer-based set prediction for object detection. *arXiv preprint arXiv:2011.10881*, 2020. 2

[45] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877*, 2020. 1, 2

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. 1, 2

[47] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 4

[48] Guan'an Wang, Shuo Yang, Huanyu Liu, Zhicheng Wang, Yang Yang, Shuliang Wang, Gang Yu, Erjin Zhou, and Jian Sun. High-order information matters: Learning relation and topology for occluded person re-identification. In *CVPR*, pages 6449–6458, 2020. 6

[49] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *ACM MM*, pages 274–282, 2018. 2, 6

[50] Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Max-deeplab: End-to-end panoptic segmentation with mask transformers. In *CVPR*, 2021. 2

[51] Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In *ECCV*, 2020. 1, 2

[52] Jun Wang, Xiaohan Yu, and Yongsheng Gao. Feature fusion vision transformer for fine-grained visual categorization. *arXiv preprint arXiv:2107.02341*, 2021. 2, 5

[53] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, 2018. 4

[54] Lin Wu, Yang Wang, Junbin Gao, Meng Wang, Zheng-Jun Zha, and Dacheng Tao. Deep coattention-based comparator for relative representation learning in person re-identification. *IEEE T NEUR NET LEAR*, 32(2):722–735, 2020. 4

[55] Cheng Yan, Guansong Pang, Xiao Bai, Changhong Liu, Ning Xin, Lin Gu, and Jun Zhou. Beyond triplet loss: person re-identification with fine-grained difference-aware pairwise loss. *TMM*, 2021. 6

[56] Ze Yang, Tiange Luo, Dong Wang, Zhiqiang Hu, Jun Gao, and Liwei Wang. Learning to navigate for fine-grained classification. In *ECCV*, 2018. 1, 5

[57] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. *arXiv preprint arXiv:2101.11986*, 2021. 2

[58] Lianbo Zhang, Shaoli Huang, Wei Liu, and Dacheng Tao. Learning a mixture of granularity-specific experts for fine-grained categorization. In *ICCV*, 2019. 1, 2, 5

[59] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *IJCV*, pages 1–19, 2021. 4

[60] Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo. Learning multi-attention convolutional neural network for fine-grained image recognition. In *ICCV*, 2017. 1, 5

[61] Heliang Zheng, Jianlong Fu, Zheng-Jun Zha, and Jiebo Luo. Looking for the devil in the details: Learning trilinear atten-

tion sampling network for fine-grained image recognition. In *CVPR*, 2019. 1, 2, 5

[62] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, pages 1116–1124, 2015. 4

[63] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *CVPR*, 2021. 2

[64] Zhedong Zheng, Tao Ruan, Yunchao Wei, Yi Yang, and Tao Mei. Vehiclenet: Learning robust visual representation for vehicle re-identification. *TMM*, 2020. 4

[65] Mohan Zhou, Yalong Bai, Wei Zhang, Tiejun Zhao, and Tao Mei. Look-into-object: Self-supervised structure modeling for object recognition. In *CVPR*, 2020. 5

[66] Kuan Zhu, Haiyun Guo, Zhiwei Liu, Ming Tang, and Jinqiao Wang. Identity-guided human semantic parsing for person re-identification. In *ECCV*, pages 346–363. Springer, 2020. 6

[67] Kuan Zhu, Haiyun Guo, Shiliang Zhang, Yaowei Wang, Gaopan Huang, Honglin Qiao, Jing Liu, Jinqiao Wang, and Ming Tang. Aaformer: Auto-aligned transformer for person re-identification. *arXiv preprint arXiv:2104.00921*, 2021. 2, 6

[68] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*, 2021. 1, 2

[69] Peiqin Zhuang, Yali Wang, and Yu Qiao. Learning attentive pairwise interaction for fine-grained classification. In *AAAI*, 2020. 1, 2, 5