Visual Attention Consistency under Image Transforms for Multi-Label Image Classification

Hao Guo[‡], Kang Zheng[‡], Xiaochuan Fan[‡], Hongkai Yu[♯], Song Wang[†,‡,*]
[†]Tianjin University, [‡]University of South Carolina, [♯]University of Texas - Rio Grande Valley
{hguo, zheng37}@email.sc.edu, efan3000@gmail.com, hongkai.yu@utrgv.edu, songwang@cec.sc.edu

Abstract

Human visual perception shows good consistency for many multi-label image classification tasks under certain spatial transforms, such as scaling, rotation, flipping and translation. This has motivated the data augmentation strategy widely used in CNN classifier training – transformed images are included for training by assuming the same class labels as their original images. In this paper, we further propose the assumption of perceptual consistency of visual attention regions for classification under such transforms, i.e., the attention region for a classification follows the same transform if the input image is spatially transformed. While the attention regions of CNN classifiers can be derived as an attention heatmap in middle layers of the network, we find that their consistency under many transforms are not preserved. To address this problem, we propose a two-branch network with an original image and its transformed image as inputs and introduce a new attention consistency loss that measures the attention heatmap consistency between two branches. This new loss is then combined with multi-label image classification loss for network training. Experiments on three datasets verify the superiority of the proposed network by achieving new state-of-the-art classification performance.

1. Introduction

As an important computer vision task, multi-label image classification [51, 60] aims to tell whether an image contains certain attributes, objects, etc., each of which is denoted by a label. Typical applications of multi-label image classification include human attribute recognition [1, 10, 18, 31, 34, 61], scene understanding [45], multi-object recognition [6], facial attribute recognition [19], etc. While recent progress on deep neural networks has improved the performance of multi-label image classification significantly, it is still a very challenging problem due to appearance complexities, intra-label variation, and unsatisfactory image qualities [17, 55, 3, 33, 65, 21, 32, 53, 67] .

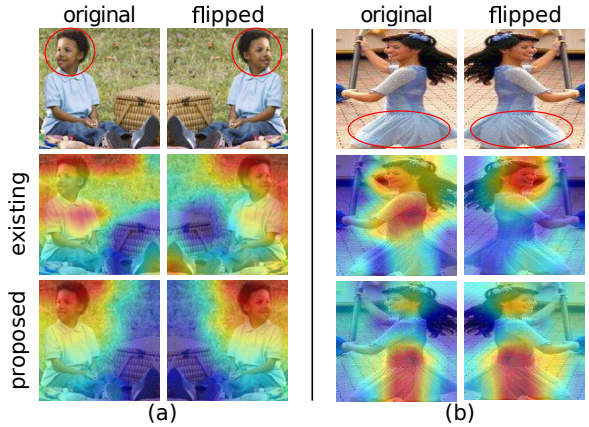

Figure 1. An illustration of attention heatmaps for classifying labels (a) "face mask" and (b) "jeans" from original and horizontally flipped images, using an existing CNN (the middle row) and the proposed method (the bottom row).

Human visual perception is consistent for many multi-label image classification tasks under certain spatial transforms, such as scaling, rotation, flipping and translation. For example, these transforms usually do not vary human recognition of "sunglasses" in an image. This consistency has motivated the *data augmentation* strategy [26], which has been widely used in training CNN classifiers – for each original image with ground-truth labels, we can transform the image to construct a new training image by assigning the same ground-truth labels. Data augmentation reduces overfitting problem of training CNN models for classification tasks with perceptual consistency under spatial transforms.

The perceptual consistency assumed in data augmentation is a high-level representation, at the stage of final classification. Actually, image classification is usually only relevant to certain attention regions for both human vision (according to studies on human cognitive [39, 27] and neuroscience [11]) and CNN models [66]. In this paper, we further assume a new perceptual consistency of visual attention, a middle-level representation, under the above spatial

*Corresponding author.

transforms in multi-label image classification and incorporate it into CNNs for enhancing classifier training. Here, we define ***visual attention consistency*** as: *attention regions for image classification follow the same transform if the image is spatially transformed*. As shown in the top row of Fig. 1, when the image is flipped horizontally, the human attention regions are also flipped horizontally to keep focusing on the face and leg regions to tell the presence for labels of "face mask" and "jeans", respectively. The proposed visual attention consistency can be considered as a visual property of "equivariance" [28]. Different from other studies of equivariance [29, 7, 50, 49, 38, 43, 57, 56, 13], this paper enforces equivariance at a specific level of CNN attention.

Previous works on CNN classifiers have shown that the attention regions can be derived as attention heatmaps in middle layers of the network, with only image-level supervision [66, 41] and can be used to re-weight the extracted image features to enhance CNN-based image classification [47, 24, 52, 58, 22]. However, we find that current CNN classifiers do not preserve the attention consistency under many of the above spatial transforms, even if training images are augmented by these transforms. As shown in the middle row of Fig. 1, the attention regions (in red) are inconsistent under the horizontal flipping transform using ResNet50 with training data augmented by flipping. Besides, the CNN attention may also cover regions irrelevant to the label "face mask" and the label "jeans", respectively. Therefore, we expect better visual perceptual plausibility and better multi-label image classification by considering visual attention consistency under spatial transforms.

For this purpose, we propose a new network with two identical branches taking the original and transformed images as two inputs. The output of each branch is the label predictions of the input image. In the middle of each branch, we use Class Activation Mapping (CAM) [66] to compute the attention heatmaps for each label on the corresponding input image. Then, we define a new attention consistency loss as a distance between the **transformed attention heatmaps of the original image** and the **attention heatmaps of the transformed image**. This loss is then combined with multi-label image classification loss for network training to improve the visual attention consistency under image transforms. As illustrated in the bottom row of Fig. 1, attention regions of the proposed network for both "face mask" and "jeans" become more consistent under image flipping. Meanwhile, these attention regions are more label-relevant, focusing on face regions for "face mask" and leg regions for "jeans".

We evaluate the proposed method for different multi-label image classification tasks on three datasets: WIDER Attribute [34], MS-COCO [35], and PA-100K [36]. The experiments show that our method achieves state-of-the-art performances on these datasets. We also conduct ablative study to verify the significant performance gains by incorporating the proposed new attention consistency.

2. Related work

2.1. Multi-label image classification

As reviewed in [51, 60], multi-label classification problem has been widely explored, with progress on both label-separate and label-correlated methods. Label-separate methods use binary relevance strategy [2] to convert multi-label image classification to multiple binary image classification problem. With great success of using CNNs [26, 46, 20, 23] for single-label image classification [9], multi-label image classification has been improved significantly. Besides, deep convolutional ranking [17] optimizes top-k ranking loss on convolutional architectures to learn a better feature representation. Hypotheses-CNN-Pooling [55] aggregates object segmentation hypotheses with max pooling to generate multi-label predictions.

Much progress has been made on label-correlated multi-label image classification in recent years. Many methods, such as matrix completion [3], probabilistic label enhancement [33], RGNN [65], SINN [21], Conditional Graphical Lasso [32], and CNN-RNN [53] are proposed to model the semantic correlations between labels for multi-label image classification. Furthermore, Spatial Regularization Network [67] captures both semantic and spatial correlations between labels. Label balancing [19] is also used for improving multi-label image classification.

In this paper, we propose to enforce consistency of attention regions under certain image transforms to improve multi-label image classification, which provides a new perspective to improve the visual perception plausibility of the CNNs for promoting the classification performance.

2.2. Attention mechanism for classification

As an intermediate result, attention of CNNs has been used for various computer vision tasks [63, 58, 24, 47, 52, 22, 40, 5, 4, 54, 12, 62, 25, 44, 14]. For the task of image classification, attention of CNNs reflects the image regions that CNNs use as the evidence to classify images [41, 66, 59]. For the multi-label image classification task, SRN [67] learns attention heatmap to specify spatial relations between labels. But the relevance of attention regions to each label is not considered by SRN. To address this problem, one straightforward idea is to learn accurate attention regions similar to semantic segmentation [37] and saliency detection [64], which requries infeasible pixel-level annotations. One potential solution to reduce annotation labors is eye-tracking [42], which is somewhat noisy and inconsistent from different observers, due to not well defined label-relevant regions. The attention heatmap can also be refined [18] by driving it to concentrate on a single

compact region instead of many fragmented regions, which, however, is not applicable to labels with multiple relevant regions in images. In this paper, we propose an indirect way to focus attention of CNNs on regions more label-relevant by enforcing consistency of attention regions under certain image transforms.

3. Proposed Method

In this section, we first describe the background for the proposed network and then elaborate on the proposed two-branch network. We construct a set of spatial transforms under which the visual attention is consistent and embed them into the proposed network.

3.1. Background

3.1.1 Class activation mapping

Because of its simplicity and capability of visualizing attention regions for classification, we apply class activation mapping (CAM) [66] to extract attention heatmaps. Typical CNN architectures, such as ResNet [20], DenseNet [23], and Inception [48], all start with convolutional layers. A global average pooling (GAP) is then performed on the feature maps $\boldsymbol{F} \in \mathbb{R}^{C\times H\times W}$ from the last convolutional layer, where C, H, W are the number of channels, height, and width of the feature maps, respectively. The pooled features are further fed into the final output layer, a fully connected (FC) layer with weights $\boldsymbol{W} \in \mathbb{R}^{L\times C}$ (L is the number of labels), for classification. CAM computes the attention heatmaps by linearly weighted sum of all channels:

$$\boldsymbol{M}_j(m,n) = \sum_{k=1}^{C} \boldsymbol{W}(j,k)\boldsymbol{F}_k(m,n), \tag{1}$$

where $\boldsymbol{M}_j(m,n)$ indicates the attention heatmap at spatial location (m,n) for label j, $\boldsymbol{W}(j,k)$ represents the weight corresponding to label j for channel k of feature maps, $\boldsymbol{F}_k(m,n)$ represents the feature maps of channel k from the last convolutional layer at spatial location (m,n). In the following, we use $\boldsymbol{M} = g(\boldsymbol{I})$ to represent the attention heatmaps of image $\boldsymbol{I}$. Note that the size of attention heatmaps in Eq. (1) is $H \times W$, which is smaller than the input image size. To visualize the attention regions on the image, bilinear interpolation is used to upsample the attention heatmaps to the input image size.

3.1.2 Multi-label image classification loss

Several different loss functions have been used for multi-label image classification in previous works, such as rank loss [8], cross entropy loss [34, 18, 30, 36, 31], etc. Note that multi-label classification is formulated as multiple binary classification problems when using the cross-entropy loss. For simplicity and effectiveness, in this paper we adopt the *weighted sigmoid cross entropy loss* in [30]:

$$\begin{aligned}\ell_c = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{L}\omega_{ij}\Bigg(y_{ij}\log\frac{1}{1+e^{-x_{ij}}} \\ +(1-y_{ij})\log\frac{e^{-x_{ij}}}{1+e^{-x_{ij}}}\Bigg)\end{aligned} \tag{2}$$

$$\omega_{ij} = \begin{cases} e^{1-p_j} & \text{if } y_{ij}=1 \\ e^{p_j} & \text{if } y_{ij}=0 \end{cases}, \tag{3}$$

where N is the number of images, L is the number of labels, $x_{ij} \in \mathbb{R}$ is the predicted presence of label j in image i and it is further normalized to a presence score $1/(1+e^{-x_{ij}}) \in [0,1]$, $y_{ij} \in \{0,1\}$ is the ground truth of the presence of label j in image i, p_j is the proportion of positive samples with label j in the training set and it is used to define the weight ω_{ij} for balancing training samples. This loss function is modified from cross entropy loss and has been used in several prior works on multi-label image classification, such as RAP [31] and HP-Net [36]. In the later experiments, we use this loss for both baselines and the proposed methods for fair comparisons, by excluding the performance difference resulting from the use of different loss functions.

3.2. Proposed network

In general, the plausibility of attention heatmaps can reflect the performance of the CNN classifier – if the attention heatmaps highlight the regions that are semantically relevant to the considered labels, we can expect better CNN classification performance. Two examples are shown in Fig. 2. With the increase of the training iterations, the predicted presence score increases (decreases) when the attention heatmaps highlight the desired relevant regions for positive (negative) samples. This suggests that "good" attention regions usually result in "good" classification results.

One straightforward approach to improve the plausibility of attention heatmaps is to impose explicit supervision of label-relevant regions in CNN training. However, it is highly laborious to accurately annotate label-relevant regions on a large set of training images. Besides, label-relevant regions may not be well defined: different annotators may not have an agreement on the relevant regions for some labels, such as "Age between 18 and 60" in an image. In this paper, we propose to improve CNN's capability to focus attention on label-relevant regions in an indirect way, i.e., enforcing CNN attention to be consistent under certain image transforms. In the following, we first introduce the proposed network for visual attention consistency and the considered image transforms will be discussed in detial in Section 3.3.

Figure 2. Attention heatmaps for label "sunglasses" in different iterations of CNN (ResNet50) model training, where face is the desired label-relevant region. The number above each attention heatmap represents the predicted presence score (in $[0, 1]$) in the corresponding iteration.

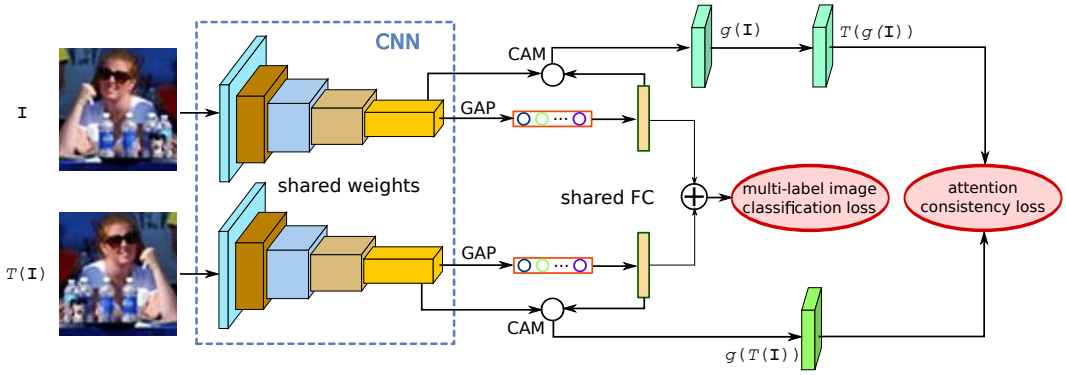

Figure 3. An illustration of the proposed two-branch network.

As shown in Fig. 3, the proposed network consists of two identical branches. Each branch starts with convolutional layers and ends with GAP-FC (fully connected layer after global average pooling) structure (e.g., ResNet, DenseNet). The parameters of two branches are shared. The two branches take an image $\boldsymbol{I}$ and the transformed image $\boldsymbol{I'} = T(\boldsymbol{I})$ as inputs, respectively. Feature maps from the last convolutional layer of two branches are $\boldsymbol{F}, \boldsymbol{F'} \in \mathbb{R}^{C\times H\times W}$, respectively. The corresponding spatial averages of feature maps after GAP are used for multi-label image classification by final fully connected layer (FC) with weights $\boldsymbol{W} \in \mathbb{R}^{L\times C}$. Meanwhile, the attention heatmaps for each input and each label are extracted by CAM. Specifically, by expanding feature maps $\boldsymbol{F}$, $\boldsymbol{F'}$ into shape of $1 \times C \times H \times W$, and FC weights $\boldsymbol{W}$ into shape of $L \times C \times 1 \times 1$, we conduct channel-wise multiplication to linearly combine feature maps for each label, and sum along dimension C of combined feature maps, as in Eq. (1). The resulting attention heatmaps $\boldsymbol{M} = g(\boldsymbol{I})$ and $\boldsymbol{M'} = g(T(\boldsymbol{I}))$, where $g(\cdot)$ represents the process of computing attention heatmap with CAM, are both in shape of $L \times H \times W$.

Based on our definition of attention consistency, the attention heatmaps, $g(\boldsymbol{I})$ and $g(T(\boldsymbol{I}))$, for the original and the transformed images, respectively, need to be equivariant [28] under the particular image transform, which can be formulated as:

$$T(g(\boldsymbol{I})) = g(T(\boldsymbol{I})). \tag{4}$$

Therefore, to enforce the attention consistency, we define an attention consistency loss using the mean square difference between the transformed heatmaps $\hat{\boldsymbol{M}} = T(\boldsymbol{M}) = T(g(\boldsymbol{I}))$ of the original image and the heatmaps of the transformed image $\boldsymbol{M'}$, i.e.,

$$\ell_a = \frac{1}{NLHW}\sum_{i=1}^{N}\sum_{j=1}^{L}\|\hat{\boldsymbol{M}}_{ij} - \boldsymbol{M'}_{ij}\|_2, \tag{5}$$

where $\boldsymbol{M}_{ij}$ indicates the attention heatmap for image i and label j. We linearly combine the multi-label image classification loss in Eq. (2) and the attention consistency loss in Eq. (5) to train the network:

$$\ell = \ell_c + \lambda\ell_a, \tag{6}$$

where λ is a hyper-parameter for balancing the two losses.

For testing, we only use one branch for multi-label image classification, since the network parameters are shared by each branch. The outputs from the last FC layer indicate the confidence values of the presence of each label. A sigmoid function is used to normalize the values as presence scores of each label in range $[0, 1]$. If the presence score is greater than 0.5, the label is predicted as being present.

3.3. Image transforms

Different spatial transforms can be considered in the proposed visual attention consistency, only if they do not change the human visual perception of the image, i.e., the presence of the class labels. We denote the set of such image transforms as $\mathbb{U}$ and any transform in this set can be embedded in the proposed two-branch network for enhancing multi-label image classification. Specifically, we focus on a subset of frequently used transforms

$\{translation, rotation, flipping, scaling\} \subset \mathbb{U}$ in this paper to justify the effectiveness of the proposed network. It is quite intuitive that human visual perception of many class labels keeps unchanged when the input image undergoes the translation, rotation, flipping and/or scaling transforms. Certainly, in some extreme cases of these transforms, e.g., down-scaling the input image to a very small size, the visual perception of the image may totally change. In this paper, we choose appropriate parameters for these transforms to avoid such extreme cases.

Difference from data augmentation: We can notice that the above subset of transforms in $\mathbb{U}$ can also be used for data augmentation in training CNN classifiers. However, data augmentation considers the classification consistency under these transforms which is imposed on the final output, a high-level representation, of the network, while the proposed method enforces the attention consistency under the same transforms on the intermediate result, a middle-level representation, of the network. In general, enforcing classification consistency at the high-level representation has less impact to the network parameters than enforcing attention consistency at middle-level representation. We will show in later experiments that the proposed method can train network with better classification performance than data augmentation.

Attention consistency under certain transforms: Any transform $T \in \mathbb{U}$ can be embeded in the proposed network. Let's take horizontal flipping transform as an example. With $T : \boldsymbol{I} \rightarrow \boldsymbol{I'}$, we have $\boldsymbol{I'}(m, n) = \boldsymbol{I}(W_I - m, n)$, where (m, n) indicates spatial location in images, while W_I represents the width of the original image. After computing the attention heatmaps $\boldsymbol{M} = g(\boldsymbol{I})$ and $\boldsymbol{M'} = g(\boldsymbol{I'}) = g(T(I))$, respectively, the same transform T is applied on $\boldsymbol{M}$ so that $\hat{\boldsymbol{M}}(m, n) = \boldsymbol{M}(W_M - m, n)$, where (m, n) indicates spatial location in attention heatmaps, and W_M represents the width of attention heatmaps. Then Eq. (5) is used to calculate the attention consistency loss. Similar procedures can be applied when attention consistency under image translation, rotation, or scaling is considered in the proposed network.

Besides, since the attention heatmaps computed by CAM (Eq. (1)) are downsized from the input image size, e.g., 7×7 from 224×224 and 6×6 from 192×192, there is a trick for embedding scaling in the proposed network. Suppose the dimensions of attention heatmaps $\boldsymbol{M}$ and $\boldsymbol{M'}$ are $L \times H_M \times W_M$ and $L \times H_{M'} \times W_{M'}$, where $H_M \neq H_{M'}$ and $W_M \neq W_{M'}$. Since H_M and W_M may not be divisible by $H_{M'}$ and $W_{M'}$, respectively, it may be inappropriate to re-scale $H_M \times W_M$ to $H_{M'} \times W_{M'}$ directly. To quantify the inconsistency, we upscale both the attention heatmaps to the same size based on the lowest common multiples in width and height dimension, respectively, e.g., upscaled to height 42 for $H_M = 7$ and $H_{M'} = 6$.

Attention consistency under combined transform: We can also embed more than one transforms in $\mathbb{U}$ to the proposed network. For example, considering two transforms, $T_1, T_2 \in \mathbb{U}$, the attention consistency loss defined in Eq. (5) can be simply calculated by

$$\ell_a = \ell_{a,T_1} + \ell_{a,T_2} \tag{7}$$

4. Experiments

We adopt ResNet [20] as our base architecture to implement the proposed network because of its excellent performance in image-related recognition tasks. The proposed network is fine-tuned from models pre-trained on ImageNet [9] using Stochastic Gradient Descent for optimization, with initial learning rate 10^{-3}. We evaluate it for multi-label image classification on three datasets: WIDER Attribute [34], PA-100K [36], and MS-COCO [35]. **WIDER Attribute** is proposed for human attribute recognition. It contains 13,789 images with 57,524 annotated human bounding boxes. Each human in a bounding box is annotated with 14 human attributes. The train-val set includes 28,345 human bounding boxes, while the test set includes 29,179 human bounding boxes. **PA-100K** is a large-scale pedestrian attribute dataset. It consists of 100,000 pedestrian images, each of which is annotated with 26 human attributes. The training, validation and test sets are split with a ratio of $8 : 1 : 1$. **MS-COCO** is originally collected for object recognition tasks in the context of scene understanding. It is also frequently used for multi-label image classification task. It contains 82,783 images in training set, and 40,504 images in validation set. Each image is annotated with 80 object labels. Since ground-truth labels of test set are not available, we train our network on training set and evaluate on validation set.

Two sets of metrics are introduced in [60] for multi-label image classification evaluation. 1) Label-based metrics include mean Average Precision (mAP), mean accuracy (mA), macro and micro precision/recall/F1-score (denoted as P-C, R-C, F1-C, P-O, R-O, F1-O, respectively). Macro metrics ("*-C") are evaluated by averaging per-label metrics, while micro metrics ("*-O") are overall measures, which count true predictions for all images over all labels, as in [67]. 2) Example-based metrics [31] include Accuracy (Acc), Precision (Prec), Recall, and F1-score.

4.1. Ablative analysis

We first conduct experiments to justify that attention consistency under certain image transforms in the proposed network can benefit multi-label image classification. We conduct two sets of ablative experiments on WIDER

Attribute dataset, with ResNet50 (**R50**) and ResNet101 (**R101**) as the backbones separately for the proposed method. The baseline methods use the original ResNet50 and ResNet101 with only weighted sigmoid cross entropy loss as in Eq. (2). The input images are resized to 224×224. The hyper-parameter in Eq. (6) is set to 1.

Table 1. Performance (%) on WIDER Attribute dataset in terms of label-based metrics. The best results are highlighted in bold font and red color, while the second bests are in blue.

model	mAP	mA	F1-C	P-C	R-C	F1-O	P-O	R-O
R50	83.4	82.0	73.9	79.5	69.4	79.4	82.3	76.6
R50+t	83.7	83.4	74.1	75.6	72.8	79.5	80.6	78.4
R50+r	83.2	82.8	73.2	75.9	71.1	78.5	81.0	76.1
R50+s	83.9	83.0	74.4	77.7	71.7	79.4	81.3	77.6
R50+f	84.2	82.8	74.6	79.5	70.7	80.0	82.9	76.9
R50+ACt	83.9	84.0	74.2	74.5	74.2	79.2	79.7	78.7
R50+ACr	85.0	83.3	75.1	79.2	71.8	80.2	82.3	77.9
R50+ACs	85.6	82.7	75.3	81.9	70.1	80.6	**84.5**	77.1
R50+ACf	86.3	**84.5**	76.4	78.9	**74.3**	81.2	82.6	**79.8**
R50+ACfs	**86.8**	83.7	**76.5**	**82.4**	72.1	**81.8**	84.4	79.3
R101	84.8	83.2	75.5	80.5	71.5	80.6	83.6	77.8
R101+ACt	84.6	83.5	75.3	79.1	71.9	80.1	83.1	77.3
R101+ACr	86.0	84.2	76.2	79.5	73.6	81.2	83.2	79.4
R101+ACs	86.5	83.6	76.5	**82.4**	71.9	81.6	**85.1**	78.3
R101+ACf	87.1	84.7	77.4	80.9	74.5	82.1	83.8	80.5
R101+ACfs	**87.5**	**85.0**	**77.6**	81.3	**74.8**	**82.4**	84.1	**80.7**

For experiments using ResNet50 as backbone, the baseline model is trained from the original ResNet50 without any data augmentation, denoted as ***R50***. For comparison, we futher train model R50 by using certain image transforms as data augmentation. These transforms are 32-pixel translation, 90° rotation, down-scaled to 192×192, and horizontal flipping, from which we get the models with data augmentation as ***R50+t*** (translation), ***R50+r*** (rotation), ***R50+s*** (scaling), ***R50+f*** (flipping), respectively. When using the proposed method that enforces the attention consistency (**AC**) under these four image transforms, we get the trained models: ***R50+ACt***, ***R50+ACr***, ***R50+ACs***, and ***R50+ACf***, respectively. The testing performance of the above models in terms of label-based metrics are shown in the upper part of Table 1.

Comparing models R50+t, R50+r, R50+s, R50+f with model R50 shows that if the transforms are only used as data augmentation, there are only minor performance gains. When the attention consistency is considered in models R50+ACr, R50+ACs, R50+ACf, the performance is improved significantly, e.g., mAP is improved from 83.5% for model R50 to 85.0%, 85.6%, and 86.3% for models R50+ACr, R50+ACs, and R50+ACf, respectively.

To verify that the attention consistency leads to significant improvement over data augmentation with the same transform, we compare models R50+ACr, R50+ACs, R50+ACf with R50+r, R50+s, R50+f and the gains of mAP are 1.8%, 1.7%, 2.1%, respectively. Note that the attention consistency under translation only results in slight performance improvement, e.g., 83.7% mAP for model R50+t against 83.9% for model R50+ACt. This is due to the fact that most CNNs learn invariant representations to image translation by using convolution and pooling operations.

Table 2. The quantified attention inconsistency under flipping and scaling.

models	flip	scale
baseline	93.23	64.34
proposed	2.85	0.74

Table 3. Performance (%) of enforcing consistency at different levels on R50+ACf.

	feature level	attention level	label level
mAP	85.1	86.3	85.4

Furthermore, as attention consistency under image scaling (model R50+ACs) and flipping (model R50+ACf) achieves significant performance gains comparing with baseline model R50 in Table 1, we combine attention consistency under both image scaling and flipping to train model ***R50+ACfs***. The mAP performance is further improved to 86.8%, which is 3.4% higher than that from the original ResNet50.

For experiments using ResNet101 as backbone, models ***R101***, ***R101+ACt***, ***R101+ACr***, ***R101+ACs***, ***R101+ACf***, ***R101+ACfs*** are trained in a similar way as models R50, R50+ACt, R50+ACr, R50+ACs, R50+ACf, and R50+ACfs, respectively. Evaluation results are reported in the lower part of Table 1. Similar performance gains are obtained by the proposed method as shown in the upper part of Table 1. The proposed network using ResNet101 as backbone finally improves the mAP by 2.7% from the original ResNet101.

For in-depth exploration of performance gains of the proposed method, we show average precision (AP) gain for each label achieved by models R50+t, R50+r, R50+s, R50+f, R50+ACt, R50+ACr, R50+ACs, R50+ACf and R50+ACfs compared with the baseline model R50 in Fig. 4. Compared with baseline model R50, models R50+t, R50+r, R50+s and R50+f use each image transform as data augmentation only without enforcing visual attention consistency. Thus, there is minor AP gain from these models for each label. As the attention consistency under the image transforms is considered, the AP gains from models R50+ACr, R50+ACs, R50+ACf and R50+ACfs of the proposed network are significant for most labels. Still, the AP gain from model R50+ACt for each label is minor since attention consistency under image translation is already preserved by the baseline model R50. Besides, the AP gains from the models R50+ACr, R50+ACs, R50+ACf and R50+ACfs of the proposed network for label 1, 4, 6, and 10 are actually limited since the baseline model R50 has already achieved performance of APs around 95% for these labels and there is not much space for performance improvement.

To further clarify the effect of attention consistency, we quantify the attention inconsistency, measured by Eq. (5), on WIDER test set under flipping and scaling for both baseline (R50) and the proposed method (R50+ACfs) in Table 2. By enforcing attention consistency in CNN training, we can notice that the inconsistency value under each transform of the proposed method is much lower than that of the base-

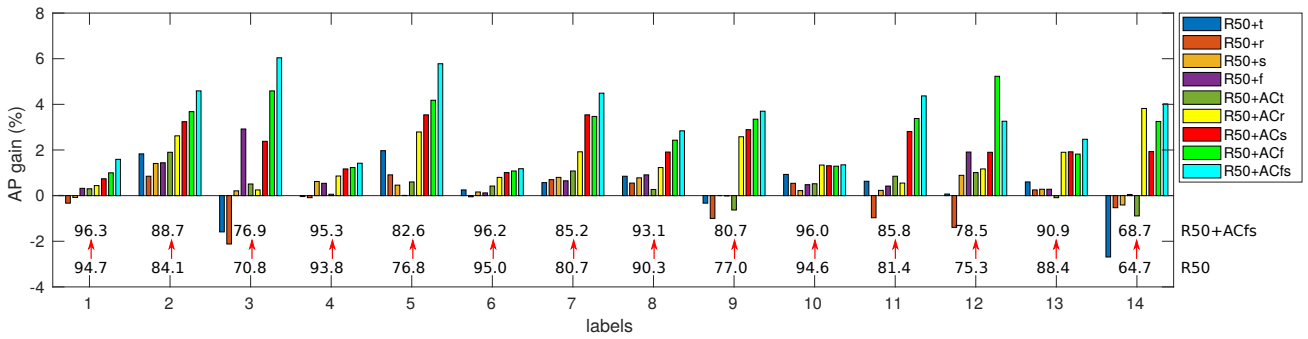

Figure 4. The average precision (AP) gain for each label on WIDER compared to model R50. Legends in the figure: models are the same a in Table 1. Two sets of APs from model R50 and model R50+ACfs are also displayed under each set of AP gains.The labels 1 to 14 are: male, long hair, sunglasses, hat, t-shirt, long sleeves, formal, short, jeans, long pants, skirt, face mask, logo and stripe.

line. Besides, to illustrate the advantage of using the consistency at attention level in this paper, we conduct comparison experiment by considering consistency at different levels under image flipping. As shown in Table 3, comparing with enforcing consistency early, e.g., in feature maps of the last convolutional layer (feature level), or late, e.g., in the final label prediction (label level), our method of enforcing consistency in the middle level, i.e., attention heatmaps, leads to the best performance.

4.2. Comparison with state of the arts

To verify that our method can achieve state-of-the-art results, we compare multi-label image classification performance of the proposed network with several state-of-the-art methods on WIDER, PA-100K, and MS-COCO. Using the same training strategy as in Section 4.1, we train the proposed network using attention consistency under different image transforms for each dataset. Different CNN architectures, ResNet50 and/or ResNet101, are used as backbone in the proposed network for different datasets. Baseline models are also trained accordingly as in Section 4.1 for each dataset.

Table 4 shows the label-based evaluation results of the comparison methods and the proposed method on WIDER. Prior to our method, VAA [44] achieves the best performance on this dataset with an mAP of 86.4%. Note that VAA uses ResNet101 as backbone and our implementation of ResNet101, i.e., model R101, achieves an mAP of 84.8%. We can see that even models of the proposed network using ResNet50 as backbone (mAP 86.8% for model R50+ACfs) can slightly outperform previous state of the arts. With ResNet101 as the backbone, the mAP of the proposed method is further increased to 87.1% when considering horizontal flipping (model R101+ACf), and 87.5% when considering both horizontal flipping and image scaling (model R101+ACfs).

Table 5 shows the evaluation results of the comparison methods and the proposed method on PA-100K dataset.

Table 4. Performance (%) of the comparison methods and the proposed method on WIDER in terms of label-based metrics. The method ResNet101* represents the baseline used in work [67] implemented from the original ResNet101 [20] with multiple data augmentations.

	method	mAP	F1-C	P-C	R-C	F1-O	P-O	R-O
	R-CNN [15]	80.0	-	-	-	-	-	-
	R*CNN [16]	80.5	-	-	-	-	-	-
	DHC [34]	81.3	-	-	-	-	-	-
	AR [18]	82.9	-	-	-	-	-	-
	ResNet101* [67]	85.0	74.7	-	-	80.4	-	-
	SRN [67]	86.2	75.9	-	-	81.3	-	-
	VAA [44]	86.4	-	-	-	-	-	-
Ours	R50	83.4	73.9	79.5	69.4	79.4	82.3	76.6
Ours	R50+ACs	85.6	75.3	81.9	70.1	80.6	84.5	77.1
Ours	R50+ACf	86.3	76.4	78.9	74.3	81.2	82.6	79.8
Ours	R50+ACfs	86.8	76.5	**82.4**	72.1	81.8	84.4	79.3
Ours	R101	84.8	75.5	80.5	71.5	80.6	83.6	77.8
Ours	R101+ACs	86.5	76.5	**82.4**	71.9	81.6	**85.1**	78.3
Ours	R101+ACf	87.1	77.3	80.9	74.5	82.1	83.8	80.5
Ours	R101+ACfs	**87.5**	**77.6**	81.3	**74.8**	**82.4**	84.1	**80.7**

Table 5. Performance (%) of the comparison methods and the proposed method on PA-100K.

	method	mA	Acc	Prec	Recall	F1-score
	DM [30]	72.7	70.39	82.24	80.42	81.32
	HP-Net [36]	74.21	72.19	82.97	82.09	82.53
Ours	R50	78.12	75.23	88.47	83.41	85.86
Ours	R50+ACs	77.46	78.25	89.96	83.97	86.86
Ours	R50+ACf	79.05	**79.46**	**90.21**	85.10	87.58
Ours	R50+ACfs	**79.16**	79.44	88.97	**86.26**	**87.59**

The prior state-of-the-art performance is achieved by HP-Net [36]. We use the ResNet50, i.e., model R50 in Table 5, as our baseline model, which has already outperformed HP-Net. As the attention consistency under different image transforms is considered, our models R50+ACs, R50+ACf, and R50+ACfs achieve better performance of F1-score than model R50.

On MS-COCO dataset, we show the label-based evaluation results of the comparison methods and the proposed method in Table 6. For fair comparison, we evaluate these metrics both with and without top-3 label constraint, which means top-3 labels with the highest presence

Table 6. Performance (%) of the comparison methods and the proposed method on MS-COCO dataset with label-based metrics. The method ResNet101* represents the baseline used in work [67] implemented from the original ResNet101 [20] with complex data augmentations.

Method		All							top-3					
		mAP	F1-C	P-C	R-C	F1-O	P-O	R-O	F1-C	P-C	R-C	F1-O	P-O	R-O
WARP [17]		-	-	-	-	-	-	-	55.7	59.3	52.5	60.7	59.8	61.4
CNN-RNN [53]		-	-	-	-	-	-	-	60.4	66.0	55.6	67.8	69.2	**66.4**
ResNet101* [67]		75.2	69.5	80.8	63.4	74.4	82.1	68.0	65.9	84.3	57.4	71.7	86.5	61.3
ResNet101-SRN [67]		77.1	71.2	81.6	65.4	75.8	82.7	69.9	67.4	85.2	58.8	72.9	87.4	62.5
baseline	ResNet101	74.9	69.7	70.1	69.7	73.7	73.6	73.7	66.1	77.7	59.8	71.2	82.2	62.8
Ours	ResNet101-ACs	76.8	70.1	**83.3**	62.1	74.9	**85.7**	66.5	66.3	**87.6**	56.3	72.0	**89.6**	60.1
	ResNet101-ACf	77.3	71.9	73.5	**71.0**	75.7	76.5	**74.9**	67.9	81.9	**61.0**	73.0	84.5	64.2
	ResNet101-ACfs	**77.5**	**72.2**	77.4	68.3	**76.3**	79.8	73.1	**68.0**	85.2	59.4	**73.1**	86.6	63.3

prediction scores are obtained for each image even if their score values are lower than 0.5, as in [53, 67]. ResNet101-SRN [67] achieves the state-of-the-art performance of mAP 77.1%, and its baseline model, ResNet101*, achieves mAP of 75.2% by using multiple data augmentations for training, including mirror and multi-scale four-corner and central crop operations. To achieve comparable baseline performance without such complex data augmentations, we simply resize the input images to 288×288 when training our baseline model, i.e., R101, using ResNet101. We train the proposed network using the same strategy as described in Section 4.1. Even though our baseline model R101 (mAP 74.9%) achieves slightly worse performance than ResNet101* (mAP 75.2%), our models R101-ACs and R101-ACf of the proposed network achieve comparable performance to ResNet101-SRN (mAP 77.1%). Furthermore, our model R101-ACfs (mAP 77.5%) of the proposed network outperforms the previous state-of-the-art methods. Besides, compared to our baseline model R101, our model R101+ACfs shows a clear performance improvement by considering attention consistency – it improves mAP by 2.6%, F1-C by 2.5%, F1-O by 2.6%, F1-C (top-3) by 1.9%, F1-O (top-3) by 1.9%.

4.3. Qualitative comparison

To verify that the attention heatmaps are refined by attention consistency, we compare the attention heatmaps extracted from the original, flipped and scaled images for the same label using baseline model and models of the proposed network. Figure 5 shows an example of attention heatmaps for the label "T-shirt". Attention regions from model R50 are inconsistent under both horizontal flipping and image scaling transforms. Our model R50+ACf produces highly consistent attention regions under image flipping, but slightly inconsistent ones under image scaling. Contrarily, our model R50+ACs produces highly consistent attention regions under image scaling, but inconsistent ones under image flipping. By considering attention consistency under both flipping and scaling, R50+ACfs produces highly consistent attention regions under both transforms. Comparing the attention heatmaps in columns 3, 4, 5 of row 2 to the one in column 2 of row 2, attention regions produced by the proposed network are more semantically relevant to the label "T-shirt". These qualitative results demonstrate that the proposed network can focus attention on regions more label-relevant by enforcing attention consistency under certain image transforms.

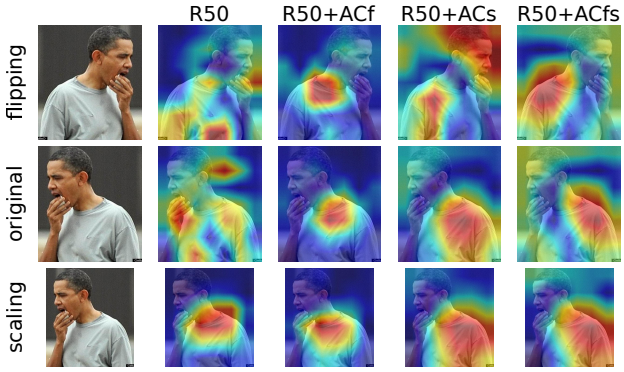

Figure 5. Attention heatmaps for classifying label "T-shirt" from flipped (row 1), original (row 2), and scaled (row 3) images using different models.

5. Conclusion

Motivated by the observations that human visual perception is consistent in classifying images under certain spatial transforms, in this paper we further assumed the consistency of CNN attention regions for image classification under such transforms, i.e., the attention region for a classification follows the same transform if the input image is spatially transformed. We found that such consistency is usually not well preserved for many CNN classifiers. To address this problem, we proposed a two-branch network, as well as an attention consistency loss, for multi-label image classification. We conducted experiments on three public datasets and the experiment results verified the effectiveness of the proposed method by achieving the new state-of-the-art performances on all three datasets.

Acknowledgment: This work was supported, in part, by NSF-1658987, NSFC61672376, and NSFC-U1803264.

References

[1] Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Describing people: A poselet-based approach to attribute classification. In *IEEE International Conference on Computer Vision*, pages 1543–1550. IEEE, 2011. 1

[2] Matthew R Boutell, Jiebo Luo, Xipeng Shen, and Christopher M Brown. Learning multi-label scene classification. *Pattern Recognition*, 37(9):1757–1771, 2004. 2

[3] Ricardo Cabral, Fernando De la Torre, Joao Paulo Costeira, and Alexandre Bernardino. Matrix completion for weakly-supervised multi-label image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(1):121–135, 2015. 1, 2

[4] Arslan Chaudhry, Puneet K Dokania, and Philip HS Torr. Discovering class-specific pixels for weakly-supervised semantic segmentation. *arXiv preprint arXiv:1707.05821*, 2017. 2

[5] Xiao Chu, Wei Yang, Wanli Ouyang, Cheng Ma, Alan L Yuille, and Xiaogang Wang. Multi-context attention for human pose estimation. *arXiv preprint arXiv:1702.07432*, 1(2), 2017. 2

[6] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: a real-world web image database from national university of singapore. In *the ACM International Conference on Image and Video Retrieval*, page 48. ACM, 2009. 1

[7] Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International Conference on Machine Learning*, pages 2990–2999, 2016. 2

[8] Krzysztof Dembczynski, Wojciech Kotlowski, and Eyke Hüllermeier. Consistent multilabel ranking through univariate losses. *arXiv preprint arXiv:1206.6401*, 2012. 3

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009. 2, 5

[10] Yubin Deng, Ping Luo, Chen Change Loy, and Xiaoou Tang. Pedestrian attribute recognition at far distance. In *the ACM International Conference on Multimedia*, pages 789–792. ACM, 2014. 1

[11] Robert Desimone and John Duncan. Neural mechanisms of selective visual attention. *Annual review of neuroscience*, 18(1):193–222, 1995. 1

[12] Ali Diba, Vivek Sharma, Ali Mohammad Pazandeh, Hamed Pirsiavash, and Luc Van Gool. Weakly supervised cascaded convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 3, page 9, 2017. 2

[13] Sander Dieleman, Jeffrey De Fauw, and Koray Kavukcuoglu. Exploiting cyclic symmetry in convolutional neural networks. *arXiv preprint arXiv:1602.02660*, 2016. 2

[14] Weifeng Ge, Sibei Yang, and Yizhou Yu. Multi-evidence filtering and fusion for multi-label classification, object detection and semantic segmentation based on weakly supervised learning. *arXiv preprint arXiv:1802.09129*, 2018. 2

[15] Ross Girshick. Fast r-cnn. *arXiv preprint arXiv:1504.08083*, 2015. 7

[16] Georgia Gkioxari, Ross Girshick, and Jitendra Malik. Contextual action recognition with r* cnn. In *IEEE International Conference on Computer Vision*, pages 1080–1088, 2015. 7

[17] Yunchao Gong, Yangqing Jia, Thomas Leung, Alexander Toshev, and Sergey Ioffe. Deep convolutional ranking for multilabel image annotation. *arXiv preprint arXiv:1312.4894*, 2013. 1, 2, 8

[18] Hao Guo, Xiaochuan Fan, and Song Wang. Human attribute recognition by refining attention heat map. *Pattern Recognition Letters*, 94:38–45, 2017. 1, 2, 3, 7

[19] Emily M Hand, Carlos Castillo, and Rama Chellappa. Doing the best we can with what we have: Multi-label balancing with selective learning for attribute prediction. In *the Association for the Advancement of Artificial Intelligence*. AAAI, 2018. 1, 2

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 2, 3, 5, 7, 8

[21] Hexiang Hu, Guang-Tong Zhou, Zhiwei Deng, Zicheng Liao, and Greg Mori. Learning structured inference neural networks with label relations. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2960–2968, 2016. 1, 2

[22] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. *arXiv preprint arXiv:1709.01507*, 7, 2017. 2

[23] Gao Huang, Zhuang Liu, Kilian Q Weinberger, and Laurens van der Maaten. Densely connected convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, page 3, 2017. 2, 3

[24] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015. 2

[25] Dahun Kim, Donghyeon Cho, Donggeun Yoo, and In So Kweon. Two-phase learning for weakly supervised object localization. *arXiv preprint arXiv:1708.02108*, 2017. 2

[26] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 1, 2

[27] Nilli Lavie. Distracted and confused?: Selective attention under load. *Trends in cognitive sciences*, 9(2):75–82, 2005. 1

[28] Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 991–999, 2015. 2, 4

[29] Karel Lenc and Andrea Vedaldi. Learning covariant feature detectors. In *European Conference on Computer Vision*, pages 100–117. Springer, 2016. 2

[30] Dangwei Li, Xiaotang Chen, and Kaiqi Huang. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In *Asian Conference on Pattern Recognition*, pages 111–115. IEEE, 2015. 3, 7

[31] Dangwei Li, Zhang Zhang, Xiaotang Chen, Haibin Ling, and Kaiqi Huang. A richly annotated dataset for pedestrian attribute recognition. *arXiv preprint arXiv:1603.07054*, 2016. 1, 3, 5

[32] Qiang Li, Maoying Qiao, Wei Bian, and Dacheng Tao. Conditional graphical lasso for multi-label image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2977–2986, 2016. 1, 2

[33] Xin Li, Feipeng Zhao, and Yuhong Guo. Multi-label image classification with a probabilistic label enhancement model. In *Uncertainty in Artificial Intelligence*, volume 1, page 3, 2014. 1, 2

[34] Yining Li, Chen Huang, Chen Change Loy, and Xiaoou Tang. Human attribute recognition by deep hierarchical contexts. In *European Conference on Computer Vision*, pages 684–700. Springer, 2016. 1, 2, 3, 5, 7

[35] 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. 2, 5

[36] Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Junjie Yan, and Xiaogang Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *IEEE International Conference on Computer Vision*, pages 1–9, 2017. 2, 3, 5, 7

[37] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015. 2

[38] Diego Marcos, Michele Volpi, Nikos Komodakis, and Devis Tuia. Rotation equivariant vector field networks. In *IEEE International Conference on Computer Vision*, pages 5048–5057, 2017. 2

[39] Jeffrey Moran and Robert Desimone. Selective attention gates visual processing in the extrastriate cortex. *Science*, 229(4715):782–784, 1985. 1

[40] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pages 483–499. Springer, 2016. 2

[41] Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 685–694, 2015. 2

[42] Dim P. Papadopoulos, Alasdair D. F. Clarke, Frank Keller, and Vittorio Ferrari. Training object class detectors from eye tracking data. In *European Conference on Computer Vision*, pages 361–376, 2014. 2

[43] Siamak Ravanbakhsh, Jeff Schneider, and Barnabas Poczos. Equivariance through parameter-sharing. In *International Conference on Machine Learning*, pages 2892–2901. JMLR. org, 2017. 2

[44] Nikolaos Sarafianos, Xiang Xu, and Ioannis A Kakadiaris. Deep imbalanced attribute classification using visual attention aggregation. *arXiv preprint arXiv:1807.03903*, 2018. 2, 7

[45] Jing Shao, Kai Kang, Chen Change Loy, and Xiaogang Wang. Deeply learned attributes for crowded scene understanding. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4657–4666, 2015. 1

[46] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[47] Marijn F Stollenga, Jonathan Masci, Faustino Gomez, and Jürgen Schmidhuber. Deep networks with internal selective attention through feedback connections. In *Advances in neural information processing systems*, pages 3545–3553, 2014. 2

[48] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich, et al. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2015. 3

[49] James Thewlis, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object frames by dense equivariant image labelling. In *Advances in neural information processing systems*, pages 844–855, 2017. 2

[50] James Thewlis, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks by factorized spatial embeddings. In *IEEE International Conference on Computer Vision*, pages 5916–5925, 2017. 2

[51] Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, 3(3), 2006. 1, 2

[52] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. *arXiv preprint arXiv:1704.06904*, 2017. 2

[53] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. Cnn-rnn: A unified framework for multi-label image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2285–2294. IEEE, 2016. 1, 2, 8

[54] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, page 3, 2017. 2

[55] Yunchao Wei, Wei Xia, Junshi Huang, Bingbing Ni, Jian Dong, Yao Zhao, and Shuicheng Yan. Cnn: single-label to multi-label. *arXiv preprint arXiv:1406.5726*, 2014. 1, 2

[56] Daniel Worrall and Gabriel Brostow. Cubenet: Equivariance to 3d rotation and translation. In *European Conference on Computer Vision*, pages 567–584, 2018. 2

[57] Daniel E Worrall, Stephan J Garbin, Daniyar Turmukhambetov, and Gabriel J Brostow. Harmonic networks: Deep translation and rotation equivariance. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5028–5037, 2017. 2

[58] Tianjun Xiao, Yichong Xu, Kuiyuan Yang, Jiaxing Zhang, Yuxin Peng, and Zheng Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 842–850, 2015. 2

[59] Jianming Zhang, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation

backprop. In *European Conference on Computer Vision*, pages 543–559. Springer, 2016. 2

[60] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837, 2014. 1, 2, 5

[61] Ning Zhang, Manohar Paluri, Marc'Aurelio Ranzato, Trevor Darrell, and Lubomir Bourdev. Panda: Pose aligned networks for deep attribute modeling. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1637–1644, 2014. 1

[62] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas Huang. Adversarial complementary learning for weakly supervised object localization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[63] Bo Zhao, Xiao Wu, Jiashi Feng, Qiang Peng, and Shuicheng Yan. Diversified visual attention networks for fine-grained object classification. *IEEE Transactions on Multimedia*, 19(6):1245–1256, 2017. 2

[64] Rui Zhao, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Saliency detection by multi-context deep learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1265–1274, 2015. 2

[65] Rui-Wei Zhao, Jianguo Li, Yurong Chen, Jia-Ming Liu, Yu-Gang Jiang, and Xiangyang Xue. Regional gating neural networks for multi-label image classification. In *British Machine Vision Conference*, 2016. 1, 2

[66] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929. IEEE, 2016. 1, 2, 3

[67] Feng Zhu, Hongsheng Li, Wanli Ouyang, Nenghai Yu, and Xiaogang Wang. Learning spatial regularization with image-level supervisions for multi-label image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5513–5522, 2017. 1, 2, 5, 7, 8