applied sciences

MDPI

Article

Multi-Task Learning Using Task Dependencies for Face Attributes Prediction

Di Fan [1], Hyunwoo Kim [1,2,*], Junmo Kim [3], Yunhui Liu [4] and Qiang Huang [1,2]

[1] School of Mechatronical Engineering, Beijing Institute of Technology, Beijing 100081, China; fandi0126@bit.edu.cn (D.F.); qhuang@bit.edu.cn (Q.H.)

[2] Beijing Advanced Innovation Center for Intelligent Robots and Systems, Beijing Institute of Technology, Beijing 100081, China

[3] School of Electrical Engineering, Korea Advanced Institute of Science and Technology, 291 Daehak-ro, Yuseong-gu, Daejeon 34141, Korea; junmo.kim@kaist.ac.kr

[4] Department of Mechanical and Automation Engineering, The Chinese University of Hong Kong, Shatin, NT, Hong Kong SAR, China; yhliu@mae.cuhk.edu.hk

* Correspondence: eugene.hwkim@gmail.com

Received: 21 May 2019; Accepted: 19 June 2019; Published: 21 June 2019

check for updates

Abstract: Face attributes prediction has an increasing amount of applications in human–computer interaction, face verification and video surveillance. Various studies show that dependencies exist in face attributes. Multi-task learning architecture can build a synergy among the correlated tasks by parameter sharing in the shared layers. However, the dependencies between the tasks have been ignored in the task-specific layers of most multi-task learning architectures. Thus, how to further boost the performance of individual tasks by using task dependencies among face attributes is quite challenging. In this paper, we propose a multi-task learning using task dependencies architecture for face attributes prediction and evaluate the performance with the tasks of smile and gender prediction. The designed attention modules in task-specific layers of our proposed architecture are used for learning task-dependent disentangled representations. The experimental results demonstrate the effectiveness of our proposed network by comparing with the traditional multi-task learning architecture and the state-of-the-art methods on Faces of the world (FotW) and Labeled faces in the wild-a (LFWA) datasets.

Keywords: multi-task learning; task dependencies; attention; face attributes prediction; deep convolutional neural network

1. Introduction

Face attributes are useful to achieve detailed description of human faces (e.g., smile, gender, age, etc.). Face attributes prediction has applications in human–computer interaction, face verification [1,2] and video surveillance [3,4]. Face variations in pose, illumination, scale and occlusion increase the difficulty of face attributes prediction. The performance of face attributes prediction has been improved by using deep convolutional neural networks (DCNNs) [5–10]. Face attributes prediction is trained separately in these networks, but the inherent correlation between the face attributes has been ignored.

Various studies show that dependencies exist in face attributes [11–15]. Multi-task learning networks can improve the performance of individual tasks by jointly learning correlated tasks. In traditional multi-task learning architectures, the shared layers learn general representations for all the tasks by parameter sharing while the following task-specific representations are learned in the task-specific layers. However, the dependencies between the tasks have been ignored in the task-specific layers. Accordingly, further improving the performance of individual tasks by using task dependencies among face attributes in the task-specific layers of the multi-task learning architecture is a challenge problem.

We propose a multi-task learning using task dependencies architecture for face attributes prediction and evaluate the performance with the tasks of smile and gender prediction. Our proposed architecture splits into two task-specific branches after the shared layers. In the task-specific branches, we establish the task dependencies in the task-specific layers by incorporating attention mechanism. The fully connected layers in the task-specific layers are transformed by using the designed attention modules for learning task-dependent disentangled representations, where the task-dependent disentangled representations denote the representations [16,17] of one task that are disentangled [18] by depending on another task. The transformed fully connected layers that contain task-dependent disentangled representations are fed into softmax layers to predict the final face attributes. In experiments, we demonstrate the effectiveness of our proposed network by comparing with the traditional multi-task learning architecture and the state-of-the-art methods on FotW and LFWA datasets.

The rest of this paper is organized as follows: Section 2 briefly reviews related works. Section 3 describes the proposed multi-task learning using task dependencies architecture in detail. Section 4 describes the experimental configuration; the results on FotW and LFWA datasets are also presented and discussed in Section 4. Section 5 concludes the paper.

2. Related Work

Multi-task learning. Caruana [19] first analyzed multi-task learning in detail. Since then, multi-task learning has been adopted for solving different computer vision problems. Gkioxari et al. used a convolutional neural network (CNN) for pose prediction and action classification of people in unconstrained images [20]. Eigen et al. proposed a multi-scale convolutional architecture for predicting depth, surface normals and semantic labels [21]. Misra et al. presented cross-stitch units to learn shared representations for multi-task learning in ConvNets [22]. Kokkinos et al. presented a CNN that jointly handles low-, mid-, and high-level vision tasks in a unified architecture [23]. Mallya et al. studied a method for performing multiple tasks in a single deep neural network by iteratively pruning and packing the network parameters [24]. Kim et al. proposed a novel architecture containing multiple networks of different configurations termed deep virtual networks with respect to different tasks and memory budgets [25]. Recently, multi-task learning with DCNNs have also been studied and applied to face attributes prediction. Levi et al. used a deep convolutional neural network(DCNN) for age and gender classification [26]. Liu et al. proposed a novel deep learning framework for attribute prediction in the wild [27]. Ranjan et al. presented a DCNN for face analysis utilizing transfer learning from a face recognition model [28]. Hyun et al. proposed a method to multi-attribute recognition of facial images based on a deep learning network that automatically learns the exclusive and joint relationship among attribute recognition tasks [29]. In multi-task learning, when the prediction of one task which will be used as condition is accurate, other tasks can be formulated by using conditional probability. For example, in [30], the experimental results on the MORPH-II dataset show that the multitask method achieves 98% gender recognition accuracy, thus the age probability $P(A(X) = a)$ can be calculated using the gender-conditioned probability $P(A(X) = a \mid G(X) = g)$ and the marginal gender probability $P(G(X) = g)$ in their proposed conditional multitask learning method. However, the error predicted gender $G(X) = g$ will lead to incorrect calculation of $P(A(X) = a \mid G(X) = g)$ and $P(G(X) = g)$; therefore, their method cannot be used when the multitask method cannot predict gender accurately on other datasets.

Attention mechanism. Human perception is similar to the attention mechanism that selects specific parts of the input information, rather than using all input information. In neural networks, attention mechanism can be used as feature selectors that can determine the importance of each feature for the particular task. The attention mechanism has been studied and applied to recurrent neural networks (RNNs) and long short term memory (LSTM) for sequential tasks [31–33]. The attention mechanism with DCNNs have been applied to vision-related tasks. Tang et al. proposed a deep-learning based generative framework with visual attention [34]. Xiao et al. applied visual attention to fine-grained classification task using DCNN [35]. Xu et al. presented an attention based model that automatically learns to describe the content of images [36]. Zhao et al. proposed a diversified visual attention network for fine-grained object classification [37]. Inspired by the attention mechanism, we propose a multi-task learning using task dependencies architecture for face attributes prediction in this paper.

The main contributions of this paper are summarized as follows:

1. A multi-task learning using task dependencies architecture for face attributes prediction in end-to-end manner. The designed attention modules in our proposed architecture are used for learning task-dependent disentangled representations. We evaluate the performance with the tasks of smile and gender prediction.
2. We present experimental results which demonstrate that our proposed architecture outperforms the traditional multi-task learning architecture and show the effectiveness in comparison with the state-of-the-art methods on FotW and LFWA datasets.

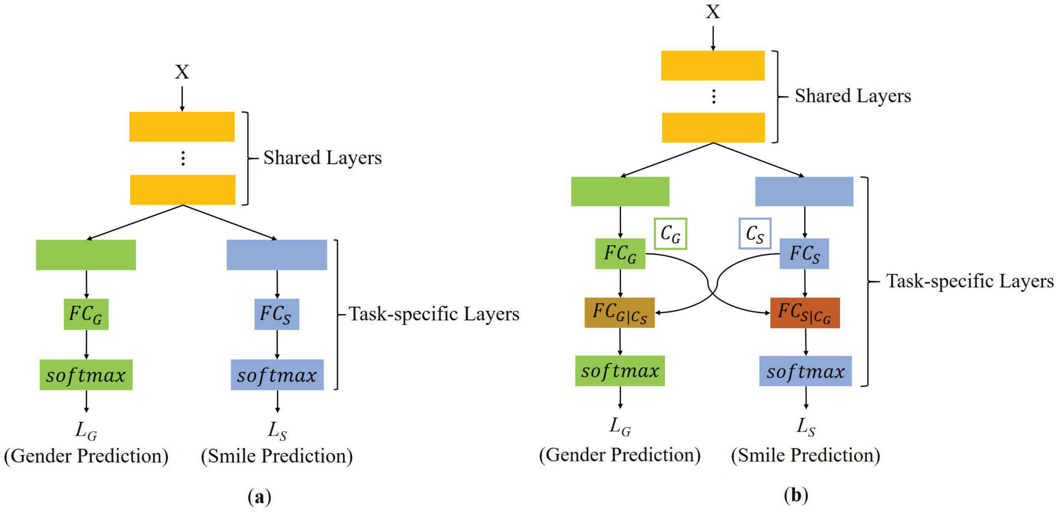
3. Proposed Method

3.1. Modeling

Formally, the smile/non-smile prediction of the input face X is defined as $S(X)$. The expected smile/non-smile prediction of the input X is defined as follows:

$$E[S(X)] = \sum_{s \in S} s \cdot P(S(X) = s), \tag{1}$$

where $P(S(X) = s)$ is the probability that the smile/non-smile prediction of the input X is s, where $s \in S$. We define S = {'non-smile', 'smile'}.



(**a**) (**b**)

Figure 1. Comparison of traditional multi-task learning and our proposed multi-task learning architecture. (**a**) traditional multi-task learning; (**b**) our proposed multi-task learning.

We assume that the predicted smile/non-smile is dependent on the gender of the input X. Compared to the traditional multi-task learning architecture as shown in Figure 1a, the FC_S layer has been transformed into $FC_{S|C_G}$ layer in the multi-task learning architecture we proposed (shown in Figure 1b). FC_S denotes the fully connected layer that contains $K(K \in \mathbb{N})$ smile/non-smile representation units. $FC_{S|C_G}$ denotes the transformed gender dependent fully connected layer that contains K gender dependent smile/non-smile representation units, where C_G is the gender context.

We feed the transformed $FC_{S|C_G}$ layer into the softmax layer to predict the final smile/non-smile. The probability $P(S(X) = s)$ in Equation (1) can be modeled as follows:

$$P(S(X) = s) = softmax(FC_{S|C_G}(X)). \tag{2}$$

The gender context C_G contains K gender context units C_{Gi}, where C_{Gi} is the i-th ($i = 1, 2, \ldots K$) gender context unit that is automatically chosen from the gender representation units in the FC_G layer. FC_G denotes the fully connected layer that contains K gender representation units.

The dependency score function $score(x_{Sj}, C_{Gi})$ that takes a conjunction of the j-th($j = 1, 2, \ldots K$) input smile/non-smile representation unit x_{Sj} from the FC_S layer and the i-th gender context unit C_{Gi} from the FC_G layer to score the dependency between x_{Sj} and C_{Gi}. The dependency score function can be formulated as follows:

$$score(x_{Sj}, C_{Gi}) = tanh(W_S x_{Sj} + W_G C_{Gi}). \tag{3}$$

The probability $P(d = j \mid \boldsymbol{x_S}, C_{Gi})$ reveals the relative importance of x_{Sj} based on C_{Gi}, where d indicates which input smile/non-smile representation unit in $\boldsymbol{x_S}$ is important based on C_{Gi}, where $\boldsymbol{x_S}$ contains K input smile/non-smile representation units. The probability $P(d = j \mid \boldsymbol{x_S}, C_{Gi})$ can be calculated using the dependency score function as follows:

$$P(d = j \mid \boldsymbol{x_S}, C_{Gi}) = \frac{exp(score(x_{Sj}, C_{Gi}))}{\sum\limits_{j=1}^{K} exp(score(x_{Sj}, C_{Gi}))}. \tag{4}$$

The importance probability distribution $P(d \mid \boldsymbol{x_S}, C_{Gi})$ is defined as follows:

$$P(d \mid \boldsymbol{x_S}, C_{Gi}) = [P(d = j \mid \boldsymbol{x_S}, C_{Gi})]_{j=1}^{K}. \tag{5}$$

The gender dependent smile/non-smile representation units in the transformed $FC_{S|C_G}$ layer can be defined as follows:

$$\hat{S}_i = E_{\boldsymbol{x_S} \sim P(d|\boldsymbol{x_S}, C_{Gi})}(\boldsymbol{x_S}) = \sum_{j=1}^{K} P(d = j \mid \boldsymbol{x_S}, C_{Gi}) x_{Sj}, \tag{6}$$

where $\hat{S}_i$ is the i-th ($i = 1, 2, \ldots K$) gender dependent smile/non-smile representation unit that is the weighted average of the input smile/non-smile representation units. $\hat{S}_i$ can be formulated as the expectation of $\boldsymbol{x_S}$ according to the importance probability distribution $P(d \mid \boldsymbol{x_S}, C_{Gi})$. The transformed $FC_{S|C_G}$ layer is generated by concatenating K gender dependent smile/non-smile representation units.

The gender prediction of the input face X is defined as $G(X)$. The expected gender prediction is defined as follows:

$$E[G(X)] = \sum_{g \in G} g \cdot P(G(X) = g), \tag{7}$$

where $P(G(X) = g)$ is the probability that the gender prediction of the input X is g, where $g \in G$. We define G ={*'male'*, *'female'*}.

We also assume that the predicted gender is dependent on the smile/non-smile of the input X. The probability $P(G(X) = g)$ in Equation (7) can be modeled as follows:

$$P(G(X) = g) = softmax(FC_{G|C_S}(X)), \tag{8}$$

where $FC_{G|C_S}$ denotes the transformed smile/non-smile dependent fully connected layer that contains K smile/non-smile dependent gender representation units, where C_S is the smile/non-smile context chosen from the smile/non-smile representation units in the FC_S layer.

The calculation of the smile/non-smile dependent gender representation units in the transformed $FC_{G|C_S}$ layer is similar to calculating the gender dependent smile/non-smile representation units in the transformed $FC_{S|C_G}$ layer. The smile/non-smile dependent gender representation units can be calculated using Equations (9)–(12) as follows:

$$score(x_{Gj}, C_{Si}) = tanh(W_G x_{Gj} + W_S C_{Si}), \tag{9}$$

$$P(d = j \mid \boldsymbol{x_G}, C_{Si}) = \frac{exp(score(x_{Gj}, C_{Si}))}{\sum\limits_{j=1}^{K} exp(score(x_{Gj}, C_{Si}))}, \tag{10}$$

$$P(d \mid \boldsymbol{x_G}, C_{Si}) = [P(d = j \mid \boldsymbol{x_G}, C_{Si})]_{j=1}^{K}, \tag{11}$$

$$\hat{G}_i = E_{\boldsymbol{x_G} \sim P(d|\boldsymbol{x_G}, C_{Si})}(\boldsymbol{x_G}) = \sum_{j=1}^{K} P(d = j \mid \boldsymbol{x_G}, C_{Si}) x_{Gj}. \tag{12}$$

3.2. Network Architecture

The multi-task learning architecture we proposed is shown in Figure 2. The ResNet50 [38] Network is adapted as the baseline architecture. We share the parameters from its first 46 layers for all the tasks. We evaluate our proposed architecture with the tasks of smile and gender prediction. Thus, the network splits into two task-specific branches corresponding to smile and gender prediction. We attach a fully connected layer FC_S that contains 64 smile/non-smile representation units and a fully connected layer FC_G that contains 64 gender representation units respectively to 'res5c1' and 'res5c2', where 'res5c1' and 'res5c2' are residual blocks in ResNet50. The smile/non-smile representation units in FC_S layer and the i-th ($i = 1, 2, \ldots$ 64) gender context unit C_{Gi} are fed into the i-th ($i = 1, 2, \ldots$ 64) gender context attention module Att_C_{Gi} (shown in Figure 3), where Att_C_{Gi} is designed to learn the i-th ($i = 1, 2, \ldots$ 64) gender dependent smile/non-smile representation unit $\hat{S}_i$ by using Equations (3)–(6). The transformed $FC_{S|C_G}$ layer is generated by concatenating 64 gender dependent smile/non-smile representation units. We feed the transformed $FC_{S|C_G}$ layer into the softmax layer to predict the final smile/non-smile. The procedure of predicting the final gender is similar to that of predicting the final smile/non-smile in our proposed architecture. The i-th ($i = 1, 2, \ldots$ 64) smile/non-smile context attention module Att_C_{Si} (shown in Figure 4) is designed to learn the i-th ($i = 1, 2 \ldots 64$) smile/non-smile dependent gender representation unit $\hat{G}_i$ by using Equations (9)–(12).

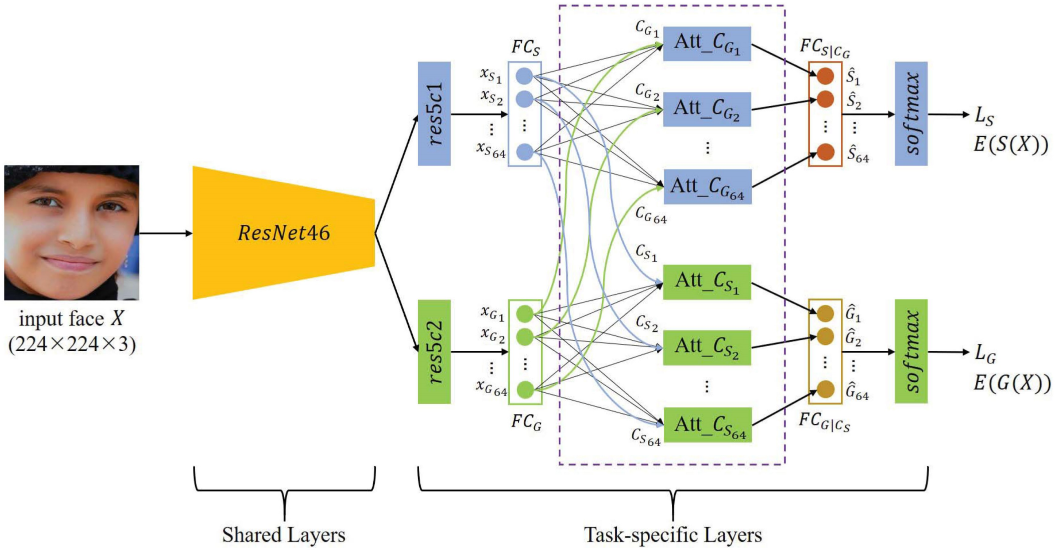


Figure 2. The architecture of the proposed multi-task learning convolutional neural network.

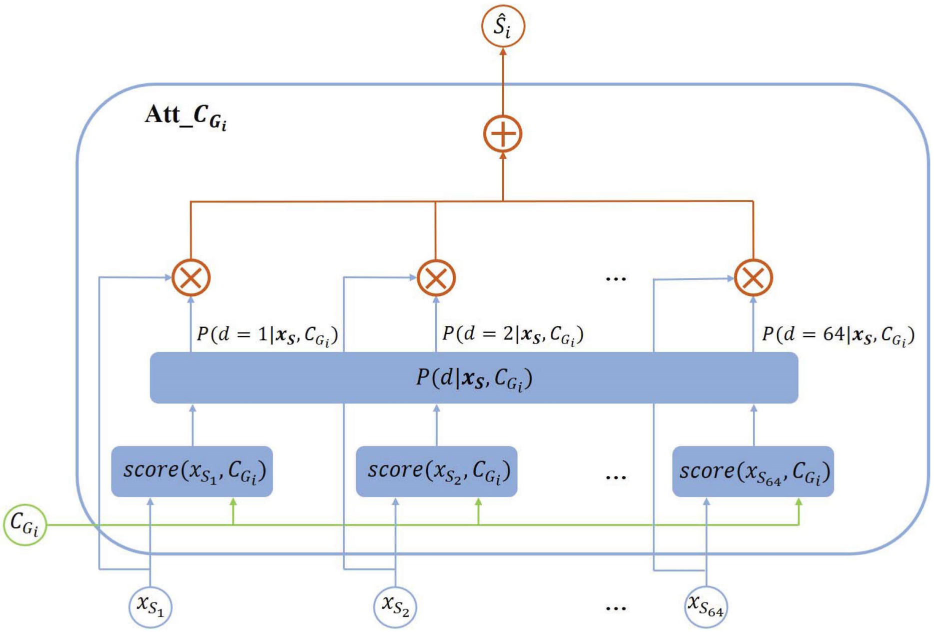

$\hat{S}_i$
Att_C_{G_i}
$P(d=1|\mathbf{x_S}, C_{G_i})$
$P(d=2|\mathbf{x_S}, C_{G_i})$
$P(d=64|\mathbf{x_S}, C_{G_i})$
$P(d|\mathbf{x_S}, C_{G_i})$
$score(x_{S_1}, C_{G_i})$
$score(x_{S_2}, C_{G_i})$
$score(x_{S_{64}}, C_{G_i})$
C_{G_i}
x_{S_1}
x_{S_2}
$x_{S_{64}}$

Figure 3. The designed gender context attention module.

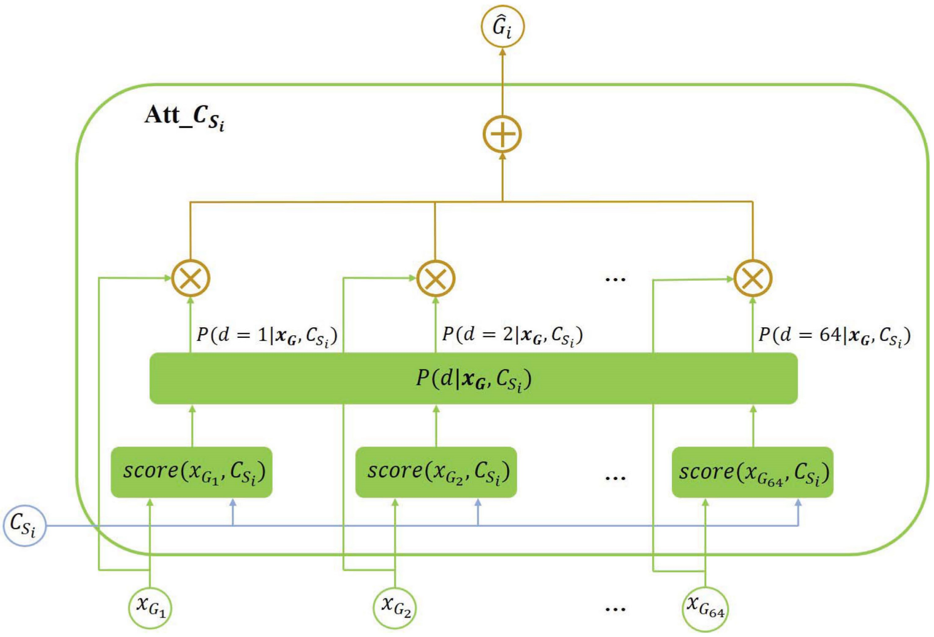

$\hat{G}_i$
Att_C_{S_i}
$P(d=1|\mathbf{x_G}, C_{S_i})$
$P(d=2|\mathbf{x_G}, C_{S_i})$
$P(d=64|\mathbf{x_G}, C_{S_i})$
$P(d|\mathbf{x_G}, C_{S_i})$
$score(x_{G_1}, C_{S_i})$
$score(x_{G_2}, C_{S_i})$
$score(x_{G_{64}}, C_{S_i})$
C_{S_i}
x_{G_1}
x_{G_2}
$x_{G_{64}}$

Figure 4. The designed smile/non-smile context attention module.

3.3. The Model Objective

We use the cross-entropy loss for training the smile prediction task. The loss function L_S is formulated as follows:

$$L_S = -s \cdot log(p_s) - (1-s) \cdot log(1-p_s), \tag{13}$$

where $s=1$ for a smiling face and $s=0$, otherwise. p_s is the final predicted probability that the input is a smiling face.

We also use the cross-entropy loss for training the gender prediction task. The loss function L_G is formulated as follows:

$$L_G = -g \cdot log(p_g) - (1-g) \cdot log(1-p_g), \tag{14}$$

where $g = 0$ if the gender is male and $g = 1$ if the gender is female. p_g is the final predicted probability that the input face is a female.

The total loss L is the weighted sum of the individual losses. L is defined as follows:

$$L = \lambda_s \cdot L_S + \lambda_g \cdot L_G, \tag{15}$$

where λ_s and λ_g are weight parameters corresponding to smile and gender prediction task, respectively.

4. Experiments

The proposed multi-task learning using task dependencies architecture is evaluated with the tasks of smile and gender prediction. The architecture in which we feed FC_S and FC_G layers directly into softmax layers to predict the final smile/non-smile and gender respectively as shown in Figure 1a is called TMTL (Traditional Multi Task Learning). We select TMTL architecture as the comparison baseline.

4.1. Datasets

We evaluate the smile and gender prediction performance on Faces of the World (FotW) [39] and Labeled Faces in the Wild-a (LFWA) [40] datasets. Both FotW and LFWA datasets cover large variations in pose, illumination and scale of faces. The FotW dataset contains 9130 images, each of which is labeled with non-smile/smile and male/female. The FotW dataset has been split into 6078 images for training and 3052 images for validation. The LFWA dataset contains 13,143 images, each of which is labeled with non-smile/smile, male/female and thirty-eight other face attributes. The LFWA dataset has been split into 6263 images for training and 6880 images for validation.

4.2. Experimental Configuration

For the FotW dataset, we crop the faces from the original images using the provided coordinates of the bounding box and resize the cropped face images to 224 × 224 × 3. For the LFWA dataset, we directly resize the face images to 224 × 224 × 3.

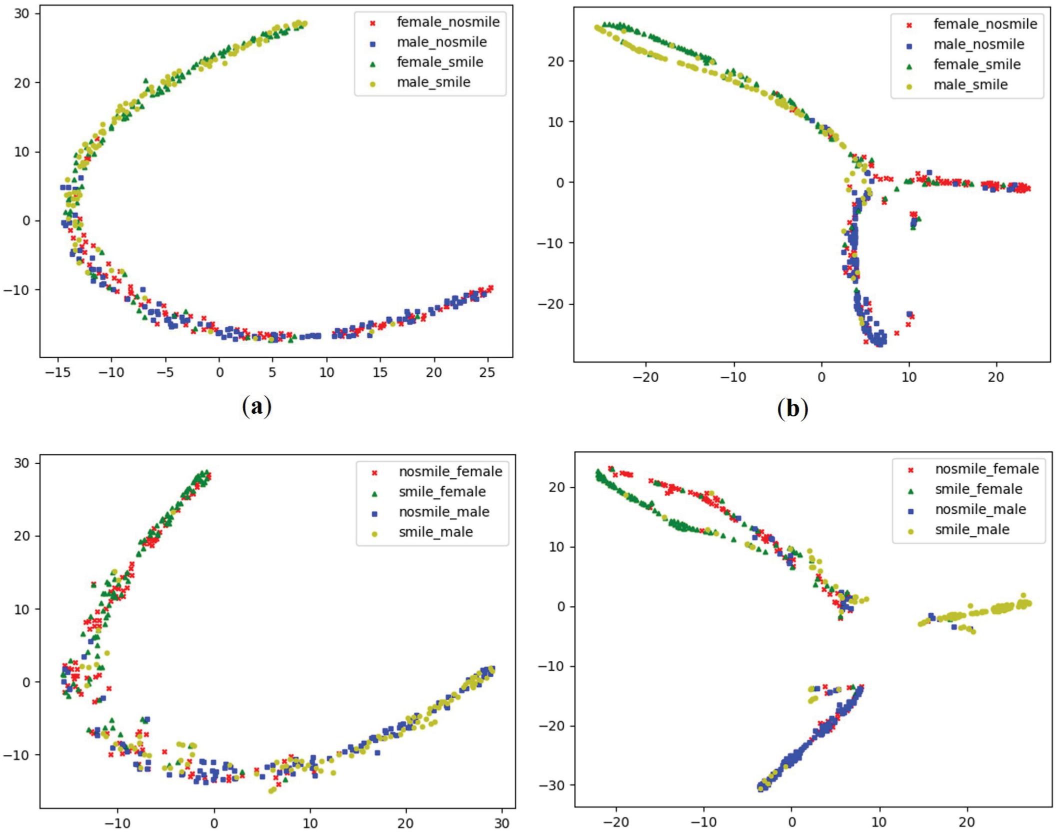
All the architectures are trained using the keras [41] framework. Data augmentation such as horizontal flip, horizontal shift and vertical shift are adopted to prevent overfitting. We train all the architectures using Adam with a mini-batch size of 64. The initial learning rate is set to 0.001. The learning rate will decrease to 0.0001 after training 25 epochs. The weight parameters are decided based on the importance of the task in the overall loss. We assume that the smile prediction task and the gender prediction task have the same importance in our proposed architecture due to both of the tasks being binary classification problems. Therefore, we set the weight parameters $\lambda_s = 1$, $\lambda_g = 1$. For FotW dataset and LFWA datasets, we adopt he_normal as the weight initialization method and train TMTL architecture 40 epochs (overfitting after 40 epochs) and 30 epochs (overfitting after 30 epochs), respectively. For all the datasets, we initialize our proposed architecture with trained weights from TMTL architecture and train 30 epochs, respectively.

4.3. The Effectiveness of Multi-Task Learning Using Task Dependencies

We evaluate the contribution of multi-task learning using task dependencies. Disentangling the underlying structure of representations into disjoint parts can benefit for solving a diverse set of tasks in a data-efficient manner. The disentangled representations are vector representations with respect to a particular decomposition of a group into subgroups using the group and representation theory [42]. Table 1 shows that our proposed architecture in comparison with TMTL architecture on FotW and LFWA datasets, respectively. Our proposed architecture produces performance gains over TMTL architecture because our proposed architecture disentangles smile/non-smile and gender representations into gender dependent smile/non-smile and smile/non-smile dependent gender representations, respectively, by establishing the task dependencies between smile and gender prediction tasks in the task-specific layers. We combine (smile/non-smile × gender) into four groups. For each of the groups, we randomly sample 100 images from the FotW validation dataset. The t-distributed stochastic neighbor embedding (t-SNE) [43] on the sampled FotW validation dataset show the distributions of the representations in FC_S and FC_G layers, respectively, in Figure 5a,c, and show the distributions of the gender dependent smile/nonsmile and smile/non-smile dependent gender representations in $FC_{S|C_G}$ and $FC_{G|C_S}$ layers, respectively, in Figure 5b,d. Clusters in Figure 5b,d are disentangled by gender and smile/non-smile more explicitly compared to those in Figure 5a,c. The procedure of achieving the sampled LFWA validation dataset is the same as that of achieving the sampled FotW validation dataset. The t-SNE on the sampled LFWA validation dataset shows the distributions of the representations in FC_S and FC_G layers, respectively, in Figure 6a,c, and show the distributions of the gender dependent smile/nonsmile and smile/non-smile dependent gender representations in $FC_{S|C_G}$ and $FC_{G|C_S}$ layers, respectively, in Figure 6b,d. Clusters in Figure 6b,d are also disentangled by gender and smile/non-smile more explicitly compared to those in Figure 6a,c.

Table 1. Comparison results for smile and gender prediction on the FotW dataset and the LFWA dataset.

Dataset	Architecture	Smile	Gender
FotW	TMTL	86.83%	82.54%
	Ours	**88.53%**	**84.83%**
LFWA	TMTL	90.74%	91.80%
	Ours	**91.13%**	**92.49%**



(**a**)

(**b**)

(**c**)

(**d**)

Figure 5. t-SNE visulization on the sampled FotW validation dataset. (**a**) t-SNE visualization of FC_S; (**b**) t-SNE visualization of $FC_{S|C_G}$; (**c**) t-SNE visualization of FC_G; (**d**) t-SNE visualization of $FC_{G|C_S}$.

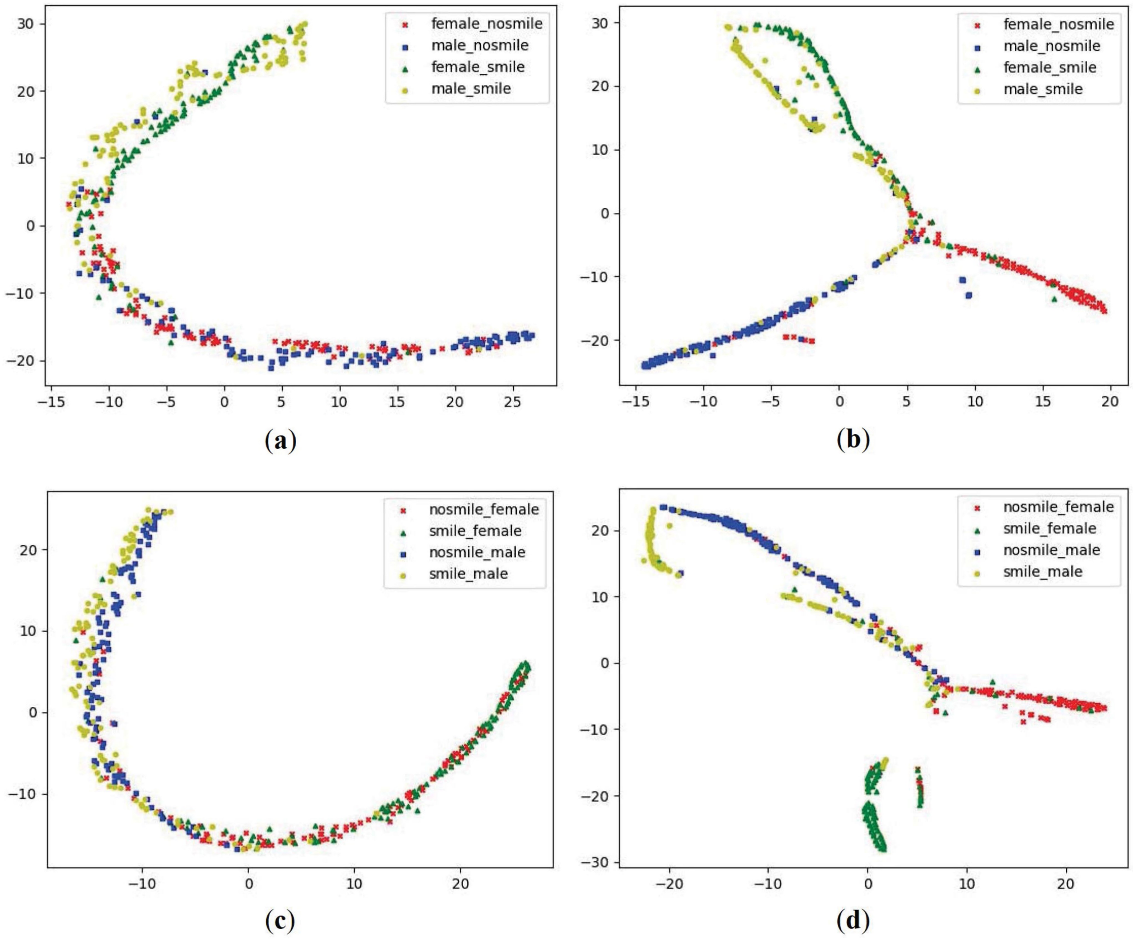


Figure 6. t-SNE visulization on the sampled LFWA validation dataset. (**a**) t-SNE visulization of FC_S; (**b**) t-SNE visualization of $FC_{S|C_G}$; (**c**) t-SNE visualization of FC_G; (**d**) t-SNE visualization of $FC_{G|C_S}$.

4.4. Comparison with Previous Approaches

We initialize our proposed architecture using the weights from ResNet50 pre-trained on ImageNet [44]. Tables 2 and 3 compare our results with those of previous methods on FotW and LFWA datasets, respectively. Our average accuracy is lower than SIAT_MMLAB on the FotW dataset and LNets+ANet on the LFWA dataset. The SIAT_MMLAB architecture is composed of GNet for gender classification and two SNets for smile classification. GNet and two SNets are trained with different face cropping schemes for better performance. The SIAT_MMLAB architecture adopts the VGG-Faces [45] model, which is pre-trained on a large-scale face identification dataset for face identification and face verification. They use a general-to-specific fine-tuning scheme that fine-tunes the model three times on CelebA [27] (with forty attribute annotations), CelebA (with smile and gender annotations) and FotW (with smile and gender annotations) datasets, respectively. The LNets+ANet architecture integrates two CNNs LNet and ANet, where LNet locates the entire face region and ANet extracts features for attribute recognition. LNet is pre-trained on ImageNet and fine-tuned by image-level attribute tags. ANet is pre-trained on the CelebA dataset and fine-tuned by attribute tags. Our proposed architecture can perform smile and gender prediction tasks in the end-to-end manner using a single deep neural network. The input face images are processed as mentioned in experimental configurations with no extra face cropping and localization steps. We only use the weights from ResNet50 pre-trained on ImageNet for weight initialization. The results also show the effectiveness of our proposed architecture in comparison with previous state-of-the-art methods.

Table 2. Performance comparison for smile and gender prediction on FotW datasets.

Architecture	Smile	Gender	Average
SMILELAB NEU [39]	81.48%	89.99%	85.74%
DMTL [46]	87.30%	84.90%	86.10%
IVA_NLPR [47]	82.52%	91.52%	87.02%
SIAT_MMLAB [48]	89.34%	91.66%	90.50%
Ours	**89.38**%	87.48%	88.43%

Table 3. Performance comparison for smile and gender prediction on LFWA datasets.

Architecture	Smile	Gender	Average
LNets+ANet(w/o) [27]	88%	91%	89.50%
PANDA-1 [49]	89%	92%	90.50%
MCFA [50]	88%	93%	90.50%
MNet [51]	89.49%	92.20%	90.85%
LNets+ANet [27]	91%	94%	92.50%
Ours	**91.38**%	92.50%	91.94%

5. Conclusions

In this paper, we have proposed a novel multi-task learning using task dependencies architecture for face attributes prediction and evaluated the performance with the tasks of smile and gender prediction. We transformed the fully connected layers by using the designed attention modules for learning task-dependent disentangled representations. The transformed fully connected layers were fed into softmax layers to predict the final face attributes. The experimental results demonstrate the effectiveness of our proposed network by comparing with the traditional multi-task learning architecture and the state-of-the-art methods on FotW and LFWA datasets. In the future, we will evaluate the performance of our proposed architecture with more tasks of face attributes prediction. We also plan to apply the attention module to more fully connected layers or convolution layers and try to use dynamic weights for performing more face attributes' prediction tasks.

Author Contributions: Conceptualization, D.F. and H.K.; methodology, D.F. and H.K.; software, D.F.; formal analysis, J.K.; writing—original draft preparation, D.F. and Y.L.; writing—review and editing, H.K., J.K. and H.Q.; visualization, Y.L.; funding acquisition, H.Q.

Funding: This work was supported by the National Natural Science Foundation of China (Grant No. 91748202).

Conflicts of Interest: The authors declare that they have no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

DCNNs	Deep convolutional neural networks
CNN	Convolutional neural networks
DCNN	Deep convolutional neural network
RNNs	Recurrent neural networks
LSTM	Long short term memory
TMTL	Traditional multi task learning
FotW	Faces of the world
LFWA	Labeled faces in the wild-a
t-SNE	T-distributed stochastic neighbor embedding

References

1. Kumar, N.; Berg, A.C.; Belhumeur, P.N.; Nayar, S.K. Attribute and simile classifiers for face verification. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 29 September–2 October 2009; pp. 365–372.
2. Song, F.; Tan, X.; Chen, S. Exploiting relationship between attributes for improved face verification. *Comput. Vis. Image Underst.* **2014**, *122*, 143–154. [CrossRef]
3. Vaquero, D.A.; Feris, R.S.; Tran, D.; Brown, L.; Hampapur, A.; Turk, M. Attribute-based people search in surveillance environments. In Proceedings of the 2009 Workshop on Applications of Computer Vision (WACV), Snowbird, UT, USA, 7–8 December 2009; pp. 1–8.
4. Li, D.; Chen, X.; Huang, K. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In Proceedings of the 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR), Kuala Lumpur, Malaysia, 3–6 November 2015; pp. 111–115.
5. Glauner, P.O. Deep convolutional neural networks for smile recognition. *arXiv* **2015**, arXiv:1508.06535.
6. Chen, J.; Ou, Q.; Chi, Z.; Fu, H. Smile detection in the wild with deep convolutional neural networks. *Mach. Vis. Appl.* **2017**, *28*, 173–183. [CrossRef]
7. Nian, F.; Li, L.; Li, T.; Xu, C. Robust gender classification on unconstrained face images. In Proceedings of the 7th International Conference on Internet Multimedia Computing and Service, Zhangjiajie City, China, 19–21 August 2015; p. 77.
8. Mansanet, J.; Albiol, A.; Paredes, R. Local deep neural networks for gender recognition. *Pattern Recognit. Lett.* **2016**, *70*, 80–86. [CrossRef]
9. Rothe, R.; Timofte, R.; van Gool, L. Dex: Deep expectation of apparent age from a single image. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Santiago, Chile, 7–13 December 2015; pp. 10–15.
10. Liu, X.; Li, S.; Kan, M.; Zhang, J.; Wu, S.; Liu, W.; Han, H.; Shan, S.; Chen, X. Agenet: Deeply learned regressor and classifier for robust apparent age estimation. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Santiago, Chile, 7–13 December 2015; pp. 16–24.
11. Brody, L.R.; Hall, J.A. Gender and emotion in context. *Handb. Emot.* **2008**, *3*, 395–408.
12. Bilinski, P.; Dantcheva, A.; Brémond, F. Can a smile reveal your gender? In Proceedings of the 2016 International Conference of the Biometrics Special Interest Group (BIOSIG), Darmstadt, Germany, 21–23 September 2016; pp. 1–6.
13. Dantcheva, A.; Brémond, F. Gender estimation based on smile-dynamics. *IEEE Trans. Inf. Forensics Secur.* **2017**, *12*, 719–729. [CrossRef]
14. Desai, S.; Upadhyay, M.; Nanda, R. Dynamic smile analysis: Changes with age. *Am. J. Orthod. Dentofac. Orthop.* **2009**, *136*, 310. [CrossRef]
15. Guo, G.; Dyer, C.R.; Fu, Y.; Huang, T.S. Is gender recognition affected by age? In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops, Kyoto, Japan, 27 September–4 October 2009; pp. 2032–2039.
16. Bengio, Y.; Courville, A.; Vincent, P. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1798–1828. [CrossRef]
17. He, R.; Wu, X.; Sun, Z.; Tan, T. Learning invariant deep representation for nir-vis face recognition. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017.
18. Rifai, S.; Bengio, Y.; Courville, A.; Vincent, P.; Mirza, M. Disentangling factors of variation for facial expression recognition. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 808–822.
19. Caruana, R. Multitask learning. *Mach. Learn.* **1997**, *28*, 41–75. [CrossRef]
20. Gkioxari, G.; Hariharan, B.; Girshick, R.; Malik, J. R-cnns for pose estimation and action detection. *arXiv* **2014**, arXiv:1406.5212.
21. Eigen, D.; Fergus, R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 11–18 December 2015; pp. 2650–2658.

22. Misra, I.; Shrivastava, A.; Gupta, A.; Hebert, M. Cross-stitch networks for multi-task learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 3994–4003.
23. Kokkinos, I. Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6129–6138.
24. Mallya, A.; Lazebnik, S. Packnet: Adding multiple tasks to a single network by iterative pruning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7765–7773.
25. Kim, E.; Ahn, C.; Torr, P.H.; Oh, S. Deep virtual networks for memory efficient inference of multiple tasks. *arXiv* **2019**, arXiv:1904.04562.
26. Levi, G.; Hassner, T. Age and gender classification using convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Boston, MA, USA, 7–12 June 2015; pp. 34–42.
27. Liu, Z.; Luo, P.; Wang, X.; Tang, X. Deep learning face attributes in the wild. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 11–18 December 2015; pp. 3730–3738.
28. Ranjan, R.; Sankaranarayanan, S.; Castillo, C.D.; Chellappa, R. An all-in-one convolutional neural network for face analysis. In Proceedings of the 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), Washington, DC, USA, 30 May–3 June 2017; pp. 17–24.
29. Hyun, C.; Seo, J.; Lee, K.E.; Park, H. Multi-attribute recognition of facial images considering exclusive and correlated relationship among attributes. *Appl. Sci.* **2019**, *9*, 2034. [CrossRef]
30. Yoo, B.; Kwak, Y.; Kim, Y.; Choi, C.; Kim, J. Deep facial age estimation using conditional multitask learning with weak label expansion. *IEEE Signal Process. Lett.* **2018**, *25*, 808–812. [CrossRef]
31. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv* **2014**, arXiv:1409.0473.
32. Luong, M.-T.; Pham, H.; Manning, C.D. Effective approaches to attention-based neural machine translation. *arXiv* **2015**, arXiv:1508.04025.
33. Yang, Z.; Hu, Z.; Deng, Y.; Dyer, C.; Smola, A. Neural machine translation with recurrent attention modeling. *arXiv* **2016**, arXiv:1607.05108.
34. Tang, Y.; Srivastava, N.; Salakhutdinov, R.R. Learning generative models with visual attention. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 1808–1816.
35. Xiao, T.; Xu, Y.; Yang, K.; Zhang, J.; Peng, Y.; Zhang, Z. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 842–850.
36. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 6 2015; pp. 2048–2057.
37. Zhao, B.; Wu, X.; Feng, J.; Peng, Q.; Yan, S. Diversified visual attention networks for fine-grained object classification. *IEEE Trans. Multimedia* **2017**, *19*, 1245–1256. [CrossRef]
38. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
39. Escalera, S.; Torres, M.T.; Martinez, B.; Baró, X.; Escalante, H.J.; Guyon, I.; Tzimiropoulos, G.; Corneou, C.; Oliu, M.; Bagheri, M.A.; et al. Chalearn looking at people and faces of the world: Face analysis workshop and challenge 2016. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1–8.
40. Huang, G.B.; Mattar, M.; Berg, T.; Learned-Miller, E. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In Proceedings of the Workshop on faces in'Real-Life'Images: Detection, Alignment, and Recognition, Marseille, France, 17–20 October 2008.
41. Chollet, F. Keras. 2015. Available online: https://github.com/keras-team/keras (accessed on 1 March 2018).
42. Higgins, I.; Amos, D.; Pfau, D.; Racaniere, S.; Matthey, L.; Rezende, D.; Lerchner, A. Towards a definition of disentangled representations. *arXiv* **2018**, arXiv:1812.02230.

43. Maaten, L.V.D.; Hinton, G. Visualizing data using t-sne. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
44. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Li, F.-F. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
45. Parkhi, O.M.; Vedaldi, A.; Zisserman, A. Deep face recognition. *BMVC* **2015**, *1*, 6.
46. Han, H.; Jain, A.K.; Wang, F.; Shan, S.; Chen, X. Heterogeneous face attribute estimation: A deep multi-task learning approach. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 2597–2609. [CrossRef] [PubMed]
47. Li, C.; Kang, Q.; Ge, G.; Song, Q.; Lu, H.; Cheng, J. Deepbe: Learning deep binary encoding for multi-label classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 39–46.
48. Zhang, K.; Tan, L.; Li, Z.; Qiao, Y. Gender and smile classification using deep convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 34–38.
49. Zhang, N.; Paluri, M.; Ranzato, M.; Darrell, T.; Bourdev, L. Panda: Pose aligned networks for deep attribute modeling. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1637–1644.
50. Zhuang, N.; Yan, Y.; Chen, S.; Wang, H. Multi-task learning of cascaded cnn for facial attribute classification. In Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018; pp. 2069–2074.
51. Zhuang, N.; Yan, Y.; Chen, S.; Wang, H.; Shen, C. Multi-label learning based deep transfer neural network for facial attribute classification. *Pattern Recognit.* **2018**, *80*, 225–240. [CrossRef]