Multi-task Relative Attributes Prediction by Incorporating Local Context and Global Style Information Features

Yuhang He
www.heyuhang.com
Long Chen*
www.carlib.net
Jianda Chen
www.carlib.net

School of Data and Computer Science
Sun Yat-sen University
Guangzhou, P.R China

Abstract

Relative attribute represents the correlation degree of one attribute between an image pair (*e.g.* one car image has more seat number than the other car image). While appearance highly and directly correlated relative attribute is easy to predict, fine-grained or appearance insensitive relative attribute prediction still remains as a challenging task. To address this challenge, we propose a multi-task trainable deep neural networks by incorporating an object's both local context and global style information to infer the relative attribute. In particular, we leverage convolutional neural networks (CNNs) to extract feature, followed by a ranking network to score the image pair. In CNNs, we treat features arising from intermediate convolution layers and full connection layers in CNNs as local context and global style information, respectively. Our intuition is that local context corresponds to bottom-to-top localised visual difference and global style information records high-level global subtle difference from a top-to-bottom scope between an image pair. We concatenate them together to escalate overall performance of multi-task relative attribute prediction. Finally, experimental results on 5 publicly available datasets demonstrate that our proposed approach outperforms several other state-of-the-art methods and further achieves comparable results when comparing to very deep networks, like 152-ResNet [19] and inception-v3 [8].

1 Introduction

Attributes, a genre of human observable mid-level semantic property in an image, have received much attention in recent years [13][1]. It describes an image in a more human understandable and preferable way. For instance, object, scene and action form an attribute triple to describe an image semantically [2]. However, conventional visual attribute easily collapses into categorical or binary since it merely indicates the presence/absence of one attribute or which attribute an image part belongs to, few extra information goes beyond that. It works well when handling simple tasks like classification, detection or recognition [25] but it also easily falls short if more complex tasks or realistic applications come into mind. For

*Dr. Long Chen is the corresponding author.

Pages 131.1-131.12
DOI: https://dx.doi.org/10.5244/C.30.131

example, seldom are consumers interested in the categorical property of each shoe in Fig.1 left, instead, they pick out the shoe to buy by comparing various attributes among shoe pairs.

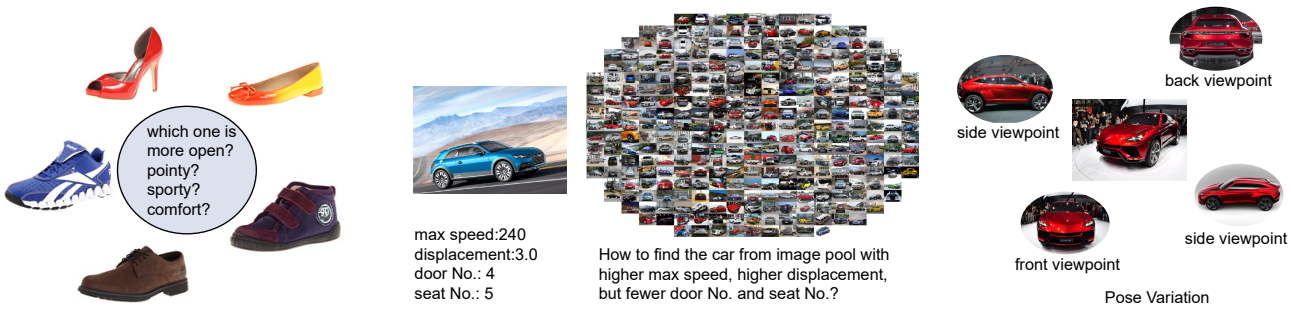

Figure 1: Left: among a bunch of shoe images, the "pointy" and "open" attributes are much more easier to determine than "comfort" and "sporty" attributes. Middle: an image may contain multiple attributes, these attributes are mutually intertwined and need to be compared in a multi-task manner. Right: an object may represent various poses or viewpoints in an image, in this case, local context alone often cannot handle this variation.

To fulfil those heterogenous requirements and diversify attribute properties, relative attribute was introduced [36][11], it strides across the barrier of humdrum presence/absence or yes/no description of an attribute to an entity (one-vs-one) and has dramatically enriched attribute pool, enabling attributes to be meaningful in cross-entity domain and to mine more meaningful information both mutually and individually. Relative attributes have been successfully applied to a variety of tasks, including online image search and retrieval [3], zero-shot learning [7][11]. Moreover, relative attribute predication is ubiquitous in our daily life, an obvious example of this is that, if a customer owes a car with known attributes, he wants to buy a new car with higher maximum speed, higher displacement, but low door number and seat number. How can he directly retrieve the appropriate car from the car image pool? (see Fig.1 Middle)

Current methods on relative attribute prediction almost unanimously follow the pipeline: feature extraction and image ranking with these features. For feature extraction, hand-crafted and engineered features, such as HOG [10] and GIST [4], as well as more discriminative features learned by Convolutional Neural Networks (CNNs) are incorporated. As for the ranking, D. Parish and K. Grauman [11] applied zero-short learning to learn a global linear ranking function for each attribute. S. Li *et al*. [27] trained hierarchical rankers with non-linear functions. In [34], Y. Souri *et al*. combined convolutional neural networks (CNNs) and RankNet [6] to directly rank each attribute individually. These approaches, however, build on two main assumptions. The first one is that relative attribute has inherent correlations with image's visual appearance and visual difference between an image pair trustworthily corresponds to relative attribute strength. The visually difference can be perfectly modelled by either hand-crafted features [11][36] or single-path CNNs. The second one is that all relative attributes are individually independent. They did not take intra relative attributes mutual effects into account. We argue that relative attributes are heavily influenced by each other and they can be multi-task behaviourally modelled and predicted. An obvious example is that higher maximum speed attribute car often occupies higher displacement attribute, and further, these two attributes are reflected by the visual appearance of seat number and door number in the car image (see Fig.1 Middle). The first assumption works well on simple dataset, but often fails to accommodate cases in which more abstract or appearance insensitive attributes are involved. For example, though it is easy to infer which shoe is more

"pointy" than the other one because "pointy" feature mostly localises around shoe's heel, we can hardly specify which part of the query image determines "comfort" "sporty" relative attribute strength since they virtually invisible and require analysis over the whole image (see Fig. 1 Left). What's more, visual appearance is not always reliable, pose variation or viewpoint difference often generate different visual appearance for the same attributes. Thus, a more robust way to address this problem is to involve more powerful feature taking both local and global information into account, and harness the mutual influence between different attributes.

Therefore, we propose a multi-task relative attribute prediction framework which incorporates both local context and global style information. Our intuition is that the relative attributes emerge from both local context and image global information (we call it global style information), both of them can be used to assist relative attribute prediction. We leverage both intermediate layers (local context) and full connection layers (global style information), and concatenate them together to form final image feature vector. Current work [5][14][37][12][30][26] demonstrate that features arising from intermediate CNNs layers escalate various tasks at a large scale by simply concatenating them together [5], or reshaping them to an uniform shape [14] or encoding them via other methods [37][26], like VLAD [17]. S. Liu *et al*. [28] and F. Yang *et al*. [12] show that intermediate layers delineate mid-level feature (motif, object, semantics), while later layers represent high-level and more abstract features. We follow these two arguments and propose a special CNNs architecture, which automatically aggregates features from both intermediate layers and final layers to capture both local context and global style information to improve relative attribute prediction. After feature representation, another ranking network is grafted to CNNs to directly rank each attribute. What's more, we take intra effect among various relative attributes into consideration and formulate our framework to model multi-task relative attributes simultaneously. Currently, harnessing deep learning methods to achieve multi-task attributes learning have been successfully explored [29][15][21]. We build on them to embrace CNNs to learn multi-task relative attribute prediction.

The main contributions of this paper lie in 1) we construct a multi-task trainable deep neural networks to formulate relative attribute prediction problem, which is able to predict multiple relative attributes at the same time. 2) we embrace both local context and global style information to infer relative attributes, and take the advantage of different layers of CNNs to extract features. 3) Our proposed methods achieve impressive result in both appearance sensitive and fine-grained relative attribute prediction tasks, comparing with very deep neural networks.

2 Relative Attribute Prediction via Ranking

Our proposed framework follows the pipeline shown in Fig. 2. Given a set of image pairs $\{\{I_1^i, I_2^i\}|i \in \{1,2,\cdots,n\}\}$, and their corresponding relative attributes label $\{l_k^i|i \in \{1,2,\cdots,n\}, k \in \{1,2,\cdots,K\}\}$, where $l_k^i = \{0, 0.5, 1.0\}$ indicates the strength of relative attribute a_k of image I_1^i over image I_2^i. We designate 0, 0.5 and 1.0 to be *less*, *equal*, *more* of the relative attribute, respectively. Suppose that we have trained an image feature extractor f, f converts image pairs set into feature vector sets $\{\{\psi_1^i, \psi_2^i\}|i \in \{1,2,\cdots,n\}\}$, our goal is to find a ranking function $\mathcal{R}$ which ranks feature vectors according the input labels

$$\mathcal{R}(I_1^i) > \mathcal{R}(I_2^i) \;\; if \; l_k^i = 1; \quad \mathcal{R}(I_1^i) = \mathcal{R}(I_2^i) \;\; if \; l_k^i = 0.5; \quad \mathcal{R}(I_1^i) < \mathcal{R}(I_2^i) \;\; if \; l_k^i = 0 \qquad (1)$$

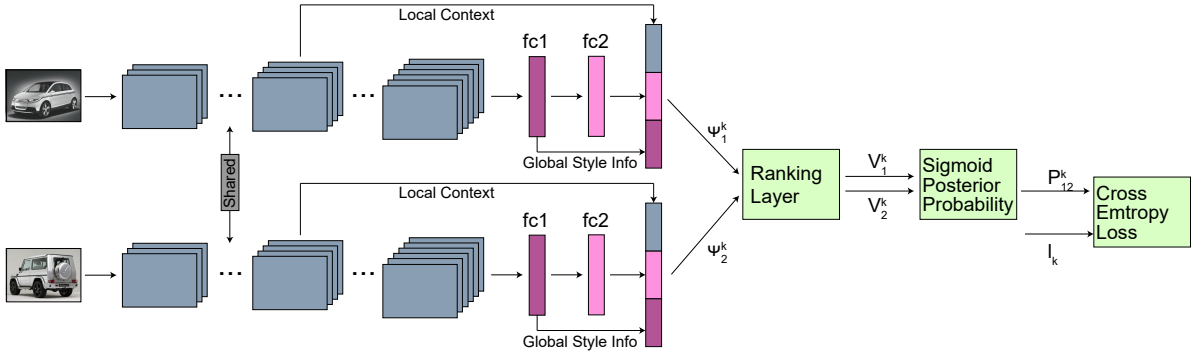

Figure 2: The pipeline of our proposed framework: we feed the image pair to two CNNs with the same network architecture and shared parameters. Features learned by CNNs intermediate layers and final several full connection layers, which are deemed to store global context and global style information, respectively, are concatenated together to form the final feature. The feature pair is further fed to the ranking layer to score each attribute.

In order to make our model end-to-end trainable, we leverage the ranking network proposed in [6] because it can be trained with gradient descent approach. It directly maps the input feature vector to k real value pairs $\{(v_1^i, v_2^i)|i \in \{1,2,\cdots,K\}\}$, each of which corresponds to a relative attribute. We calculate the posterior probability for each relative attribute and squash it between $[0-1]$ via a sigmoid function.

$$P_{1,2}^k = \frac{1}{1+e^{-(v_1^i - v_2^i)}} \tag{2}$$

Finally, we utilise cross entropy loss to rank each relative attribute as follows

$$\mathcal{L}_i = \sum_{k=1}^{K} l_k^i \log(P_{1,2}^k) - (1-l_k^i)\log(-P_{1,2}^k) \tag{3}$$

Note that by minimising the cross entropy loss in Eqn.3, we can force the whole neural network to learn discriminative image features so that they are compatible with the given multiple relative attributes ordering. Note that the feature vector learned by CNNs is mapped to a real value (for single relative attribute prediction) or a real value vector (for multi-task relative attribute prediction) through a mapping matrix W and a bias term b: $v = W \cdot \psi + b$.

3 Feature Learning by incorporating Local Context and Global Style Information

We embrace the power of CNNs to learn features as they have shown state-of-the-art performance on various vision related tasks [16][9]. CNNs excel at learning discriminative features at various granularities. Current researches in CNNs related vision tasks either try to build much deeper neural networks as various experiments on large public datasets shows that deeper network enables to model to learn more robust and discriminative features [19][18], yet it is time-consuming and requires multiple powerful machines to train the model, or fully exploit intermediate layer features to improve overall performance [5][14][37]. Rather than delving into very deep neural networks, we are more interested in improving relative attribute prediction accuracy merely from relatively shallow neural networks but leverage both local

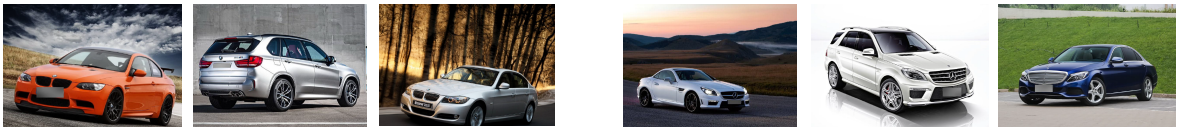

Figure 3: Style information of BMW and Mercedes-Benz. While BMW emphasizes on sporty elegance and dynamics, Mercedes-benz highlights aerodynamic principles and streamlined outlook. We call these abstract and manufacturer unique characteristic global style information.

context and global style information carried by the neural network. This is desirable because it is both time and training hardware cost-effective.

As CNNs learn feature through a coarse-to-fine or concrete-to-abstract process, we can safely assume that while the earliest layers capture low-level and basic features (*e.g.* edges, contours or texture), intermediate layers records mid-level feature, such as motifs, object and semantics, the last few layers are responsible for extracting high-level, global and abstract information [5][37][12]. We build on these theories and propose to combine features from intermediate layers and last two full connection layers. So intermediate layers and the last few layers capture local context and global style information, respectively. Our intuition is that relative attributes stem from both local context and global style information. Local context is bottom-to-top and corresponds to appearance sensitive attributes. For example, we can easily decide which shoe is "pointy" than the other shoe simply by its hell height metric. Global style information is top-to-bottom and corresponds to high-level semantics, which escalates fine-grained or large pose variational relative attribute prediction. Besides, global style information is prevalent in product manufacturing industry. Each manufacturer has its own unique design philosophy, which is blended into its product's sophisticated design feature and overall outlook. For example, automobiles designed from BMW look sportily elegant and dynamic, this character is manifested by automobiles' sophisticated overall modelled surface, like short overhangs, long bonnet and wheelbase. On the contrary, Mercedes-benz highlights aerodynamic principles and streamlined elements in its automobile (See Fig.3). We call these abstract global semantics as "style information" and it has strong correlation with internal relative attributes. Experts can easily decide what type of the car it is after taking a glance at the car's image and further, and compare their relative attributes among a pool of car pairs. We utilise this property to incorporate "global style" expert-like information into our process. Thus, our feature learning process works as

$$\psi^i = \psi^i_{fc} + \psi^i_{local} + \psi^i_{global} \tag{4}$$

where fc indicates the last full connection layer. ψ^i indicates the final concatenated learned feature vector. ψ^i_{global} indicates the penultimate full connections layer and ψ^i_{local} indicates the intermediate convolution layer. We extract these features and reshape them into a vector, and then concatenate them together to form one whole vector, serving as the final image feature. The final feature vector is further fed to ranking layer for relative attribute ranking by mapping the feature vector to a real value (or a real value vector) with a matrix W and bias b.

In general, as shown in Fig. 2, given an image pair, we train them with two CNNs architectures. The two CNNs architectures share the same structures as well as parameters. The

Attributes	Max. Speed						Displacement					
Viewpoint	NoView	F	R	S	FS	RS	No	F	R	S	FS	RS
ReAtr[11]	43.22	47.33	45.33	48.01	46.37	46.20	40.39	40.82	40.11	43.09	43.22	40.10
LoLe[36]	46.32	47.74	46.11	47.68	47.99	47.00	42.77	43.11	43.00	44.09	43.98	43.11
DepRe[34]	52.33	54.27	51.28	55.01	52.09	51.07	58.33	49.32	48.27	49.57	48.97	48.30
NaPa[35]	42.13	42.78	42.10	41.07	45.78	44.23	42.78	43.11	43.90	43.27	43.17	43.91
152-ResNet[19]	55.14	57.28	56.10	56.78	58.22	57.77	49.10	49.01	47.82	49.92	49.72	48.00
inception-v3[8]	54.19	56.29	54.78	57.20	57.38	57.00	49.31	49.57	48.87	49.37	49.27	48.38
Ours_indiv	46.78	46.92	43.27	50.33	48.00	48.71	44.20	43.82	44.78	42.10	41.10	44.78
Ours_none	51.20	53.18	52.77	52.78	49.03	48.77	46.78	47.23	46.32	47.07	47.00	46.69
Ours_local	52.13	53.11	51.34	53.07	52.01	52.00	48.32	48.10	47.29	48.01	47.77	47.38
Ours_global	53.07	54.78	53.10	54.09	54.22	53.00	48.77	48.97	47.88	47.79	47.29	48.00
Ours_total	**55.07**	**56.26**	**55.78**	**56.34**	**55.78**	**55.21**	**50.01**	**49.79**	**49.80**	**50.32**	**50.37**	**50.11**
Attributes	Door Number						Seat Number					
Viewpoint	NoView	F	R	S	FS	RS	No	F	R	S	FS	RS
ReAtr[11]	67.72	53.20	50.00	88.90	86.71	84.20	60.38	62.33	63.74	63.37	61.00	59.78
LoLe[36]	66.32	50.19	49.32	87.32	84.28	83.77	59.11	61.22	63.00	63.12	61.00	58.78
DeepRe[34]	77.19	60.32	60.47	93.01	87.77	85.78	75.00	69.38	67.29	87.00	83.00	88.27
NaPa[35]	69.00	51.22	51.01	86.78	84.04	85.99	58.99	60.18	62.99	60.98	63.00	60.00
152-ResNet[19]	78.10	62.22	61.89	94.37	89.76	88.39	78.22	73.00	68.88	89.00	85.32	89.90
inception-v3[8]	77.79	61.10	60.88	94.00	88.72	87.35	79.00	74.00	69.11	88.88	84.34	87.78
Ours_indiv	75.22	61.11	51.00	90.00	88.27	85.43	76.01	72.89	65.19	82.11	84.10	85.18
Ours_none	75.31	61.00	52.11	91.00	88.27	85.34	76.66	73.10	65.22	84.00	84.11	85.29
Ours_local	75.99	62.10	54.09	91.87	88.67	86.56	76.32	72.99	66.89	85.90	85.00	86.04
Ours_global	76.10	62.22	53.10	92.11	88.33	86.78	76.80	73.28	65.78	84.99	84.33	86.00
Ours_total	**82.41**	**70.01**	**63.10**	**94.11**	**89.00**	**87.22**	**80.14**	**73.90**	**65.78**	**86.98**	**85.40**	**89.30**

Table 1: Result on CompCar [22] dataset. (%)

learned features, containing local context and global style information, are further put into ranking layer to score all relative attributes.

4 Experiment

We implement our method in open-source Caffe [33] deep learning framework. To augment train data, we apply various data augmentation methods, such as vignetting, casting, random crop and rotation to get more training data. We deploy Microsoft 34-layer residual network with identity mapping [19] and have pre-trained it on ILSVRC 2014 dataset [23]. During training, we use stochastic gradient descent with RMSProb [31] to fine-tune the parameters w.r.t. different relative attributes prediction tasks. Besides, we set the minibatch size as 64 and the learning rate as 10^{-2}. We also apply l_2 regularisation and weight w and bias b are initialised by Xavier [32] and 0, respectively. All models are trained on 4 Nvidia Titan X.

To quantitatively and comprehensively evaluate our proposed method, we conduct experiment on five publicly available datasets: CompCar [22], Zappos50K [36], LFW-10 [24], PubFig and OSR [11]. We mostly focus on CompCar [22] dataset because each car image contains 4 attributes: maximum speed, displacement, door number and seat number. Besides, extra metadata, including five viewpoints: front (F), rear (R), side (S), front-side (FS) and rear-side (RS) viewpoint. This allows us to model multi-task relative attribute prediction and further investigate the impact of pose and viewpoint variation on relative attributes. Zappos50K [36] contains commercial shoes for both coarse and fine-grained relative attribute comparison, each of them is associated with one or more attributes from a total of 4 attributes: open, comfort, pointy and sporty. LFW-10 [24] contains 2,000 face images and 10 attributes are available for comparison (note that not every image associates with 10 attributes). PubFig [11] consists of 800 face images and OSR [11] consists of 2,688 outdoor scene images. We report mean accuracy of the percentage of correctly ordered image pairs as the evaluation metric, as adopted by other relevant methods [11][36][24][34][35]. To validate the involvement of local context and global style information indeed improves relative attribute prediction, we compare our algorithm from four perspectives: without local context or global style information (Ours_none), with only local context (Ours_local), with only global style infor-

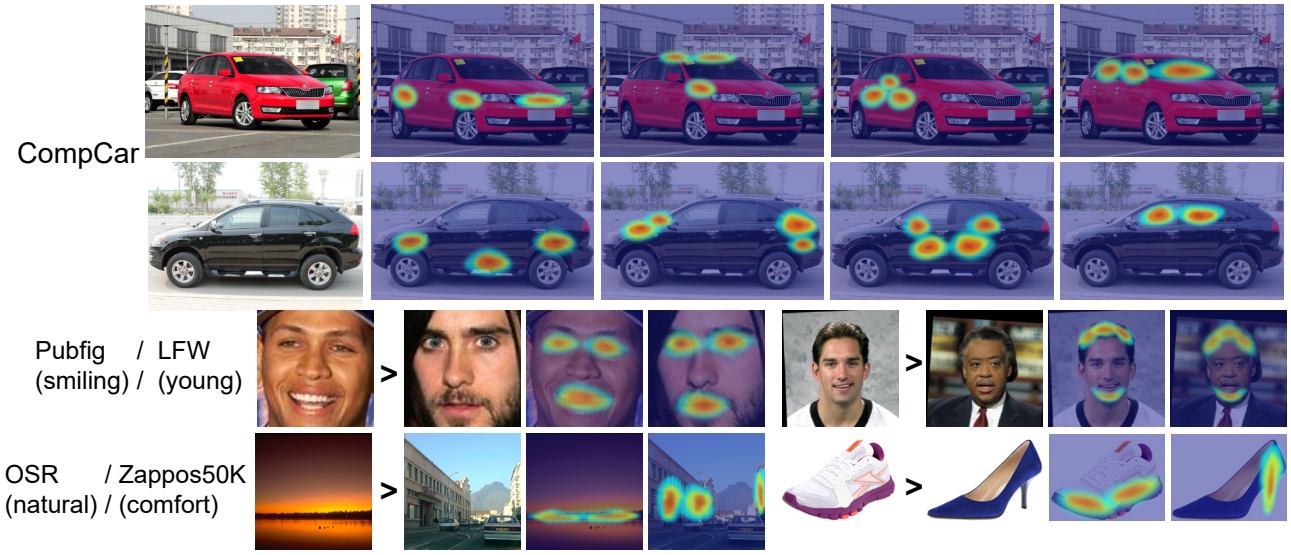

Figure 4: Experimental result samples on 5 datasets and their corresponding saliency map. In CompCar dataset, the four saliency maps are maximum speed, displacement, door number and seat number, respectively. (from left to right). Note that the black car shows more max. speed, less displacement and equal door number and seat number than the red car.

mation (Ours_global) and with both local context and global style information (Ours_total). We compare our algorithm with four other algorithms: hand-crafted feature based methods: ReAtr [11], LoLe [36] and NaPa [35], and CNNs based methods: DepRe [34]. Recent study shows that deeper network learns more discriminative feature [19][18]. In order to test whether our proposed framework enables shallows neural network to achieve comparable result when comparing to very deep neural network by incorporating both local context and global style information presented in this paper, we trained another two very deep neural networks that currently have achieved state-of-the-art performance in various vision tasks: 152 residual network (152-ResNet) [19] and google inception-v3 [8], for fair comparison. 152-ResNet with identity mapping [18] enables earlier learned localised features to directly flow to any deep layer, and inception-v3 percepts an image at various scales. Therefore, we assume both of them automatically learn local context and global style information.

CompCar [22] dataset used here contains a total of 136,727 web-nature images, covering 161 car makers and 1,687 car models. We conducted two experiments in CompCar dataset. In the first one, we neglect the viewpoint variation and has randomly created 46,869 image pairs. We split it into 38,000 for train, 2,000 for validation and 6,869 for testing, respectively. In the second experiment, we take viewpoint variation into consideration and separately conduct experiment for front view (F), rear view (R), side view (S), font-side view (FS) and rear-side view (RS), respectively. Besides, in order to test whether our proposed multi-task framework outperforms modelling each relative attribute separately, we have also separately trained models to predict each relative attribute individually (Ours_indiv). We fine-tune with Microsoft 34-layer residual network pre-trained on ILSVRC 2014 dataset [23] with identity mapping [19][18]. The accuracy result is reported in Table 1. We can see that CNNs based methods [34][19][8], including our proposed method, outperform hand-crafted feature based methods ReAtr [11], LolLe [36] and NaPa [35] by a large margin, demonstrating the powerful CNNs to learn discriminative feature for relative attribute prediction. Relative attributes "maximum speed" and "displacement" suffer most in relative attribute separately-trained situation, with an average of 10 percent prediction accuracy less than multi-task training

Dataset	Zappos50K-1					Zappos50K-2				
Attribute	Open	Sporty	Comfort	Pointy	Mean	Open	Sporty	Comfort	Pointy	Mean
ReAtr[11]	87.77	91.20	89.93	89.37	89.57	60.18	62.70	64.04	59.56	61.62
LoLe[36]	90.67	92.67	92.37	90.83	91.64	71.91	64.54	62.51	63.74	66.43
DeepRe[34]	93.00	95.56	93.22	92.11	93.47	71.21	67.81	65.88	66.64	67.88
Napa[35]	63.10	61.55	62.33	60.13	61.78	76.20	64.80	63.60	65.30	67.50
152-ResNet[19]	95.14	95.77	93.74	93.11	94.44	72.22	69.34	66.34	69.39	70.33
inception-v3[8]	95.33	96.30	94.01	92.00	94.41	71.22	68.99	67.03	68.32	68.42
Ours_none	93.10	95.07	93.55	93.12	93.71	71.30	68.03	69.72	66.98	69.01
Ours_local	94.10	95.10	93.57	95.38	94.54	73.12	68.13	69.80	71.33	70.60
Ours_global	93.98	97.13	95.44	94.39	95.24	73.92	70.32	70.99	69.98	71.30
Ours_total	**95.50**	**97.56**	**96.00**	**95.98**	**96.26**	**74.10**	**71.92**	**71.34**	**69.99**	**71.84**

Table 2: Result on Zappos [36] dataset. (%)

scheme. The reason is that "maximum speed" and "displacement" of a car show little correlation with car's visual appearance, and it is somewhat difficult to predict them merely from visual information. However, involving other appearance sensitive attributes such as "door number" and "seat number" greatly aid the two attributes prediction because they are positively correlated with each other. Besides, involving local context or global style information do escalate overall performance when comparing to using single-path CNNs alone (Ours_indiv), which shows that local context and global style information extracted from CNNs intermediate layers are very important to escalate relative attribute prediction. In general, our proposed strategy which uses relative shallow CNNs by incorporating local context and global style information achieves state-of-the-art results when comparing with several other relevant methods, and also achieves competitive results comparing with current most powerful very deep neural network 152-ResNet [19] and Google inception-v3 [8]. Moreover, we find that "maximum speed" and "displacement" show subtle difference w.r.t viewpoint variation, but "door number" and "seat number" relative attributes can be best predicted from S, FS and RS viewpoint. This is easy to understand because a side relative viewpoint reveals the maximum part of a car's doors or seats.

We want to figure out which part of a car image corresponds to the attribute we are modelling. To this end, we adopt the method [20] to visualise the saliency map of each attribute. As is shown in Fig. 4, we can clearly see that while "door number" and "seat number" mainly lie close to car's window and door area, which correspond to local context discussed in this paper, the extra 2 attributes "maximum speed" and "displacement" lie across the whole car torso, which indicates the global style information. This again demonstrates the importance of involving both local context and global style information to predict relative attributes.

Zappos50K [36] is collection of dataset of 50,025 shoe catalog images from Zappos.com and relative labels of 4 attributes: open, sporty, comfort and pointy. It contains two sub-datasets: coarse and relatively simple dataset Zappos50K-1, containing approximately 1,500 annotated image pairs, fine-grained and hard dataset Zappos50K-2 with approximately 4,300 pairs. The result is shown in Table 2, from which we can see that our proposed framework outperforms other methods by a large margin in both coarse and fine-grained dataset. Involving local context specially improves "open" and "pointy" relative attributes prediction. Incorporating global style information mainly escalates more abstract relative attributes prediction, such as "comfort" and "sporty". It is naturally easy to understand this phenomena because, as we discussed above, global style information mainly corresponds to abstract and high-level feature and local context specifies more localised feature.

LFW-10 [24] has a total of 2,000 face images and 10 attributes in total. In our experiment, we have found about 500 image pairs for each attribute. **PubFig**[11] contains 800 face images of celebrates and a total of 11 attributes. **OSR** [11] consists of more than 2,500

Attribute	Bald	DkHair	Eyes	GdLook	Mascu	Mouth	Smile	Teeth	FrHead	Young	Mean
ReAtr[11]	70.4	75.7	52.6	68.4	71.3	55.0	54.6	56.0	64.5	65.8	63.4
LoLe[36]	67.9	73.6	49.6	64.7	70.1	53.4	59.7	53.5	65.6	66.2	62.4
DeepRe[34]	81.27	88.92	91.98	72.03	95.40	89.04	84.75	89.33	84.11	73.35	85.02
Napa[35]	78.8	72.4	70.7	67.6	84.5	67.8	67.4	71.7	79.3	68.4	72.9
152-ResNet[19]	82.33	89.94	92.33	75.04	96.00	89.10	84.77	92.34	85.72	74.40	86.28
inception-v3[8]	83.00	90.11	92.00	76.21	95.78	89.33	85.72	93.27	85.79	75.20	86.63
Ours_none	82.33	89.01	91.39	72.04	95.67	89.09	84.80	89.35	84.11	73.72	85.15
Ours_local	82.30	88.72	92.90	72.05	95.92	91.01	85.00	92.33	84.33	73.20	85.74
Ours_global	82.72	89.94	91.90	75.06	97.70	89.07	86.20	89.53	85.78	75.35	86.33
Ours_total	**83.09**	**90.01**	**93.14**	**75.70**	**97.93**	**89.12**	**89.50**	**85.89**	**86.11**	**75.58**	**86.60**

Table 3: Result on LFW-10 [24] dataset. (%)

Attribute	Natural	Open	Perspective	Large Size	Diag	ClsDepth	Mean
ReAtr[11]	95.03	90.77	86.73	86.23	86.50	87.53	88.80
LoLe[36]	95.70	94.10	90.43	91.10	92.43	90.47	92.37
DeepRe[34]	97.96	94.48	92.37	92.70	95.14	91.44	93.98
Napa[35]	94.98	92.32	91.98	92.77	95.10	91.67	93.14
152-ResNet[19]	98.80	96.77	92.58	94.74	96.34	93.10	95.39
inception-v3[8]	99.10	97.71	91.00	95.12	97.10	92.30	95.39
Ours_none	97.00	93.92	92.48	92.78	96.30	91.98	94.08
Ours_local	98.02	94.52	93.27	92.50	96.00	92.00	94.66
Ours_global	98.45	96.00	93.64	94.01	96.88	91.86	95.14
Ours_total	**98.91**	**96.32**	**94.20**	**94.93**	**97.01**	**92.29**	**95.62**

Table 4: Result on OSR [11] dataset. (%)

outdoor scene images, with 6 relative attributes. We follow the pre-defined train/test split of both PubFig and OSR datasets. LFW-10 result is given in Table 3. Similarly, we can observe that local features extracted from CNNs intermediate layers improves most in more localised and specified relative attribute prediction, for example, the "teeth", "eyes" and "bald". Other more abstract attributes such as "young", "good look" and "masculine" benefit most from involving global style information. The results of OSR dataset [11] and PubFig dataset [11] are given in Table 4 and Table 5, respectively. They represent similar experimental results with Zappos5K and LFW-10 dataset. That is, our proposed framework achieves state-of-the-art performance in almost all relative attribute prediction tasks. Note that in PubFig dataset, our methods performs slightly inferior to inception-v3. We think the reason is that "young" attribute requires much deeper neural network to comprehensively represent it. Global style information extracted from shallow CNNs cannot completely capture the global "young" information. Again, we can clearly see the saliency map of each attribute on all datasets in Fig. 4, which help us to intuitively understand local context and global style information.

In sum, our proposed relative attribute prediction framework shows impressive performance on all the 5 publicly available datasets, which contain various scenarios, including indoor, outdoor, face, car, natural and human-designed images. The incorporation of local context and global style information defined in this paper enables our framework to successfully handle both coarse and fine-grained, even large pose variations relative attribute prediction tasks. Besides, it is noteworthy to note that our proposed framework achieves comparable results comparing with current popular very deep neural networks. Since very deep neural networks are time-consuming and require powerful machines to train, our framework is cost-effective and time-effective.

5 Conclusion

We proposed a novel framework for relative attribute prediction. We build on CNNs and exploit its intermediate layers to force CNNs to learn local context and global style information, both of which are demonstrated to improve a lot on both coarse and fine-grained, pose-variational relative attribute prediction. Future work includes research on much more

Attribute	Male	White	Young	Smiling	Chubby	Forehead	Eyebrow	Eye	Nose	Lip	Face	Mean
ReAtr[11]	81.80	76.97	83.20	79.90	76.27	87.60	79.87	81.67	77.40	79.17	82.33	80.53
LoLe[36]	91.77	87.43	91.87	87.00	87.37	94.00	89.83	91.40	89.07	90.43	86.70	89.72
DeepRe[34]	90.10	89.49	89.83	88.62	88.72	92.33	88.13	86.94	86.30	89.79	92.71	89.36
Napa[35]	82.10	78.21	83.90	78.00	76.33	86.98	79.72	80.69	75.98	79.99	83.00	80.45
152-ResNet[19]	92.33	89.99	91.30	90.03	89.00	91.11	89.47	88.94	87.34	91.09	93.11	90.59
inception-v3[8]	91.87	88.22	**91.78**	90.92	88.97	92.00	89.92	88.68	89.00	90.09	93.08	90.41
Ours_none	90.11	89.33	91.00	88.99	88.98	93.19	89.10	85.99	86.78	90.79	93.00	89.75
Ours_local	89.10	90.20	91.00	89.34	89.28	92.18	90.34	90.72	87.39	91.00	92.88	89.94
Ours_global	91.00	90.38	91.09	90.00	89.55	92.78	89.00	86.69	87.32	91.72	93.12	90.24
Ours_total	**92.39**	**90.75**	91.10	**90.24**	**93.00**	**93.00**	**91.78**	**87.62**	**88.38**	**92.84**	**93.22**	**91.28**

Table 5: Result on PubFig [11] dataset. (%)

deeper network architectures' performance on relative attribute prediction, and experiments on other intermediate feature encoding methods.

6 Acknowledgement

This research is supported by the National Natural Science Foundation of China (NSFC) under grant No. 41401525, the Natural Science Foundation of Guangdong Province under grant No. 2014A030313209.

References

[1] D. Hoiem A. Farhadi, I. Endres and D. Forsyth. Describing objects by their attributes. In *Proc. CVPR*, 2009.

[2] M. Sadeghi A. Farhadi, M. Hejrati, P. Young, and C. Raschtchian. Every picture tells a story: Generating sentences from images. In *Proc. ECCV*, 2010.

[3] D. Parikh A. Kovashka and L. Grauman. Whittlesearch: Image search with relative attribute feedback. In *Proc. CVPR*, 2012.

[4] A. Torralba A. Oliva. Modeling the shape of the scene: A holistic representation of the spatial evelope. *IJCV*, 42(3):145–175, 2001.

[5] R.B. Girshick B. Hariharan, P.A. Arbelaez and J. Malik. Hyper-columns for object segmentation and fine-grained localization. In *Proc. CVPR*, pages 447–456, 2015.

[6] E. Renshaw C. Burges, T. Shaked, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. Learning to rank using gradient descent. In *Proc. ICML*, pages 89–96, 2005.

[7] H. Nickisch C. Lampert and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE T-PAMI*, 36(3):453–465, 2014.

[8] S. Ioffe C. Szegedy, V. Vanhoucke, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. *arXiv preprint arXiv:1512.00567*, 2015.

[9] Y . Jia C. Szegedy, W. Liu, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proc. CVPR*, 2015.

[10] B. Triggs D. Nalal. Histograms of oriented gradients for human detetection. In *Proc. CVPR*, pages 886–893, 2005.

[11] Kristen Grauman Devi Parikh. Relative attributes. In *Proc. ICCV*, 2011.

[12] W. Choi F. Yang and Y. Lin. Exploit all layers: fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers. In *Proc. CVPR*, 2016.

[13] V. Ferrari and A. Zisserman. Learning visual attributes. In *Proc. NIPS*, pages 433–440, 2007.

[14] E. Shelhamer J. Long and T. Darrel. Fully convolutional networks for semantic segmentation. In *Proc. CVPR*, pages 3431–3440, 2015.

[15] C. C. Loy X. Wang J. Shao, K. Kang. Deeply learned attributes for crowded scene understanding. In *Proc. CVPR*, 2015.

[16] T. Leung J. Wang, Y. Song, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu. Proc. cvpr. Learning fine-grained image similarity with deep ranking, 2014.

[17] Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez. Aggregating local descriptors into a compact image representation. In *Proc. CVPR*, pages 3304–3311, 2010.

[18] S. Ren K. He, X. Zhang. Identity mappings in deep residual networks. *arXiv preprint arXiv:1603.05027*, 2016.

[19] S. Ren K. He, X. Zhang and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. In Arxiv.

[20] A. Vedaldi K. Simonyan and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv: 1312.6034*, 2013.

[21] L. Shao L. Liu, Y. Zhou. Dap3d-net: Where, what and how actions occur in videos? *arXiv preprint arXiv: 1602.03346*, 2016.

[22] Ping Luo Linjie Yang, Chen Change Loy, and Xiaoou Tang. A large-scale car dataset for fine-grained categorization and verification. In *Proc. CVPR*, 2015.

[23] H. Su O. Russakovsky, J. Deng, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein. Imagenet large scale visual recognition challenge. *IJCV*, pages 211–252, 2015.

[24] Y. Verma R. N. Sandeep and C. V. Jawahar. Relative parts: Distinctive parts for learning relative attributes. In *Proc. CVPR*, 2014.

[25] O. Beijbom S. Branson and S. Belongie. Efficient large scale structured learning. In *Proc. CVPR*, 2013.

[26] Y. Qiao S. Guo, W. Huang. Locally-supervised deep hybrid model for scene recognition. In *arXiv preprint arXiv: 1510.03283*, 2015.

[27] S. Shan S. Li and X Chen. Relative forest for attribute prediction. In *Proc. ACCV*, 2012.

[28] R. Qian S. Liu, X. Ou, W. Wang, and X. Cao. Makeup like a superstar: Deep localized makeup transfer network. *arXiv preprint arXiv:1604.07102*, 2016.

[29] V. K. Garg S. Shankar and R. Cipolla. Discovering visual attributes by carving deep neural nets. In *Proc. CVPR*, 2015.

[30] Z. Tu S. Xie. Holistically-nested edge detection. In *Proc. ICCV*, 2015.

[31] T. Tieleman and G. Hinton. Rmsprob: Divide the gradient by running average of its recent magnitute. Coursera: neural networks for machine learning, 2012.

[32] Y. Bengio X. Glorot. Understanding the difficulty of training deep feedforward neural networks. In *Proc. AISTATS*, pages 249–256, 2010.

[33] E. Shelhamer Y, Jia, D. Jeff, K. Sergey, L. Jonathan, G. Ross, G. Sergio, and D. Trevor. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[34] E. Adeli-Mosabbed Y. Souri, E. Noury. Deep relative attributes. *arXiv preprint arXiv:1512.04103*, 2015.

[35] C. Jawahar Y. Verma. Exploring locally rigid discriminative patches for learning relative attributes. In *Proc. BMVC*, 2015.

[36] A. Yu and K. Grauman. Fine-grained visual comparisions with local learning. In *Proc. CVPR*, 2014.

[37] A. G. Hauptmann Z. Xu, Y. Yang. A discriminative cnn video representation for event detection. In *Proc. CVPR*, pages 1798–1807, 2015.