International Journal of Computer Vision (2022) 130:1088–1106
https://doi.org/10.1007/s11263-022-01591-y

Visual Attention Consistency for Human Attribute Recognition

Hao Guo[1] · **Xiaochuan Fan**[2] · **Song Wang**[1]

Received: 6 March 2021 / Accepted: 8 February 2022 / Published online: 5 March 2022

Abstract
The recognition of a human attribute is usually determined by certain regions of the input image, e.g., certain part of the human body, and such attribute-region relevance plays an important role in human attribute recognition. In deep networks, this attribute-region relevance can be derived as an interpretive attention map, where highlighted areas indicate the most relevant regions that contribute to the final recognition. Based on the assumption that more plausible attention maps indicate better networks, in this paper, we propose a new approach for human attribute recognition by exploring and enforcing two kinds of attention consistency in network learning. One kind of consistency enforces the equivariance of the attention map when the input image undergoes certain spatial transforms, such as scaling, rotation and flipping. The other kind of the consistency is enforced between the attention maps derived from two different networks when both of them are trained for recognizing the same attribute from the same image. We formulate these two kinds of consistency as new loss functions and combine them with the traditional classification loss for network training. Experiments on three datasets of human attribute recognition verify the effectiveness of the proposed method by achieving new state-of-the-art performance.

Keywords Attention maps · Human attribute recognition · Attention consistency · Equivariance

1 Introduction

As an important task in computer vision, human attribute recognition has been well explored in recent years. It aims to tell the presence of semantic attributes, e.g., gender, clothing and hair style, for a person in an image and has drawn increasing interests with many applications, such as person re-identification (Han et al. 2018; Su et al. 2016; Lin et al. 2019), person retrieval (Feris et al. 2014) and pedestrian detection (Tian et al. 2015). Many advanced deep neural networks have been developed for enhancing the performance of human attribute recognition from different perspectives, such as exploring attribute dependencies (Wang et al. 2016, 2017; Sarfraz et al. 2017; Zhao et al. 2018; Han et al. 2019; Li et al. 2019; Tan et al. 2020) and discovering the attribute-relevant regions (Zhang et al. 2014; Li et al. 2018; Liu et al. 2018, 2017; Zhu et al. 2017; Guo et al. 2017; Sarafianos et al. 2018; Tang et al. 2019; Tan et al. 2019) or attribute-related contexts (Li et al. 2016; Wang et al. 2017). However, it remains as a very challenging task due to limited training images and highly varied appearances.

One of the most important properties of human attributes is their *local spatiality*, i.e., an attribute is usually related to particular human-body parts, local image regions, or certain contexts (Zhang et al. 2014; Li et al. 2018; Liu et al. 2018, 2017; Zhu et al. 2017; Guo et al. 2017; Sarafianos et al. 2018; Tang et al. 2019; Tan et al. 2019), e.g., the attribute of "wearing sunglasses" usually appears at the face region of a person. Prior researches in cognitive science (Moran and Desimone 1985; Lavie 2005) and neuroscience (Desimone and Duncan 1995) show that our human vision actually recognizes an attribute by discovering and focusing visual attention on such local discriminative regions. By simulating this attention in human vision (Koch et al. 2006; Eriksen and Hoffman 1972; Treisman and Gelade 1980; Koch and Ullman 1987; Connor et al. 2004), many deep networks have been designed

Communicated by Suha Kwak.

✉ Song Wang
songwang@cec.sc.edu

Hao Guo
hguo@email.sc.edu ; hguosc@gmail.com

Xiaochuan Fan
efan3000@gmail.com

[1] Department of Computer Science and Engineering, University of South Carolina, Columbia, SC 29201, USA

[2] JD.com American Technologies Corporation, Mountain View, CA, USA

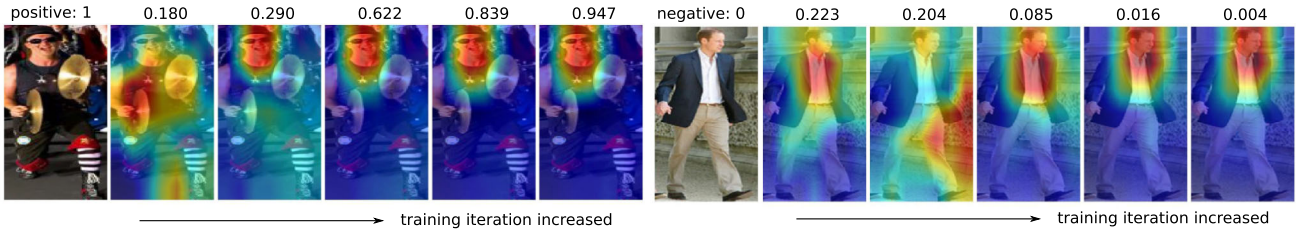

Fig. 1 Attention maps for attribute "sunglasses" in different iterations of a deep network (ResNet50) training, where face is the desired attribute-relevant region. The number above each attention map represents the predicted presence score (in [0, 1]) in the corresponding iteration

to generate attention maps by identifying the local image regions that contribute most to the final recognition. In most cases, such attention maps are computationally estimated as an intermediate result (or a byproduct) of the model prediction with only image-level supervision, and have been widely used for network interpretation (Zhou et al. 2016; Selvaraju et al. 2017; Bansal et al. 2020). In this paper, we leverage the attention map for further enhancing the performance of human attribute recognition.

Our basic assumption is that the correctness of attention maps reflects the performance of the trained deep network. An example is shown in Fig. 1, where we train a ResNet50 (He et al. 2016) network for human attribute recognition. We can see that, with more training iterations, the attention maps of attribute "sunglasses" on these two images are getting more focused to the face regions, and meanwhile, the network is getting better trained by outputting more accurate prediction scores, i.e., higher score for the attribute presence in the left image and lower score for the attribute absence in the right image. Motivated by this, we propose to incorporate the plausibility of attention maps into network training for improving human attribute recognition.

One straightforward approach to achieve this goal is to impose implicit supervision of attribute-relevant regions in deep network learning, which requires the pixel-level ground truth of attention maps. However, annotating such pixel-level ground truth for large-scale training images is infeasible due to cognitive ambiguity and intensive labor involved in manual annotations. For example, it is not a trivial work to identify regions relevant to the attribute of "Age Between 18 and 60". Besides, in the same image, different attributes have different attention maps, which further increases the difficulty and workload of manual annotation.

Motivated by human usually paying attention on consistent image regions for recognizing an attribute, in this paper we study the consistency of the attention maps when recognizing an attribute in an image to improve the plausibility of attention maps without using the ground-truth attention maps. Specifically, we consider two kinds of attention consistency: *equivariance under spatial transforms* and *invariance between different networks*. For the former, from a well trained network, the attention map of the same attribute in

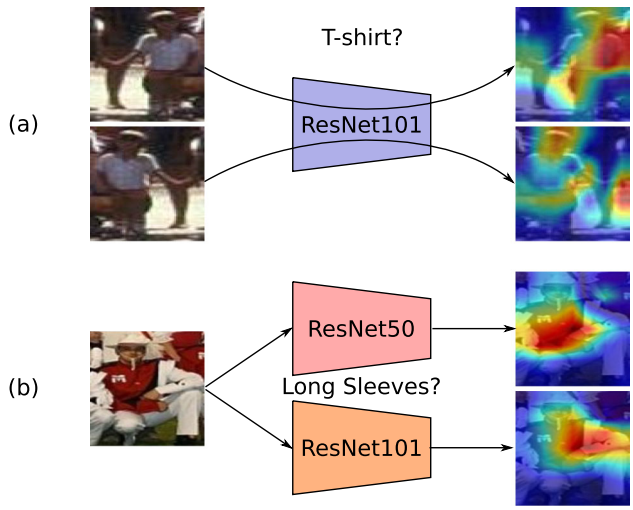

Fig. 2 An illustration of visual attention inconsistency in the current networks for human attribute recognition. **a** In recognizing the attribute "T-shirt" using a ResNet101 (He et al. 2016), the flipping of the input image does not lead to the flipping of the attention map. **b** In recognizing the attribute "Long Sleeves" in an image, two networks, ResNet50 and ResNet101, produce different attention maps

the same image shall be equivariant to certain spatial image transforms, i.e., if the input image undergoes a rotation, flipping or scaling transform, the attention maps derived from the network shall show the same transform to capture the consistency of attribute-relevant regions. For the latter, when two different networks are well trained for human attribute recognition, they shall produce identical attention maps when recognizing the same attribute in the same image, since the underlying attribute-relevant regions, even if difficult to manually annotate sometimes, is a visual perception concept independent of the adopted network. However, neither of these two kinds of consistency is well preserved in the current deep neural networks learned for human attribute recognition. More specifically, Fig. 2a shows an example where the attention consistency of equivariance under image transforms is violated and Fig. 2b shows an example where attention consistency of invariance between different networks is violated. While preliminary study of the first kind of attention consistency, i.e., equivariance under image transforms, has been studied in our early conference paper (Guo et al. 2019),

in this paper, we study both kinds of attention consistency by enforcing and combining them in a unified network training for better human attribute recognition.

To achieve these two kinds of attention consistency, we propose a two-branch framework where both branches are deep networks learned to recognize the same set of human attributes by minimizing cross-entropy-based image classification loss. Meanwhile, we use Class Activation Mapping (CAM) (Zhou et al. 2016) to estimate attention maps for each branch. For the attention consistency of equivariance under spatial transforms, we train the same network with shared parameters for the two branches, while the input image of one branch is spatially transformed as the input of the other branch. We then define a new attention consistency loss that measures the difference between the attention maps of two branches after applying the inverse spatial transform to the attention maps of the transformed image. For the attention consistency of invariance between two networks, we use different networks for two branches, which take the same image as input. We then use the new attention consistency loss to measure the difference of the two attention maps for recognizing the same attribute. Finally, we also consider the combination of two kinds of consistency by using two different networks for the two branches, and spatially transforming the input of one branch as the input of the other branch. In each case, the defined new consistency loss is added to the original classification losses for training the respective networks for human attribute recognition.

We evaluate the proposed methods on three representative datasets for human attribute recognition: WIDER Attribute (Li et al. 2016), PA-100K (Liu et al. 2017), and RAP (Li et al. 2016). The experimental results verify the effectiveness of each of the two kinds of proposed attention consistency as well as the combination of them. Our proposed methods achieve new state-of-the-art performances on these datasets. Part of this work has been published in a conference paper (Guo et al. 2019). As an extension, this paper has the following main differences from the conference version.

- Comparing to the conference paper focusing on only attention consistency of equivariance under image transforms (in the same network), this work discusses two types of attention consistency, including the equivariance under image transforms and the invariance between different networks, and their combinations for human attribute recognition.
- In this work, we unify the proposed framework for both kinds of attention consistency, with attention consistency loss update.
- Comprehensive analysis is conducted on why the attention-level consistency is selected instead of early feature-level consistency or late prediction-level consistency.
- We add more experiments to verify that the attention consistency of invariance between networks can further improve the performance of human attribute recognition comparing to attention consistency of equivariance under image transforms. Also, their combination performs even better than either of them.

The remainder of the paper is organized as follows. In Sect. 2, we conduct a literature review for the works related to human attribute recognition, deep visual attention and consistency-based network regularization. In Sect. 3, we elaborate on the proposed methods of enforcing attention consistency for human attribute recognition. In Sect. 4, we conduct comprehensive experiments to verify the effectiveness of the proposed methods, followed by a brief conclusion in Sect. 5.

2 Related work

2.1 Human attribute recognition

Human attribute recognition has been widely studied by many computer vision researchers (Wang et al. 2019). Earlier methods (Bourdev et al. 2011; Zhu et al. 2013; Deng et al. 2014) leverage hand-crafted features to recognize each attribute based on human appearance. As deep networks grow prosperous in the computer-vision field, convolutional neural networks (CNNs) (Krizhevsky et al. 2012; Simonyan et al. 2014; Szegedy et al. 2015; He et al. 2016; Huang et al. 2017) have become a standard component (Sudowe et al. 2015; Li et al. 2015) and achieved great success in human attribute recognition.

Recent methods for human attribute recognition can be classified into two main categories: attribute-correlation methods (Wang et al. 2016, 2017; Sarfraz et al. 2017; Zhao et al. 2018; Han et al. 2019; Li et al. 2019; Tan et al. 2020), which explore semantic attribute dependencies to facilitate the attribute recognition, and attribute-localization methods, which utilize the attribute-relevant image regions for spatially more focused attribute recognition. Attribute-localization methods can be further classified to two sub-categories: part-based and attention-based ones. Part-based localization (Zhang et al. 2014; Li et al. 2018; Liu et al. 2018) usually exploits the pose estimation, body-part detections or manual annotations to specify the attribute-relevant image regions, which can be sensitive to pose changes and occlusions. Attention-based localization (Liu et al. 2017; Zhu et al. 2017; Sarafianos et al. 2018; Tang et al. 2019; Tan et al. 2019; Wu et al. 2020) applies attention mechanisms to discover attribute-relevant image regions for improving attribute recognition. Recent works also attempt to refine the network attention by enhancing the attention concentration (Guo et

al. 2017; Sun et al. 2020). Our proposed methods in this paper fall in the category of attention-based localization, for which we define and enforce two new kinds of visual attention consistency for regularizing the network learning in human attribute recognition.

2.2 Visual Attention Maps

The visual attention of deep networks has drawn significant research interest in recent years. In general, prior works on deep visual attention can be categorized to either bottom-up attention or top-down attention. The bottom-up attention maps are learned during network forwarding as soft masks to actively help the network focus on discriminative regions, such as STN (Jaderberg et al. 2015), SENet (Hu et al. 2017), CBAM (Woo et al. 2018), and RAN (Wang et al. 2017). The top-down attention maps are usually inferred based on network predictions. Instead of being used for masking, top-down attention maps are more interpretative, i.e., specifying the influence of each image region to the network decisions, and therefore they are also called attribution maps (Bansal et al. 2020). In this paper, we aim to improve the plausibility of attention maps for refining the network learning, for which we use top-down, interpretative attention maps.

Three categories of methods have been used for estimating the top-down attention maps (Bansal et al. 2020). Perturbation-based methods (Zeiler and Fergus 2014; Ribeiro et al. 2016; Fong and Vedaldi 2017; Dabkowski and Gal 2017; Ribeiro et al. 2018; Zintgraf et al. 2017) usually remove a portion of an input image before feeding the image through the network to infer the effect of the removed region to the network prediction. Gradient-based methods (Simonyan et al. 2013; Shrikumar et al. 2017; Sundararajan et al. 2017; Selvaraju et al. 2017) calculate the gradients in the back-propagation to quantize the contribution of input-image pixels to certain recognition. Structure-based methods (Oquab et al. 2015; Zhou et al. 2016) produce attention maps by re-weighting the activation maps on the basis of certain network architectures, such as global average pooling (Zhou et al. 2016). In this paper, we use a structure-based method, i.e., Class Activation Mapping (CAM) (Zhou et al. 2016), for attention map estimation, since it is differentiable and computationally efficient.

2.3 Consistency-Based Network Regularization

We consider two kinds of attention consistency to regularize the network learning in this paper. They are related to existing works on deep equivariance and collaborative learning, respectively, which we briefly discussed in the following.

2.3.1 Deep Equivariance

Equivariance is studied as an important mathematical property (Lenc and Vedaldi 2015) of spatial representations. It indicates that certain spatial representations of images should follow the same transform if the image is spatially transformed. Some image representations, such as HOG (Dalal and Triggs 2005), have been proved to be adhering to this property (Lenc and Vedaldi 2015). Previous works also attempt to construct equivariant representations (Hinton et al. 2011; Kivinen and Williams 2011; Schmidt et al. 2012), including the deep convolutional representations with certain equivariance (Lenc and Vedaldi 2015). Most of existing works on applying equivariance to deep network learning are focused on the features at certain convolutional layers (Lenc and Vedaldi 2016; Dieleman et al. 2016; Cohen and Welling 2016; Worrall et al. 2017; Thewlis et al. 2017, ?; Marcos et al. 2017; Ravanbakhsh et al. 2017; Worrall et al. 2018). Differently, in this paper, we propose the attention equivariance of deep networks at a specific semantic stage of the networks—the estimated attention maps reflecting the local spatiality of the considered attribute. In later Sect. 4, we will conduct comparison experiments to show the effectiveness of the proposed attention equivariance against the existing deep equivariance in network learning for human attribute recognition.

2.3.2 Collaborative Learning

Enforcing consistent attention maps between different networks can be regarded as a kind of collaborative learning. Following the principle that "Two heads are better than one", the training of a network can be regularized by transferring information learned by another network, as verified in the research on knowledge distillation (Hinton et al. 2015; Zagoruyko and Komodakis 2016; Tarvainen et al. 2017). Different kinds of collaborative learning, e.g., deep mutual learning (Zhang et al. 2018), co-regularization (Niu et al. 2019) and co-training (Qiao et al. 2018), have been studied to transfer information between two different networks, leading to regularized training of both networks. Most of them try to minimize the final prediction difference between networks for the same input. Each network provides smoothed ground truth for the training of other network, leading to better network performance (Müller et al. 2019). Co-teaching (Han et al. 2018; Malach and Shalev-Shwartz 2017) also learns two networks simultaneously, with each network helping the other by excluding samples with noisy labels, i.e., it uses consistency of predictions for sample selection, instead of prediction supervision. Since the final prediction is aggregated for the whole image, all these methods lack consideration of local spatiality in defining the cross-network consistency. Differently, the proposed attention consistency

between networks explicitly considers local spatiality that is important for human attribute recognition during collaborative learning.

3 Proposed Method

In this section, we first briefly introduce the proposed framework (Sect. 3.1), the adopted attribute learning based on classification loss (Sect. 3.2), and the adopted visual attention estimation by CAM (Sect. 3.3). We then elaborate on the proposed two kinds of attention consistency (Sect. 3.4) and discuss the difference between our method and existing consistency-regularization methods (Sect. 3.5).

3.1 Overview

Given an image $\mathbf{x} \in \mathbb{X}$ of a person, the task of human attribute recognition aims to predict the presence of each attribute. The ground-truth attribute annotations for the image are $\mathbf{y} \in \mathbb{Y}$, with $\mathbf{y} = \{y_1, y_2, \ldots, y_K\}$ where $y_j = 1$ if attribute j is present in the image and $y_j = 0$ otherwise. K is the number of considered attributes. $\mathbb{X}$ is the set of N training images and $\mathbb{Y}$ is their corresponding set of ground-truth annotations.

Generally, as shown in Fig. 3, the proposed framework consists of two branches. Both of them are deep networks starting with convolutional layers and ending with GAP-FC (fully connected layer after global average pooling) structure, e.g., ResNet, DenseNet (Huang et al. 2017). We use the traditional binary cross entropy loss as the classification loss $\mathcal{L}_{cls}$ to learn each branch for recognizing the same set of human attributes. Based on the structure-based attention mechanism, we adopt CAM (Zhou et al. 2016) to estimate attribute-specific attention maps for each branch. To enforce the attention consistency between two branches, we introduce a new attention consistency loss $\mathcal{L}_{con}$ based on pixel-level distance between attention maps for recognizing the same attribute in an image.

Let $\mathbf{x}$ and $\mathbf{x}'$ be the inputs, f and f' be the networks of the two branches, respectively. By defining them in different ways, we can use this two-branch framework to enforce the proposed two different kinds of consistency, respectively:

(a) To enforce the attention consistency of equivariance under spatial transforms, we set $\mathbf{x}$ and $\mathbf{x}'$ as the original and transformed images, respectively, i.e., $\mathbf{x}' = T(\mathbf{x})$, where T is a spatial transform, such as flipping, scaling and rotation. Besides, the networks in two branches are identical, sharing the architecture and parameters, i.e., $f' = f$. The attention map estimated on $\mathbf{x}'$ goes through the inverse transform T^{-1} before being compared to the attention map of $\mathbf{x}$ for the calculation of attention consistency loss $\mathcal{L}_{con}$.

(b) To enforce the attention consistency of invariance between different networks, we feed the same input to two branches, i.e., $\mathbf{x}' = \mathbf{x}$, and adopt different networks with varied architecture and/or parameters, i.e., $f' \neq f$, for the two branches. In this case, the attention maps derived from the two branches are directly compared for the calculation of attention consistency loss $\mathcal{L}_{con}$.

We can also combine these two kinds of attention consistency by setting $\mathbf{x}' = T(\mathbf{x})$ and $f' \neq f$, with a unified attention consistency loss $\mathcal{L}_{con}$, in this two-branch framework. In each case, the classification loss $\mathcal{L}_{cls}$ and the attention consistency loss $\mathcal{L}_{con}$ are combined for the whole network training. During the testing, we only use one of the branches for attribute recognition for computational efficiency and fair evaluation against existing methods.

3.2 Attribute Recognition

Human attribute recognition is a specific task of multi-label recognition. Recognizing each attribute is usually formulated as a binary classification problem. We adopt the widely used cross entropy loss (Li et al. 2015, 2016; Liu et al. 2017;

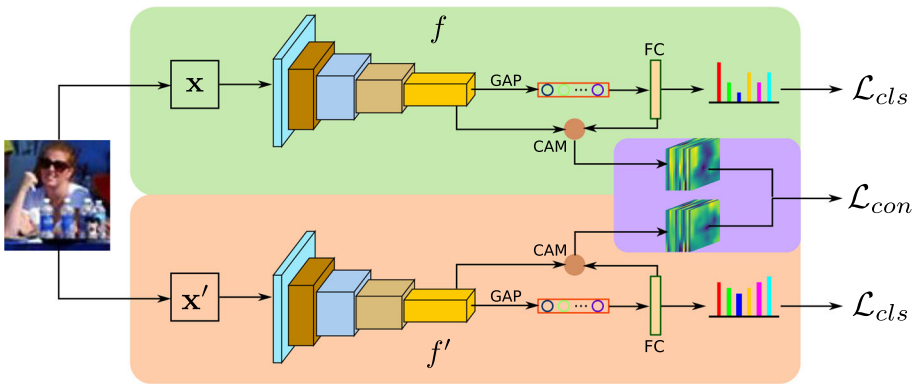

Fig. 3 An illustration of the proposed two-branch framework

Guo et al. 2019; Tan et al. 2020) for learning attributes. Let $\hat{\mathbf{y}} = [\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_K] \in \mathbb{R}^K$ be the output of the branch with $\mathbf{x}$ as input and f as the network. The classification loss would be computed as

$$\mathcal{L}_{cls}(\hat{\mathbf{y}}, \mathbf{y}) = -\frac{1}{K}\sum_{j=1}^{K} \omega_j \big(y_j \log(\sigma(\hat{y}_j)) + (1-y_j)\log(1-\sigma(\hat{y}_j))\big), \tag{1}$$

where

$$\sigma(\hat{y}_j) = 1/(1+e^{-\hat{y}_j}), \tag{2}$$

and $\hat{y}_j \in \mathbb{R}$ indicates the predicted score for attribute j being present in image $\mathbf{x}$. ω_j is used for weighting the loss from attribute j to alleviate the imbalance among different attributes. We consider two weighting strategies in the network training. The exponential strategy (Li et al. 2015) produces relatively smooth attribute weights:

$$\omega_j^e = \begin{cases} e^{1-\rho_j} & \text{if } y_j = 1, \\ e^{\rho_j} & \text{if } y_j = 0, \end{cases} \tag{3}$$

where ρ_j is the ratio of positive samples for attribute j. The square root strategy (Tan et al. 2020) heavily emphasizes the attributes with rare positive samples:

$$\omega_j^s = \begin{cases} \sqrt{1/2\rho_j} & \text{if } y_j = 1, \\ \sqrt{1/2(1-\rho_j)} & \text{if } y_j = 0. \end{cases} \tag{4}$$

Specifically, we denote $\hat{\mathbf{y}} = f(\mathbf{x})$ and $\hat{\mathbf{y}}' = f'(\mathbf{x}')$ as the output of two branches of the proposed framework, respectively. Accordingly, their classification losses can be defined as $\mathcal{L}_{cls}(\hat{\mathbf{y}}, \mathbf{y})$ and $\mathcal{L}_{cls}(\hat{\mathbf{y}}', \mathbf{y})$, respectively.

3.3 Visual Attention

We estimate the visual attention maps for attribute recognition based on the well-known class activation mapping (CAM) (Zhou et al. 2016). Feeding the input image $\mathbf{x}$ to the branch f for attribute recognition, we obtain a set of convolutional feature maps $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$, where C, H and W are the channel, height and width of the feature maps, respectively. With the Global Average Pooling (GAP), the feature maps can be aggregated to a feature vector $\mathbf{f} \in \mathbb{R}^C$, which is passed to a linear layer (Fully Connected layer, FC) for each attribute recognition. The parameters of the linear layer for recognizing multiple human attributes consist of linear weights $\mathbf{W} \in \mathbb{R}^{K\times C}$ and bias $\mathbf{b} \in \mathbb{R}^K$. The prediction for the presence of attribute j in image $\mathbf{x}$ can be written as

$$\hat{y}_j = \mathbf{w}_j \mathbf{f} + b_j, \tag{5}$$

where $\mathbf{w}_j \in \mathbb{R}^C$ is the j-th row of $\mathbf{W}$ and b_j is the j-th element of $\mathbf{b}$, and they represent the linear weights and bias for recognizing attribute j, respectively.

As each value in the feature vector $\mathbf{f}$ is aggregated from a channel (a visual pattern) of the feature maps, the learned linear weights $\mathbf{w}_j$ specify the importance of each visual pattern for recognizing the attribute j. Therefore, CAM directly maps the linear weights to the channels of feature maps $\mathbf{F}$ to estimate the $H \times W$-dimensional attention map

$$h(\mathbf{x}, j, f) = \sum_{c=1}^{C} w_{jc}\mathbf{F}_c \tag{6}$$

for attribute-j's presence in image $\mathbf{x}$, where w_{jc} is the c-th value of $\mathbf{w}_j$, and $\mathbf{F}_c \in \mathbb{R}^{H\times W}$ is the c-th channel of the feature maps $\mathbf{F}$. Note that $\mathbf{w}_j$ and $\mathbf{F}$ are derived from f and $\mathbf{x}$, respectively. Therefore, we denote CAM as a function of both the input image and the network. Similarly, the attention map estimated for the other branch can be denoted as $h(\mathbf{x}', j, f')$. Using bilinear interpolation, the attention map can be up-sampled to the input image size to indicate pixel-level evidence for or against the attribute presence, which well reflects the desired local spatiality of each attribute in the image. As shown in Fig. 3, the attention maps estimated for both branches are actually byproducts of the network prediction.

3.4 Attention Consistency

Given an attribute j, let $\mathbf{Z}_j \in \mathbb{R}^{H\times W}$ and $\mathbf{Z}'_j \in \mathbb{R}^{H\times W}$ be the *aligned* attention maps computed from the two branches, respectively, we define the attention consistency loss by

$$\mathcal{L}_{con}(\mathbf{Z}_j, \mathbf{Z}'_j) = \frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W} |z_{jhw} - z'_{jhw}|^p, \tag{7}$$

where z_{jhw} and z'_{jhw} are the elements of the aligned attention maps $\mathbf{Z}_j$ and $\mathbf{Z}'_j$ at the location (h, w), respectively, and $p > 0$ refers to a power term. Accordingly, the consistency loss on attention maps yields gradients for each attention pixel z_{jhw} as

$$\frac{\partial \mathcal{L}_{con}(\mathbf{Z}_j, \mathbf{Z}'_j)}{\partial z_{jhw}} = \frac{p}{HW}|z_{jhw} - z'_{jhw}|^{p-1}. \tag{8}$$

Similarly, the gradients for the other branch can be calculated for each attention pixels z'_{jhw}. Equation (8) indicates the pixel-level local spatiality of the attributes is well considered by optimizing the proposed attention consistency loss. In the following, we discuss the construction of the aligned attention maps $\mathbf{Z}_j$ and $\mathbf{Z}'_j$ from the estimated the CAM atten-

tion maps $h(\mathbf{x}, j, f)$ and $h(\mathbf{x}', j, f')$, respectively, to enforce the proposed two kinds of attention consistency.

3.4.1 Attention Consistency 1: Equivariance under Spatial Transforms

When attention consistency of equivariance under spatial transforms is enforced, the inputs of two branches are the original image $\mathbf{x}$ and its transformed image $\mathbf{x}' = T(\mathbf{x})$, respectively, and the two branches use the same network, i.e, $f = f'$. We need to conduct the inverse transform T^{-1} on the attention map estimated from the branch with the transformed image as input to make it spatially aligned with the attention map estimated from the branch with the original image as input, i.e.,

$$\begin{cases} \mathbf{Z}_j = h(\mathbf{x}, j, f), \\ \mathbf{Z}'_j = T^{-1}(h(T(\mathbf{x}), j, f)). \end{cases} \tag{9}$$

Here T is an image transform that does not change the visual perception, especially attention objects/contents for each attribute, in this image, such as image flipping, scaling, and rotation. While translation is also a typical spatial transform, its equivariance in both attention maps and final prediction has been well preserved in most existing deep networks, as verified in the later experiments.

3.4.2 Attention Consistency 2: Invariance between Different Networks

When enforcing attention consistency of invariance between different networks, we feed the same input, i.e., $\mathbf{x}' = \mathbf{x}$, to two branches with different networks, i.e., $f' \neq f$ and the CAM attention maps estimated from two branches are already aligned and directly comparable, i.e.,

$$\begin{cases} \mathbf{Z}_j = h(\mathbf{x}, j, f), \\ \mathbf{Z}'_j = h(\mathbf{x}, j, f'). \end{cases} \tag{10}$$

This way, two networks individually learn to recognize the same set of attributes and collaboratively learn attention maps for the same attribute from each other. Such a collaborative learning enables one network to learn missed knowledge that may be learned by the other network and vice versa, leading to enhanced learnings of both networks.

Note that our proposed method for attention consistency between networks is different from model ensemble (Zhou et al. 2002), which trains multiple networks separately and then combines the predictions. In model ensemble, all the networks must be kept in both training and testing, resulting in significantly more parameters and computational consumption. Differently, our proposed method trains two networks simultaneously by achieving consistent attention maps and in the testing stage, we only deploy one individual network. This way, our proposed method for attention consistency of invariance between two different networks has increased computation consumption in training, but uses the same number of parameters and similar computation consumption as a single network in testing. In practice, we can also use a relatively shallower network for one of two branches to avoid introducing too many new parameters in training, with the goal to only deploy the other branch in testing.

3.4.3 Combined Attention Consistency

We also consider the combined attention consistency by enforcing both the equivariance under spatial transforms and the invariance between different networks. In this case, we set $\mathbf{x}' = T(\mathbf{x})$ and $f' \neq f$, i.e., the input of one branch is spatially transformed as the input of the other branch and two branches use different networks. As in Eq. (9), we need to conduct the inverse transform T^{-1} on the attention map estimated from the branch with the transformed image as input to make it spatially aligned with the attention map estimated from the branch with the original image as input, i.e.,

$$\begin{cases} \mathbf{Z}_j = h(\mathbf{x}, j, f), \\ \mathbf{Z}'_j = T^{-1}(h(T(\mathbf{x}), j, f')). \end{cases} \tag{11}$$

This way, the unified loss Eq. (7) reflects a combination of the two kinds of consistencies.

3.5 Consistency at Different Levels

In this paper, we propose to apply consistency at the level of attention maps. Actually, consistency at other levels, such as the final prediction layer and certain feature layers, have been used in many previous works for improving deep network learning. For example, the widely used data augmentation strategy (Krizhevsky et al. 2012) assumes that a transformed image shares the same ground-truth classification as the original image and this can be regarded as enforcing the consistency of transform equivariance at the final step of recognition. Previous collaborative learning (Zhang et al. 2018; Niu et al. 2019; Qiao et al. 2018) also considers to enforce consistency between the final predictions of two different networks, as shown in Fig. 4a. One previous work (Zagoruyko and Komodakis 2016) considers the aggregated activations for information transfer between networks, which can be regarded as a feature-level consistency as shown in Fig. 4c. The attention consistency proposed in this paper is more attribute-specific on local regions, reflecting the local spatiality of attributes, as shown in Fig. 4b.

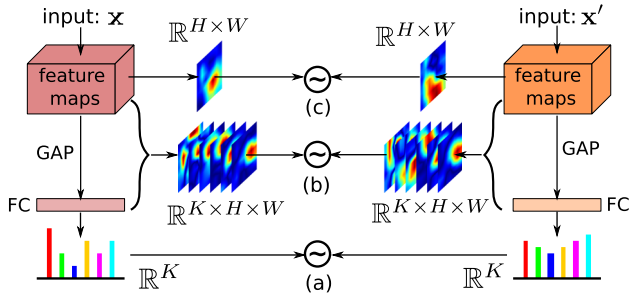

Fig. 4 An illustration of consistency at different levels: **a** final prediction, **b** attention maps, and **c** feature aggregation

More specifically, let's consider the use of prediction consistency loss of the p-th order in the previous collaborative learning, i.e.,

$$\mathcal{L}_{con}(\hat{y}_j, \hat{y}'_j) = |\hat{y}_j - \hat{y}'_j|^p. \tag{12}$$

Following Eqs. (5) and (6), we have

$$\hat{y}_j = \frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W} z_{jhw} + b_j, \tag{13}$$

and $\hat{y}'_j$ in a similar form. The produced gradients for an attention pixel would be

$$\begin{aligned}\frac{\partial \mathcal{L}_{con}(\hat{y}_j, \hat{y}'_j)}{\partial z_{jhw}} &= \frac{\partial \mathcal{L}_{con}(\hat{y}_j, \hat{y}'_j)}{\partial \hat{y}_j}\frac{\partial \hat{y}_j}{\partial z_{jhw}} \\ &= \frac{p}{HW}|\hat{y}_j - \hat{y}'_j|^{p-1}\end{aligned}. \tag{14}$$

The similar calculation can be applied to get the gradients at the pixel z'_{jhw}. Equation (14) clearly shows that, when using the prediction consistency, the gradients at different locations, i.e. with varying (h, w), are the same for both networks and the local spatiality of each attribute is not well reflected in optimizing this loss. Besides, from Eqs. (8) and (14), we can also see that when setting $p = 1$, the proposed attention consistency degrades to the prediction consistency. Therefore, in our experiments, we always choose $p > 1$.

Similarly, we can try to enforce feature consistency between $\mathbf{F}$ and $\mathbf{F}'$, i.e.,

$$\mathcal{L}_{con}(\mathbf{F}, \mathbf{F}') = \frac{1}{CHW}\sum_{c=1}^{C}\sum_{h=1}^{H}\sum_{w=1}^{W}|F_{chw} - F'_{chw}|^p. \tag{15}$$

The gradient on the pixel F_{chw} is

$$\frac{\partial \mathcal{L}_{con}(\mathbf{F}, \mathbf{F}')}{\partial F_{chw}} = \frac{p}{CHW}|F_{chw} - F'_{chw}|^{p-1}. \tag{16}$$

Since $z_{jhw} = \sum_{c=1}^{C} w_{jc}F_{chw}$ and $z'_{jhw} = \sum_{c=1}^{C} w'_{jc}F'_{chw}$, we can also calculate the gradient from the attention consistency according to Eq. (8):

$$\frac{\partial \mathcal{L}_{con}(\mathbf{Z}_j, \mathbf{Z}'_j)}{\partial F_{chw}} = \frac{pw_{jc}}{HW}\left|\sum_{c=1}^{C} w_{jc}F_{chw} - \sum_{c=1}^{C} w'_{jc}F'_{chw}\right|^{p-1}. \tag{17}$$

Comparing Eqs. (16) and (17), the feature pixel gradients are more aggregated from attention consistency than from feature consistency. It implies that the attention consistency is more semantic, i.e., attribute-j specific, than the feature consistency.

Therefore, attention consistency is more spatial than prediction consistency and more semantic than feature consistency. Moreover, we can further combine feature consistency and prediction consistency as $\mathcal{L}_{con:F+P} = \mathcal{L}_{con}(\mathbf{F}, \mathbf{F}') + \mathcal{L}_{con}(\hat{y}_j, \hat{y}'_j)$. According to Eqs. (13), (14) and (16), the gradient on pixel F_{chw} can be written as

$$\begin{aligned}\frac{\partial \mathcal{L}_{con:F+P}}{\partial F_{chw}} &= \frac{\partial \mathcal{L}_{con}(\mathbf{F}, \mathbf{F}')}{\partial F_{chw}} + \frac{\partial \mathcal{L}_{con}(\hat{y}_j, \hat{y}'_j)}{\partial F_{chw}} \\ &= \frac{p}{CHW}|F_{chw} - F'_{chw}|^{p-1} \\ &\quad + \frac{pw_{jc}}{HW}\left|\sum_{c=1}^{C}\sum_{h=1}^{H}\sum_{w=1}^{W} w_{jc}F_{chw} + b_j\right. \\ &\quad \left. - \left(\sum_{c=1}^{C}\sum_{h=1}^{H}\sum_{w=1}^{W} w'_{jc}F'_{chw} + b'_j\right)\right|^{p-1}.\end{aligned} \tag{18}$$

Equation (18) shows that directly combining the feature consistency and the prediction consistency additively leverages the spatial-level (the first term) and the semantic-level consistency (the second term). Comparing Eqs. (17) and (18), we think the attention consistency integrates feature consistency and prediction consistency organically, instead of combining them additively.

In the later Sect. 4, we will conduct comparison experiments by enforcing consistency at different levels to verify the effectiveness of the proposed methods.

3.6 End-to-End Training

Finally, the proposed two-branch network is learned in the end-to-end manner. We linearly combine the classification losses and the consistency loss by

$$\mathcal{L}_{total} = \mathcal{L}_{cls}(\hat{\mathbf{y}}, \mathbf{y}) + \mathcal{L}_{cls}(\hat{\mathbf{y}}', \mathbf{y}) + \lambda\mathcal{L}_{con}, \tag{19}$$

where λ is a hyper-parameter to balance the two kinds of losses for each network learning. Classification losses

$\mathcal{L}_{cls}(\hat{\mathbf{y}}, \mathbf{y})$ and $\mathcal{L}_{cls}(\hat{\mathbf{y}}', \mathbf{y})$ supervise the training of two branches, respectively, while the consistency loss is involved in the training of both branches.

3.7 Model Inference/Testing

After the training of the proposed two-branch framework, we use a single branch for model inference or testing to be consistent with existing methods.

(1) To enforce the attention consistency of equivariance under image transforms, two branches are identical, i.e., sharing the same architecture and parameters. Thus, we can select either branch for model inference.
(2) When enforcing the attention consistency of invariance between different networks, we define the two branches as the main branch and the auxiliary branch, respectively. The main branch is the one we want to deploy for model inference, while the auxiliary branch can be optionally deprecated after the training. Given two different branches, either the smaller branch or the larger branch can be the main branch or auxiliary branch, depending on the applications. In the following experiments, we demonstrate that the performance is improved for both the smaller branch and the larger branch. Thus, the smaller branch can be selected as the main branch for better computation efficiency, while the larger branch can be selected as the main branch for better recognition performance.

4 Experiment

4.1 Datasets and Configurations

4.1.1 Datasets

We conduct experiments on three representative human attribute datasets. *WIDER Attribute* (Li et al. 2016) consists of images with complex scene contexts and 14 human attributes. The train-val set includes 28,345 samples (22,962 images in training set for model learning), while the test set includes 29,179 samples. *PA-100K* (Liu et al. 2017) contains 100,000 human bounding boxes in total, and has the largest number of training samples in the existing attribute datasets. 26 human attributes are annotated, and the training, validation and test sets are aligned with the ratio of 8:1:1. *RAP* (Li et al. 2016) has a total number of 41,585 cropped human bounding boxes, with 69 attributes annotated, of which 51 attributes are usually recognized for evaluation. Among the existing human attribute datasets, RAP is the one with the largest number of attributes.

4.1.2 Model Training

During the model training of the proposed network, for fair comparison, we use the same backbones as the prior methods to construct the branch of the proposed method. Specifically, when enforcing attention consistency of equivariance under image transforms, denoted as VAC-TE (Transform Equivariance), the proposed network uses ResNet101 as the two identical branches to learn on WIDER dataset and ResNet50 as the two identical branches to learn on PA-100K and RAP datasets. When enforcing attention consistency of invariance between different networks, denoted as VAC-NI (Network Invariance), we use ResNet101 as the main branch on WIDER dataset and ResNet50 as the main branch on PA-100K and RAP datasets. The auxiliary branch is based on ResNet50 and ResNet34, respectively. Besides, when we conduct experiments for ablation studies, we use various networks as the backbones, which will be clarified explicitly.

We also use two sets of hyper-parameters for model training on WIDER dataset and PA-100K/RAP datasets, respectively. The reason is two-fold: (1) the dataset properties are different, such as the number of defined human attributes, the level of imbalance between positive and negative samples of each attribute, etc.; (2) we would like to align with prior state-of-the-art methods, e.g., DIAA (Sarafianos et al. 2018) and Da-HAR (Wu et al. 2020) on WIDER and ALM (Tang et al. 2019) and JLAC (Tan et al. 2020) on PA-100K and RAP datasets use different configurations. To be specific, the hyper-parameters used in our experiments are as the following.

On WIDER dataset, we use the exponential attribute loss weights of Eq. (3), initial learning rate of 0.001 (divided by 10 every 5 epochs), SGD optimizer and $p = 2$ in Eq. (7). On PA-100K and RAP, we use input image size of 256×128, the attribute loss weights of Eq. (4), initial learning rate of 0.0001, $p = 3$ in Eq. (7), and Adam optimizer. The parameter λ in Eq. (19) is set to 1 in our experiments. These configurations are mostly aligned with the comparison methods.

4.1.3 Model Testing

During model testing, our experiments mainly use a single branch of the proposed two-branch network for inference. In VAC-TE, we arbitrarily select one of the two identical branches for testing, while in VAC-NI we test the main and the auxiliary branches individually. The only exception is in comparing VAC-NI with traditional model ensemble, where we have an experiment of ensembling both VAC-NI-M and VAC-NI-A branches in model testing.

To be consistent with prior literatures, we use the metric of mean Average Precision (mAP) on WIDER dataset, and the metrics of mean Accuracy (mA), instance Accuracy (Acc.),

Table 1 Performance comparison in terms of mean Average Precision (mAP, %) between our proposed methods and existing state-of-the-art methods on WIDER dataset

Method	Backbone	Input size	1	2	3	4	5	6	7	8	9	10	11	12	13	14	mAP
R*CNN ICCV'15	VGG16	224 × 224	94	82	62	91	76	95	79	89	68	96	80	73	87	56	80.5
DHC ECCV'16	VGG16	224 × 224	94	82	64	92	78	95	80	90	69	96	81	76	88	55	81.3
SRN w/o val CVPR'17	ResNet101	224 × 224	95	87	72	92	82	95	84	92	80	96	84	76	90	66	85.1
SRN w/val CVPR'17	ResNet101	224 × 224	–	–	–	–	–	–	–	–	–	–	–	–	–	–	86.2
DIAA ECCV'18	ResNet101	224 × 224	96	88	74	93	83	96	85	93	81	96	85	78	90	68	86.4
Da-HAR AAAI'20	ResNet101	256 × 256	97	89	76	96	85	97	86	92	81	97	87	79	91	70	87.3
baseline w/o data aug	ResNet101	224 × 224	95	86	73	94	79	96	82	92	79	95	83	76	90	67	84.8
baseline w/data aug	ResNet101	224 × 224	95	87	74	95	80	96	83	92	79	96	84	77	89	65	85.2
VAC-TE (Ours)	ResNet101	224 × 224	96	89	76	96	83	97	85	94	83	96	87	79	92	69	87.5
VAC-NI-A (Ours)	ResNet50	224 × 224	97	89	77	96	84	97	86	93	82	98	87	79	91	70	87.6
VAC-NI-M (Ours)	ResNet101	224 × 224	97	90	78	96	84	97	86	93	83	98	88	80	92	71	88.1
VAC-Combine (Ours)	ResNet101	224 × 224	97	90	79	96	85	98	86	94	84	98	88	80	92	71	88.4

Our baseline method is reproduced from the baseline of Da-HAR (Wu et al. 2020) on our experimental environment, with or without random horizontal flipping as data augmentation (data aug). Attributes: 1—Male, 2—Long Hair, 3—Sunglasses, 4—Hat, 5—T-shirt, 6—Long Sleeves, 7—Formal, 8—Shorts, 9—Jeans, 10—Long Pants, 11—Skirts, 12—Face Mask, 13—Logo, 14—Plaid

Precision (P), Recall (R) and F-1 score (F1)[1] on PA-100K and RAP datasets. Compared with mA, Acc., P, R, and F1, which rely on a specific threshold, e.g., 0.5, to produce binary prediction results, mAP reflects the model performance over all possible thresholds, leading to a more comprehensive evaluation on multiple attribute recognition.

Besides, the mA and Acc. metrics on PA-100K and RAP datasets suffer from strong data imbalance and can not reflect the intrinsic dependency of multiple human attributes (Li et al. 2016). Therefore, more reliable metrics are mAP on WIDER dataset and F1 score on PA-100K and RAP datasets.

4.2 Comparison with Existing Arts

4.2.1 Performance Comparison on WIDER Dataset

Firstly, we conduct experiments to compare the proposed methods with existing state-of-the-art approaches. On WIDER dataset, we compare with R*CNN (Gkioxari et al. 2015), DHC (Li et al. 2016), SRN (Zhu et al. 2017), DIAA (Sarafianos et al. 2018) and Da-HAR (Wu et al. 2020). On WIDER dataset, we train VAC-TE by enforcing attention consistency of equivariance under a spatial transform randomly selected from *scaling* and *horizontal flipping*, with equal probability. For its scaling transform, we bi-linearly down-sample the image size from 224 × 224 to 192 × 192. Since many prior arts use ResNet101 as the backbone, we also adopt ResNet101 as the backbone of our methods for fair comparisons. In VAC-TE, both branches is constructed by an identical ResNet101 with shared parameters. In VAC-NI, one branch is constructed by ResNet101, which we regard as the *main branch*, denoted as VAC-NI-M, to learn for attribute recognition, and the other branch is constructed by ResNet50, which we regard as the *auxiliary branch*, denoted as VAC-NI-A. Using the relatively shallower auxiliary branch can help alleviate the increase of computation load in the training.

The performance comparison is reported in Table 1. Prior Da-HAR with ResNet101 as backbone achieves mAP of 87.3% over 14 human attributes. Our method VAC-TE achieves the mAP of 87.5%, while VAC-NI-M achieves the mAP of 88.1%. The comparison shows that considering attention consistency can improve the performance of human attribute recognition. Moreover, if we use the VAC-NI-A with ResNet50 for testing, the achieved performance is 87.6%, also outperforming the prior arts. As discussed in Sect. 3.4.3, we can combine these two kinds of attention consistency into a unified consistency loss, where transform T is randomly selected from scaling and flipping as mentioned above. As shown in Table 1, such combined consistency (VAC-Combine) can further improve the mAP of attribute recognition to 88.4%.

4.2.2 Performance Comparison on PA-100K Dataset

Because prior arts use ResNet50 as the backbone on PA-100K evaluation, we follow the same protocol by taking ResNet50 as the backbone in the proposed methods. On PA-100K dataset, we train VAC-TE by enforcing attention consistency of equivariance under *horizontal flipping*. Table 2 shows the performance comparison between our method and prior methods, such as DeepMar (Li et al. 2015), HPNet (Liu et al. 2017), VeSPA (Sarfraz et al. 2017), PGDM (Li et al. 2018),

[1] Detailed definitions can be found in (Li et al. 2016).

Table 2 Performance (%) comparison between our methods and prior methods on PA-100K

Method	mA	Acc.	P	R	F1
DeepMar ACPR'15	72.70	70.39	82.24	80.42	81.32
HPNet ICCV'17	74.21	72.19	82.97	82.09	82.53
VeSPA BMVC'17	76.32	73.00	84.99	81.49	83.20
PGDM ICME'18	74.95	73.08	84.36	82.24	83.29
LGNet BMVC'18	76.96	75.55	86.99	83.17	85.04
ALM ICCV'19	80.68	77.08	84.21	**88.84**	86.46
JLPLS TIP'19	81.61	78.89	86.83	87.73	87.27
JLAC AAAI'20	**82.31**	79.47	87.45	87.77	87.61
Baseline-ResNet50	81.58	78.97	86.32	86.89	86.60
VAC-TE (ours)	80.85	79.68	88.20	87.28	87.74
VAC-NI-M (ours)	82.23	80.39	88.24	88.23	88.23
VAC-Combine (ours)	82.19	**80.66**	**88.72**	88.10	**88.41**

Top performance for each metric is marked as bold

Table 3 Performance (%) comparison between our methods and prior methods on RAP dataset

Method	mA	Acc.	P	R	F1
HPNet ICCV'17	76.12	65.39	77.53	78.79	78.05
VeSPA BMVC'17	77.70	67.35	79.51	79.67	79.59
PGDM ICME'18	74.31	64.57	78.86	75.90	77.35
LGNet BMVC'18	78.68	68.00	80.36	79.82	80.09
JLPLS TIP'19	81.25	67.91	78.56	81.45	79.98
CoCNN IJCAI'19	81.42	68.37	81.04	80.27	80.65
JLAC AAAI'20	**83.69**	69.15	79.31	**82.40**	80.82
Da-HAR AAAI'20	79.44	68.86	80.14	81.30	80.72
Baseline	80.67	67.79	79.06	80.32	79.69
VAC-TE (Ours)	79.41	69.22	81.50	80.09	80.79
VAC-NI-M (Ours)	81.10	70.01	81.51	81.37	81.44
VAC-Combine (Ours)	81.30	**70.12**	**81.56**	81.51	**81.54**

Top performance for each metric is marked as bold

LGNet (Liu et al. 2018), ALM (Tang et al. 2019), JLPLS (Tan et al. 2019), and JLAC (Tan et al. 2020). Based on ResNet50, we train a baseline model with an FC–BN (Batch Normalization) structure beside the FC layer for prediction regularization. Given the input size of 256 × 128, the baseline model achieves F1 score of 86.60%. When attention consistency of transform equivariance is enforced, VAC-TE achieves F1 score of 87.74%. Considering the attention consistency of invariance between networks, VAC-NI-M based on ResNet50 as the backbone achieves new state-of-the-art performance on F1 score of 88.23%, by using ResNet34 as the auxiliary branch. Furthermore, enforcing both kinds of attention consistency, VAC-Combine further improves the performance of F1 score to 88.41%, a new state-of-the-art performance on PA-100K dataset.

4.2.3 Performance Comparison on RAP Dataset

On RAP dataset, the experiments use the same configurations as those applied to the experiments on PA-100K dataset. We also train VAC-TE by enforcing attention consistency of equivariance under *horizontal flipping*. The involved comparison methods include HPNet, VeSPA, PGDM, LGNet, JLPLS, CoCNN (Han et al. 2019), JLAC, and Da-HAR. Similar to these methods, we conduct experiments on the five different train/test splits (Li et al. 2016) and report the mean performance. As shown in Table 3, JLAC (ResNet50 backbone) and Da-HAR (ResNet101 backbone) achieve F1 scores of 80.82% and 80.72%, respectively. For fair comparison, we use the ResNet50 as our backbone. VAC-TE achieves F1 score of 80.79%, while VAC-NI-M with ResNet50 as the backbone achieves F1 score of 81.44%, of which the auxiliary branch adopts ResNet34 as the backbone. Also, when we enforce both kinds of attention consistency VAC-Combine further improves the F1 score to 81.54%. While the mA performance of our proposed VAC-TE method is lower than that of JLAC, previous researches have pointed out that mean accuracy (mA) may not reflect the intrinsic dependency among multiple attributes and suffer from the imbalance issue between positive and negative samples of each human attribute.

4.3 Ablation Studies

In the following, we conduct ablations studies to further justify the detailed settings of the proposed methods, mainly on the WIDER dataset with input image size of 224 × 224.

4.3.1 Equivariance under Different Spatial Transforms

Different spatial transforms can be considered as T of Eq. (9). Specifically, we focus on a set of frequently used transforms, including translation, rotation, scaling and flipping, for the ablation studies, since they do not change the visual perception of an image, i.e., the presence of human attributes. Certainly, in some extreme cases, e.g., down-sampling the input image to a very small size, the visual perception of an attribute may totally change. In this paper, we choose appropriate parameters for these transforms to avoid such extreme cases. The four specific transforms we involve in this ablation study are 32-pixel translation to the right with zero-padding, 90° counter-clockwise rotation, bi-linear down-scaling from 224 × 224 to 192 × 192, and horizontal flipping.

As shown in Table 4, with ResNet101 as the backbone, when there is no consideration of the attention consistency of transform equivariance in the network training for attribute recognition, the achieved mAP is 84.8%—it is slightly lower than the baseline performance of 85.2% in Table 1, because

Table 4 Performance (%) on WIDER Attribute dataset considering attention equivariance under different transforms, with ResNet101 as backbone

Transforms	mAP	F1-C	F1-O
ResNet101 w/o	84.8	75.5	80.6
Translation	84.6	75.3	80.1
Rotation	86.0	76.2	81.2
Scaling	86.5	76.5	81.6
Flipping	87.1	77.4	82.1
Scaling and flipping	**87.5**	**77.6**	**82.4**

Top performance for each metric is marked as bold
F1-C and F1-O (Zhu et al. 2017) represent the macro and micro F1 scores evaluated by averaging per attribute results and on all images over all attributes, respectively

the latter also applies random horizontal flipping as data augmentation. When attention consistency of equivariance under either rotation, scaling or flipping is adopted for network regularization, the mAP performance for attribute recognition is improved. The combination of scaling and flipping (last row of Table 4), each with a random selection probability of 50%, leads to a further improved mAP performance of 87.5% and we chose this setting of transforms in the above comparison experiments against the state of the arts on the WIDER dataset. Since deep networks are inherently equivariant to translation, by using convolution and pooling operations, further enforcement of attention consistency of equivariance under translation does not introduce clear performance improvement, as shown in Table 4.

We also conduct an experiment to compare the attention consistency of equivariance under a spatial transform with using the same transform for data augmentation only. As shown in Table 5, with ResNet50 as the backbone, enforcing attention consistency of equivariance under certain transform achieve much better performance than using the same transform for data augmentation for network learning, except for the translation. We can notice the different performance changes when enforcing attention consistency of equivariance under translation in Table 4 (−0.2% mAP) and Table 5 (+0.5% mAP, −0.2% F1-O). This could be caused by the training variance and the backbone difference. Theoretically, attention consistency of equivariance under translation should not bring in any change of performance since deep convolutional networks are inherently equivariant to translation already, as mentioned above.

Table 5 Performance (%) on WIDER Attribute dataset using certain transform for data augmentation and attention consistency of equivariance, respectively

Transform	Data Augmentation			Attention Consistency		
	mAP	F1-C	F1-O	mAP	F1-C	F1-O
ResNet50 w/o	83.4	73.9	79.4	–	–	–
Translation	83.7	74.1	79.5	83.9	74.2	79.2
Rotation	83.2	73.2	78.5	85.0	75.1	80.2
Scaling	83.9	74.4	79.4	85.6	75.3	80.6
Flipping	84.2	74.6	80.0	86.3	76.4	81.2

The backbone is ResNet50

Table 6 Performance (mAP, %) of the main branch ResNet101 when the auxiliary branch using different backbones

Auxiliary	VAC-NI-A	VAC-NI-M
Without	–	85.2
ResNet50	87.6	88.1
ResNet101	87.9	88.0
ResNet152	88.6	88.4
DenseNet121	87.5	88.3
DenseNet161	88.4	88.3

Experiments are conducted on WIDER dataset with input size of 224 × 224

4.3.2 Consistency between Different Networks

In this section, we study the influence of using different auxiliary branches when enforcing the attention consistency between two networks. For this study, we take ResNet101 as the main branch in the proposed method, and consider ResNet50, ResNet101, ResNet152, DenseNet121, and DenseNet161 as the candidate backbone of the auxiliary branch. The experiment results are reported in Table 6. Note that when using the ResNet101 as the backbone of the auxiliary branch, we initialize the ResNet101 of the main branch and the ResNet101 of the auxiliary branch differently to introduce branch diversity, i.e., main branch is initialized with parameters pretrained on ImageNet, while the auxiliary branch is initialized with parameters pretrained on WIDER dataset.

From Table 6, given the fixed main branch of ResNet101, we can measure the impact of using different networks as the auxiliary branch from the following four perspectives:

(1) *Less effective auxiliary branch can benefit the main branch*: even if the auxiliary branch itself, e.g. ResNet50 and DenseNet121, cannot achieve as good performance as the main branch, the main branch can still benefit from the proposed method by enforcing the attention consistency between the two branches.
(2) *Deeper auxiliary branch is better*: when deeper networks, e.g., ResNet152 and DenseNet161, are used as the auxiliary branch, the attribute recognition performance of the main branch can be further improved, compared to shallower networks, e.g. ResNet50 and DenseNet121,

since deeper networks may provide more robust attention maps for collaborative attention learning.

(3) *Different network structures can be used for the auxiliary branch, e.g., ResNet and DenseNet.*

(4) *Sharing architecture between main and auxiliary branches with different initialization also benefits both branches*, although the performance gain is not as good as using a different architecture as the auxiliary branch, since collaborative learning between two branches with the same architecture is easier to collapse into each other.

4.3.3 Quantitative Attention-Map Refinement

In this section, we conduct experiments to quantitatively examine whether the proposed attention consistency does improve the attention maps of the network. As mentioned earlier, constructing the ground-truth attention maps on a large-set of training images is very difficult for many attributes. Some attributes, such as "Age Between 18 and 60", may be related to ambiguous image regions and constructing its ground-truth attention map on an image may require a vision study involving a group of subjects following rigorous protocols. To quantitatively evaluate the quality of estimated attention maps, i.e., CAM, we select two attributes, "Long Hair" and "Shorts" in the WIDER dataset, with relatively unambiguous relevant regions, and manually annotate these regions. More specifically, we randomly select 200 test images for each of attributes "Long Hair" and "Shorts", and annotate the bounding boxes around the hair and shorts, respectively. By normalizing CAM attention maps to the value range of [0, 1], we define an attention response ratio as the total attention values inside the bounding box over the area of bounding box. A higher attention response ratio indicates that the obtained attention map is more aligned with the annotated attention region and therefore, shows higher quality. Table 7 shows the results of two baseline methods, where ResNet50 and ResNet101 are trained without considering attention consistency, and the proposed methods enforcing attention consistencies. For VAC-TE, we use ResNet101 with randomly selected scaling and flipping transforms as discussed above. For VAC-NI, we use ResNet101 as the main branch and ResNet50 as the auxiliary branch. Compared with the baselines, the proposed methods produce better attention maps by enforcing either type of attention consistency when recognizing these two attributes.

Table 7 Quantitative evaluation of the attention maps against the manually annotated attention regions for two attributes on selected test images in WIDER dataset

Consistency	Nets	Attention response ratio (%)	
		Long Hair	Shorts
Baselines	ResNet50	46.77	48.14
	ResNet101	47.74	48.48
VAC-TE	ResNet101	57.51	61.17
VAC-NI-A	ResNet50	59.55	58.86
VAC-NI-M	ResNet101	62.62	60.72

'Baselines' indicates that networks are trained without considering attention consistency

4.3.4 Consistency at Different Levels

In this section, we conduct experiments on WIDER dataset to compare the use of consistency at different representation levels, including feature level, attention-map level and prediction level, as discussed in Sect. 3.5.

Table 8 Performance (%) of enforcing flipping equivariance at different levels

Consistency Levels	ResNet50
w/o	83.4
Feature	85.1
Prediction	85.4
Feature + Prediction	85.7
Attention (Ours)	86.3

The result by enforcing the consistency of transform equivariance at different levels are reported in Table 8. The image transform adopted is horizontal flipping, and the backbone is ResNet50. Since the prediction for each attribute is a scalar without spatial information, we actually enforce flipping invariance of the prediction score for the prediction-level consistency. It can be regarded as an extension of the data augmentation, where the invariance is directly applied to the recognition result. For feature-level consistency, we enforce the feature equivariance of the last convolutional layer. For transform equivariance at each level, we use similar consistency loss by calculating element-wise difference, as in Eq. (7). As shown in Table 8, enforcing transform equivariance at the attention-map level achieves the best results, since the local spatiality of attribute recognition is well embedded. Also, compared with feature equivariance under transforms, attention equivariance under the same transforms encodes attribute specific spatial information in the network learning, leading to better performance. As an organic integration of feature- and prediction-level equivariance, attention-level equivariance also performs better than directly combining the feature- and prediction-level equivariance, which is shown in the row of "Feature + Prediction" in Table 8.

For the consistency between different networks, we also compare the performance by enforcing attention consistencyagainst the feature/prediction-level consistency, as

Table 9 Performance comparison (mAP(%)) of using different-level consistency for collaborative learning on WIDER dataset

Consistency levels	ResNet50	ResNet101
w/o	84.3	85.2
Feature ICLR'17	85.7	86.5
Prediction CVPR'18	86.8	87.6
Feature + Prediction	86.6	87.3
Attention (Ours)	87.6	88.1

Two networks are ResNet50 and ResNet101, and the input size is 224 × 224

discussed in Sect. 3.5. As shown in Table 9, both the considerations of feature-level consistency (Fig. 4c) and prediction-level consistency (Fig. 4a) can improve the attribute recognition performance. But the proposed method achieves the largest improvement by considering the attention-level consistency between two networks. Both experiments in Tables 8 and 9 demonstrate that attention-level consistency is superior to feature- and prediction-level consistencies for human attribute recognition. Attention invariance between different networks also performs better than directly combining feature- and prediction-level invariance, which is shown in the row of "Feature + Prediction" in Table 9.

While the direct combination of feature and prediction equivariance improves recognition performance in Table 8, the combination of feature and prediction invariance decreases the recognition performance in Table 9. A possible reason might be that, when enforcing feature invariance between different networks, the corresponding channels of two sets of feature maps from two different networks represent different visual patterns. In this case, enforcing the consistency is lack of robustness and may harm the recognition performance.

4.3.5 Comparison to Network Ensemble

In the above Sect. 3.4.2, we discuss the difference between the proposed method by enforcing attention consistency between networks and prior works on model ensemble (Zhou et al. 2002), which integrate predictions from multiple networks in the testing. Since our proposed method only deploys one branch, it has much fewer parameters and takes much less computation time than model-ensemble methods in the testing. In this section, we conduct an experiment to compare the performance of the proposed method and the model-ensemble method.

For simplicity, we average the predictions from two networks as model ensemble. As shown in Table 10, when two networks, e.g., ResNet50 and ResNet101, are trained separately, i.e., "w/o VAC", the model ensemble achieves better performance than each of them. When two networks are collaboratively learned by enforcing attention consistency by using our proposed method, either branch of our collaboratively trained networks performs better than the direct model ensemble.

If we adopt model ensemble on these two branches during testing, the performance is further improved. However, since two branches are collapsing into each other by learning from the same data collaboratively, they are trained to predict similarly for the same input. Thus, the performance improvement from model ensemble of the two branches of the proposed network is marginal. Ideally, when two branches are completely optimized to yield the exactly same attention map for the same attribute and the same input, they would predict the same result and model ensemble on them will not lead to any performance improvement. Limited by the training data volume and the network architecture capacity, the ideal case is not achieved in our experiments, leading to the above performance improvement. These results verify not only the

Table 10 Performance comparison between the proposed method and model ensemble

Datasets	Networks	w/o VAC	with VAC
WIDER (mAP, %)	ResNet50	84.3	87.6
	ResNet101	85.2	88.1
	Ensemble (50 & 101)	86.6	88.3
PA-100K (mAP, %)	ResNet34	70.91	74.07
	ResNet50	71.03	74.32
	Ensemble (34 & 50)	73.46	74.97
PA-100K (F1, %)	ResNet34	86.10	88.12
	ResNet50	86.60	88.23
	Ensemble (34 & 50)	87.83	88.35
RAP (F1, %)	ResNet34	78.98	81.02
	ResNet50	79.69	81.44
	Ensemble (34 & 50)	80.71	81.65

'VAC' indicates attention consistency between networks

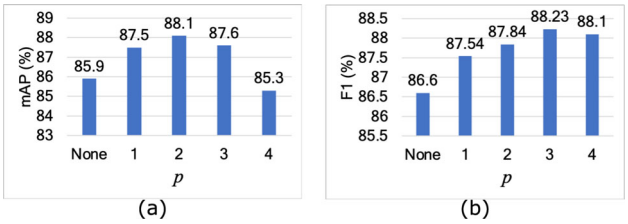

Fig. 5 Performance of attribute recognition by setting different values for p in the attention consistency between two networks

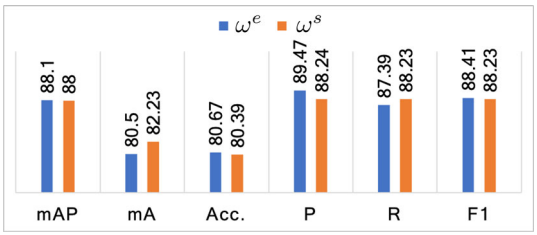

Fig. 6 Attribute recognition performance (mAP, %) by using different attribute weights in the classification loss on WIDER dataset, and mA, Acc. P, R, and F1 are reported on PA-100K

effectiveness of the proposed attention consistency between networks, but also its marginal complementarity to ensemble methods. In practice, considering the marginal performance improvement from model ensemble of the proposed method, we usually only need to use one of the branches for model inference for computation efficiency.

4.3.6 Hyper-Parameter Influence

Based on the consistency loss between two networks, we conduct experiments to investigate the influence of the power term p and show the recognition performance on two datasets in Fig. 5. Specifically, on WIDER dataset, we use ResNet101 as the main branch and ResNet50 as the auxiliary branch, while on PA-100K dataset, we use ResNet50 as the main branch and ResNet34 as the auxiliary branch. As shown in Fig. 5a, the best mAP performance on WIDER dataset (88.1%) is achieved by using $p = 2$, while an overly large power, e.g., $p = 4$, makes the consistency loss dominate the network learning, leading to reduced mAP performance. On PA-100K, there exists more severe data imbalance. A larger power term p in the attention consistency loss is desired to emphasize the pixel-wise difference between attention maps for the same attribute. As shown in Fig. 5b, the best F1 performance of attribute recognition on PA-100K is achieved by setting $p = 3$.

4.3.7 Attribute Weights in Classification Loss

Since the attribute weighting (Li et al. 2015; Tan et al. 2020) has been widely used in human attribute recognition, we regard them as a standard module in both our baselines and the proposed method. We specifically conduct an experiment to verify the effect of the attribute weighting in Eq. (3) for VAC-TE on WIDER dataset. We use the ResNet50 as the backbone of the proposed network and enforce the attention consistency of equivariance under image flipping (VAC-TEf). The results in Table 11 demonstrate that the proposed visual attention consistency improves the recognition performance either with or without the attribute weighting.

We adopt two attribute weighting strategies, as shown in Eqs. (3) and (4), respectively, on different datasets to fairly compare our method with prior works. We further compare the strategies on the same dataset in Fig. 6. The experiment results further verify that the proposed method is actually robust to both weighting strategies.

4.4 Qualitative Results

4.4.1 Qualitative Analysis

To qualitatively analyze the proposed method for attribute recognition, we visually compare the attention maps from the baseline ResNet101 without using attention consistency, and those enhanced with two kinds of attention consistency. As shown in Fig. 7, each row illustrates the attention maps for recognizing an attribute from the same image by different methods. Attention maps estimated by the baseline method without enforcing attention consistency may highlight visually irrelevant regions for certain attribute recognition, e.g., leg regions in recognizing the attribute

Table 11 Performance (%) of baseline model ResNet50 and model ResNet50-VAC-TEf with and without using attribute weighting strategy (on WIDER dataset)

Model	mAP	mA	F1-C	P-C	R-C	F1-O	P-O	R-O
ResNet50 with attribute weighting	83.4	82.0	73.9	**79.5**	69.4	79.4	82.3	76.6
ResNet50-VAC-TEf with attribute weighting	**86.3**	**84.5**	**76.4**	78.9	**74.3**	**81.2**	**82.6**	**79.8**
ResNet50 w/o attribute weighting	83.5	81.8	73.8	80.3	68.7	79.4	83.3	75.8
ResNet50-VAC-TEf w/o attribute weighting	**86.2**	**82.6**	**75.5**	**83.3**	**69.7**	**81.1**	**85.3**	**77.3**

Top performance for each metric is marked as bold

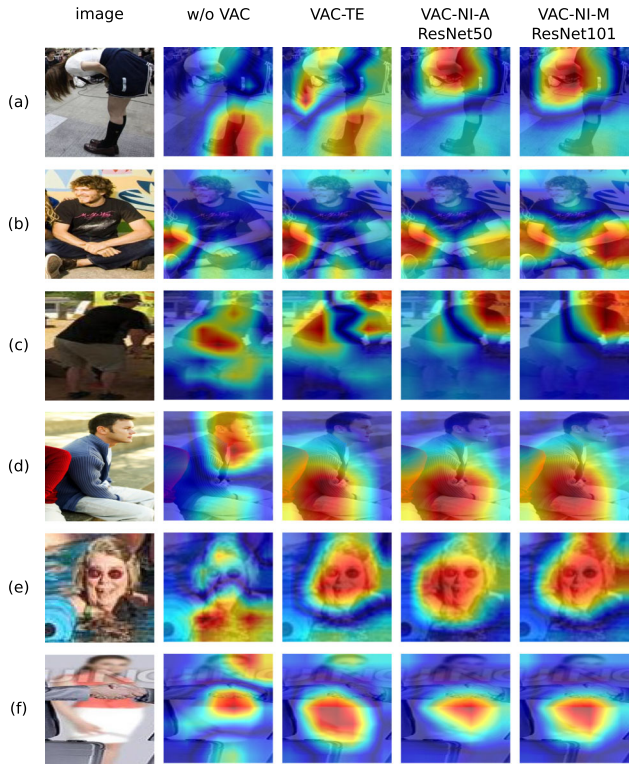

Fig. 7 Qualitative comparison of attention maps estimated in recognizing the same attribute (each row) by using different methods. The attributes to be recognized in each row are **a** T-shirt, **b** Jeans, **c** Hat **d** Long Pants, **e** Long Hair and **f** Skirt

"T-shirts" in the second column of row (a). When attention consistency of transform equivariance is adopted, the attention map is refined in the third column of row (a) by paying more attention on upper body. Furthermore, enforcing attention consistency between networks can also refine the attention maps by focusing attention on upper body as shown in the fourth and fifth columns of row (a) with ResNet50 and ResNet101 as backbones, respectively.

Moreover, for recognizing attributes "T-shirt", "Jeans" and "Hat", VAC-TE refines the corresponding attention maps, but may still miss/highlight some relevant regions/irrelevant regions for the attribute, e.g., highlighted leg regions for "T-shirt", missed left-leg regions for "Jeans" and highlighted waist regions for "Hat" in the third column of rows (a), (b) and (c), respectively. By contrast, VAC-NI-A and VAC-NI-M better highlight upper body, two legs and head regions for recognizing "T-shirt", "Jeans" and "Hat", respectively. This is aligned with the quantitative results which also show that VAC-NI achieves better performance than VAC-TE. For recognizing attributes "Long Pants", "Long Hair" and "Skirt" in rows (d), (e) and (f), respectively, enforcing either kind of attention consistency makes the corresponding attention maps to better highlight the correct image regions, e.g., leg, head and lower body. These qualitative results verify that the proposed two kinds of attention consistency can refine the visual attention map of networks in recognizing human attributes.

In addition to the above qualitative analysis, we have more qualitative results comparing the attention maps from baseline models and the proposed method, as shown in Fig. 8. Generally, for recognizing the same attribute from the same image, attention maps by enforcing consistency are usually more concentrated on the expected image regions, i.e., attribute relevant regions. These comparisons qualitatively

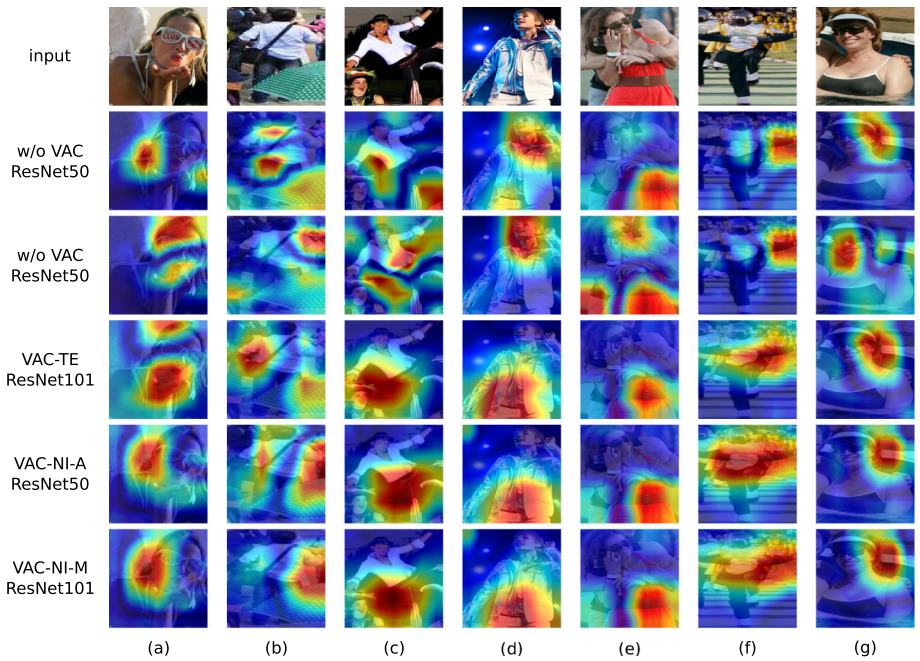

Fig. 8 More qualitative results by comparing the baselines (ResNet50 and ResNet101 without enforcing attention consistency) and the proposed method, including ResNet101 with VAC-TE, and VAC-NI with ResNet50 and ResNet101. Attributes to be recognized in each column: **a** Long Hair, **b** Long Sleeves, **c** Long Pants, **d** Long Pants, **e** Skirt, **f** Long Sleeves, and **g** Long Hair

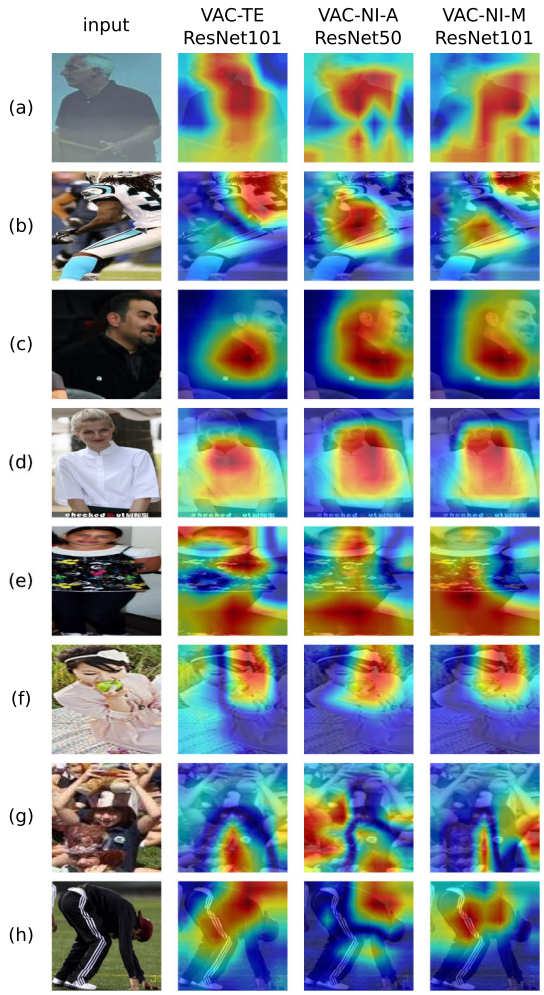

Fig. 9 Examples of failure cases when using the proposed visual attention consistency. Attributes to be recognized: **a** T-shirt, **b** Long Hair, **c** Long Hair, **d** Long Hair, **e** T-shirt, **f** Hat, **g** Long Pants, and **h** Hat

interpret the performance improvement from the proposed method.

4.4.2 Failure Cases

To further understand the limitations of the proposed method, we conduct experiments to locate several failure cases. As shown in Fig. 9, there are various challenges causing the proposed method to fail to predict the attribute. Respect to the cases shown from row (a) to (h) in Fig.9, such challenges can be (a) low image quality, (b) partial occlusion of the attribute, (c) context inference, (d) view angle, (e) image distortion after resizing, (f) insufficient robustness to distinguish similar objects, e.g., hair band vs. hat, (g) absent attribute-relevant regions, and (h) abnormal human pose. These challenges are very common in the task of human attribute recognition.

Specifically, in Fig. 9b, c, f, the proposed method almost discovers the correct attribute-specific image regions for recognition but still yields wrong predictions. This indicates that the human attribute recognition is still a very challenging task in real applications, such as video surveillance with complicated contexts.

5 Conclusion

In this paper, we proposed new methods to improve the plausibility of deep network attention maps to improve the performance of human attribute recognition. Specifically, we designed a two-branch framework to enforce the attention consistency during network learning for attribute recognition. In this framework, we formulated two kinds of attention consistency, i.e., equivariance under spatial transforms and invariance between different networks, and defined corresponding attention-consistency losses, which are combined with the initial classification loss for network learning. We conducted comprehensive experiments on three representative datasets for human attribute recognition and verified the effectiveness of enforcing attention consistency for attribute recognition by achieving new state-of-the-art performances on all these datasets.

References

Bansal, N., Agarwal, C., & Nguyen, A. (2020). SAM: The sensitivity of attribution methods to hyperparameters. In *IEEE conference on computer vision and pattern recognition* (pp. 8673–8683).

Bourdev, L., Maji, S., & Malik, J. (2011). Describing people: A poselet-based approach to attribute classification. In *IEEE international conference on computer vision* (pp. 1543–1550). IEEE.

Cohen, T., & Welling, M. (2016). Group equivariant convolutional networks. In *International conference on machine learning* (pp. 2990–2999).

Connor, C. E., Egeth, H. E., & Yantis, S. (2004). Visual attention: Bottom-up versus top-down. *Current Biology, 14*(19), R850–R852.

Dabkowski, P., & Gal, Y. (2017). Real time image saliency for black box classifiers. In *Advances in neural information processing systems* (pp. 6967–6976).

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection.

Deng, Y., Luo, P., Loy, C.C., & Tang, X. (2014). Pedestrian attribute recognition at far distance. In *ACM International conference on multimedia* (pp. 789–792). ACM.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience, 18*(1), 193–222.

Dieleman, S., De Fauw, J., & Kavukcuoglu, K. (2016). Exploiting cyclic symmetry in convolutional neural networks. arXiv preprint arXiv:1602.02660

Eriksen, C. W., & Hoffman, J. E. (1972). Temporal and spatial characteristics of selective encoding from visual displays. *Perception & Psychophysics, 12*(2), 201–204.

Feris, R., Bobbitt, R., Brown, L., & Pankanti, S. (2014). Attribute-based people search: Lessons learnt from a practical surveillance system. In *International conference on multimedia retrieval* (pp. 153–160).

Fong, R. C., & Vedaldi, A. (2017). Interpretable explanations of black boxes by meaningful perturbation. In *IEEE international conference on computer vision* (pp. 3429–3437).

Gkioxari, G., Girshick, R., & Malik, J. (2015). Contextual action recognition with r* CNN. In *IEEE international conference on computer vision* (pp. 1080–1088).

Guo, H., Fan, X., & Wang, S. (2017). Human attribute recognition by refining attention heat map. *Pattern Recognition Letters, 94*, 38–45.

Guo, H., Zheng, K., Fan, X., Yu, H., & Wang, S. (2019). Visual attention consistency under image transforms for multi-label image classification. In *IEEE conference on computer vision and pattern recognition* (pp. 729–739).

Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., & Sugiyama, M. (2018). Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in neural information processing Systems* (pp. 8527–8537).

Han, K., Guo, J., Zhang, C., & Zhu, M. (2018). Attribute-aware attention model for fine-grained representation learning. In *ACM international conference on multimedia* (pp. 2040–2048).

Han, K., Wang, Y., Shu, H., Liu, C., Xu, C., & Xu, C. (2019). Attribute aware pooling for pedestrian attribute recognition. arXiv preprint arXiv:1907.11837

He, K., Zhang, X., Ren, S., & Sun, J. (2016) Deep residual learning for image recognition. In *IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531

Hinton, G. E., Krizhevsky, A., & Wang, S. D. (2011). Transforming auto-encoders. *International conference on artificial neural networks* (pp. 44–51). Springer.

Hu, J., Shen, L., & Sun, G. (2017). Squeeze-and-excitation networks, *7*. arXiv preprint arXiv:1709.01507

Huang, G., Liu, Z., Weinberger, K.Q., & van der Maaten, L. (2017). Densely connected convolutional networks. In *IEEE conference on computer vision and pattern recognition* (Vol. 1, p. 3).

Jaderberg, M., Simonyan, K., Zisserman, A., & Kavukcuoglu, K. (2015). Spatial transformer networks. In *Advances in neural information processing systems* (pp. 2017–2025).

Kivinen, J. J., & Williams, C. K. (2011). Transformation equivariant Boltzmann machines. *International conference on artificial neural networks* (pp. 1–9). Springer.

Koch, C., & Ullman, S. (1987). Shifts in selective visual attention: Towards the underlying neural circuitry. *Matters of intelligence* (pp. 115–141). Springer.

Koch, K., McLean, J., Segev, R., Freed, M. A., Berry, M. J., II., Balasubramanian, V., & Sterling, P. (2006). How much the eye tells the brain. *Current Biology, 16*(14), 1428–1434.

Krizhevsky, A., Sutskever, I., & Hinton, G.E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Lavie, N. (2005). Distracted and confused?: Selective attention under load. *Trends in Cognitive Sciences, 9*(2), 75–82.

Lenc, K., & Vedaldi, A. (2015). Understanding image representations by measuring their equivariance and equivalence. In *IEEE conference on computer vision and pattern recognition* (pp. 991–999).

Lenc, K., & Vedaldi, A. (2016). Learning covariant feature detectors. *European conference on computer vision* (pp. 100–117). Springer.

Li, D., Chen, X., & Huang, K. (2015). Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In *Asian conference on pattern recognition* (pp. 111–115). IEEE.

Li, D., Chen, X., Zhang, Z., & Huang, K. (2018). Pose guided deep model for pedestrian attribute recognition in surveillance scenarios. In *International conference on multimedia and expo* (pp. 1–6). IEEE.

Li, D., Zhang, Z., Chen, X., Ling, H., & Huang, K. (2016). A richly annotated dataset for pedestrian attribute recognition. arXiv preprint arXiv:1603.07054

Li, Q., Zhao, X., He, R., & Huang, K. (2019). Visual-semantic graph reasoning for pedestrian attribute recognition. In *AAAI conference on artificial intelligence* (Vol. 33, pp. 8634–8641).

Li, Y., Huang, C., Loy, C. C., & Tang, X. (2016). Human attribute recognition by deep hierarchical contexts. In *European conference on computer vision* (pp. 684–700). Springer.

Lin, Y., Zheng, L., Zheng, Z., Wu, Y., Hu, Z., Yan, C., & Yang, Y. (2019). Improving person re-identification by attribute and identity learning. *Pattern Recognition, 95*, 151–161.

Liu, P., Liu, X., Yan, J., & Shao, J. (2018). Localization guided learning for pedestrian attribute recognition. arXiv preprint arXiv:1808.09102

Liu, X., Zhao, H., Tian, M., Sheng, L., Shao, J., Yan, J., & Wang, X. (2017). Hydraplus-net: Attentive deep features for pedestrian analysis. In *IEEE international conference on computer vision* (pp. 1–9).

Malach, E., & Shalev-Shwartz, S. (2017). Decoupling "when to update" from "how to update". In *Advances in neural information processing systems* (pp. 960–970).

Marcos, D., Volpi, M., Komodakis, N., & Tuia, D. (2017). Rotation equivariant vector field networks. In *IEEE international conference on computer vision* (pp. 5048–5057).

Moran, J., & Desimone, R. (1985). Selective attention gates visual processing in the extrastriate cortex. *Science, 229*(4715), 782–784.

Müller, R., Kornblith, S., & Hinton, G. E. (2019). When does label smoothing help? In *Advances in neural information processing systems* (pp. 4694–4703).

Niu, X., Han, H., Shan, S., & Chen, X. (2019). Multi-label co-regularization for semi-supervised facial action unit recognition. In *Advances in neural information processing systems* (pp. 909–919).

Oquab, M., Bottou, L., Laptev, I., & Sivic, J. (2015). Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *IEEE conference on computer vision and pattern recognition* (pp. 685–694).

Qiao, S., Shen, W., Zhang, Z., Wang, B., & Yuille, A. (2018). Deep co-training for semi-supervised image recognition. In *European conference on computer vision* (pp. 135–152).

Ravanbakhsh, S., Schneider, J., & Poczos, B. (2017). Equivariance through parameter-sharing. In *International conference on machine learning* (pp. 2892–2901). JMLR.org.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" explaining the predictions of any classifier. In *ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144).

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. In *AAAI conference on artificial intelligence*.

Sarafianos, N., Xu, X., & Kakadiaris, I. A. (2018). Deep imbalanced attribute classification using visual attention aggregation. arXiv preprint arXiv:1807.03903

Sarfraz, M. S., Schumann, A., Wang, Y., & Stiefelhagen, R. (2017). Deep view-sensitive pedestrian attribute inference in an end-to-end model. arXiv preprint arXiv:1707.06089

Schmidt, U., & Roth, S. (2012). Learning rotation-aware features: From invariant priors to equivariant descriptors. In *IEEE conference on computer vision and pattern recognition* (pp. 2050–2057). IEEE.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE international conference on computer vision* (pp. 618–626).

Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. arXiv preprint arXiv:1704.02685

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556

Su, C., Zhang, S., Xing, J., Gao, W., & Tian, Q. (2016). Deep attributes driven multi-camera person re-identification. In *European conference on computer vision* (pp. 475–491). Springer.

Sudowe, P., Spitzer, H., & Leibe, B. (2015). Person attribute recognition with a jointly-trained holistic CNN model. In *IEEE international conference on computer vision workshops* (pp. 87–95).

Sun, G., Khan, S., Li, W., Cholakkal, H., Khan, F., & Van Gool, L. (2020). Fixing localization errors to improve image classification. In *European conference on computer vision*.

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. arXiv preprint arXiv:1703.01365

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. In *IEEE conference on computer vision and pattern recognition*. IEEE.

Tan, Z., Yang, Y., Wan, J., Guo, G., & Li, S. Z. (2020). Relation-aware pedestrian attribute recognition with graph convolutional networks. In *AAAI conference on artificial intelligence* (pp. 12055–12062).

Tan, Z., Yang, Y., Wan, J., Hang, H., Guo, G., & Li, S. Z. (2019). Attention-based pedestrian attribute analysis. *IEEE Transactions on Image Processing, 28*(12), 6126–6140.

Tang, C., Sheng, L., Zhang, Z., & Hu, X. (2019). Improving pedestrian attribute recognition with weakly-supervised multi-scale attribute-specific localization. In *IEEE international conference on computer vision* (pp. 4997–5006).

Tarvainen, A., & Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in neural information processing systems* (pp. 1195–1204).

Thewlis, J., Bilen, H., & Vedaldi, A. (2017). Unsupervised learning of object frames by dense equivariant image labelling. In *Advances in neural information processing systems* (pp. 844–855).

Thewlis, J., Bilen, H., & Vedaldi, A. (2017). Unsupervised learning of object landmarks by factorized spatial embeddings. In *IEEE international conference on computer vision* (pp. 5916–5925).

Tian, Y., Luo, P., Wang, X., & Tang, X. (2015). Pedestrian detection aided by deep learning semantic tasks. In *IEEE conference on computer vision and pattern recognition* (pp. 5079–5087).

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology, 12*(1), 97–136.

Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., & Tang, X. (2017). Residual attention network for image classification. In *IEEE conference on computer vision and pattern recognition* (pp. 3156–3164).

Wang, J., Yang, Y., Mao, J., Huang, Z., Huang, C., & Xu, W. (2016). CNN-RNN: A unified framework for multi-label image classification. In *IEEE conference on computer vision and pattern recognition* (pp. 2285–2294). IEEE.

Wang, J., Zhu, X., Gong, S., & Li, W. (2017). Attribute recognition by joint recurrent learning of context and correlation. In *IEEE international conference on computer vision* (pp. 531–540).

Wang, X., Zheng, S., Yang, R., Luo, B., & Tang, J. (2019). Pedestrian attribute recognition: A survey. arXiv preprint arXiv:1901.07474

Woo, S., Park, J., Lee, J. Y., & So Kweon, I. (2018). CBAM: Convolutional block attention module. In *European conference on computer vision* (pp. 3–19).

Worrall, D., & Brostow, G. (2018). CubeNet: Equivariance to 3d rotation and translation. In *European conference on computer vision* (pp. 567–584).

Worrall, D. E., Garbin, S. J., Turmukhambetov, D., & Brostow, G. J. (2017). Harmonic networks: Deep translation and rotation equivariance. In *IEEE conference on computer vision and pattern recognition* (pp. 5028–5037).

Wu, M., Huang, D., Guo, Y., & Wang, Y. (2020). Distraction-aware feature learning for human attribute recognition via coarse-to-fine attention mechanism. In *AAAI conference on artificial intelligence* (pp. 12394–12401).

Zagoruyko, S., & Komodakis, N. (2016). Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. arXiv preprint arXiv:1612.03928

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision* (pp. 818–833). Springer.

Zhang, N., Paluri, M., Ranzato, M., Darrell, T., & Bourdev, L. (2014). Panda: Pose aligned networks for deep attribute modeling. In *IEEE conference on computer vision and pattern recognition* (pp. 1637–1644).

Zhang, Y., Xiang, T., Hospedales, T. M., & Lu, H. (2018). Deep mutual learning. In *IEEE conference on computer vision and pattern recognition* (pp. 4320–4328).

Zhao, X., Sang, L., Ding, G., Guo, Y., & Jin, X. (2018). Grouping attribute recognition for pedestrian with joint recurrent learning. In *International joint conferences on artificial intelligence* (pp. 3177–3183).

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In *IEEE conference on computer vision and pattern recognition* (pp. 2921–2929). IEEE.

Zhou, Z. H., Wu, J., & Tang, W. (2002). Ensembling neural networks: Many could be better than all. *Artificial Intelligence, 137*(1–2), 239–263.

Zhu, F., Li, H., Ouyang, W., Yu, N., & Wang, X. (2017). Learning spatial regularization with image-level supervisions for multi-label image classification. In *IEEE conference on computer vision and pattern recognition* (pp. 5513–5522).

Zhu, J., Liao, S., Lei, Z., Yi, D., & Li, S. (2013). Pedestrian attribute classification in surveillance: Database and evaluation. In *IEEE international conference on computer vision workshops* (pp. 331–338).

Zintgraf, L. M., Cohen, T. S., Adel, T., & Welling, M. (2017). Visualizing deep neural network decisions: Prediction difference analysis. arXiv preprint arXiv:1702.04595