Cross-Task Representation Learning for Anatomical Landmark Detection

Fu, Zeyu; Jiao, Jianbo; Suttie, Michael; Noble, J. Alison

Link to publication on Research at Birmingham portal

Cross-Task Representation Learning for Anatomical Landmark Detection

Zeyu Fu[1], Jianbo Jiao[1], Michael Suttie[2], and J. Alison Noble[1]

[1] Department of Engineering Science, University of Oxford, Oxford, UK
`zeyu.fu@eng.ox.ac.uk`
[2] Nuffield Department of Women's and Reproductive Health, University of Oxford, Oxford, UK

Abstract. Recently, there is an increasing demand for automatically detecting anatomical landmarks which provide rich structural information to facilitate subsequent medical image analysis. Current methods related to this task often leverage the power of deep neural networks, while a major challenge in fine tuning such models in medical applications arises from insufficient number of labeled samples. To address this, we propose to regularize the knowledge transfer across source and target tasks through cross-task representation learning. The proposed method is demonstrated for extracting facial anatomical landmarks which facilitate the diagnosis of fetal alcohol syndrome. The source and target tasks in this work are face recognition and landmark detection, respectively. The main idea of the proposed method is to retain the feature representations of the source model on the target task data, and to leverage them as an additional source of supervisory signals for regularizing the target model learning, thereby improving its performance under limited training samples. Concretely, we present two approaches for the proposed representation learning by constraining either final or intermediate model features on the target model. Experimental results on a clinical face image dataset demonstrate that the proposed approach works well with few labeled data, and outperforms other compared approaches.

Keywords: Anatomical landmark detection · Knowledge transfer

1 Introduction

Accurate localization of anatomical landmarks plays an important role for medical image analysis and applications such as image registration and shape analysis [3]. It also has the potential to facilitate the early diagnosis of Fetal Alcohol Syndrome (FAS) [11]. An FAS diagnosis requires the identification of at least 2 of 3 cardinal facial features; a thin upper lip, a smooth philtrum and a reduced palpebral fissure length (PFL) [10], which means that even a small inaccuracy in the PFL measurement can easily result in misdiagnosis. Conventional approaches for extracting anatomical landmarks mostly rely on manual examination, which is tedious and subject to inter-operator variability. To automate landmark detection, recent methods in computer vision [16, 22, 25] and

medical image analysis [3, 11, 26] have extensively relied on convolutional neural networks (CNN) for keypoint regression. Although these models have achieved promising performance, this task still remains challenging especially when handling the labeled data scarcity in medical domain, due to expensive and inefficient annotation process. Transfer learning, in particular fine-tuning pre-trained models from similar domains have been widely used to help reduce over-fitting by providing a better initialization [17]. However, merely fine-tuning the existing parameters may arguably lead to a suboptimal local minimum for the target task, because much knowledge of the pre-trained model in the feature space is barely explored [13, 14]. To address this, we explore the following question: *Is it possible to leverage the abundant knowledge from a domain-similar source task to guide or regularize the training of the target task with limited training samples?*

We investigate this hypothesis via cross-task representation learning, where "cross-task" here means that the learning process is made between the source and target tasks with different objectives. In this work, the proposed cross-task representation learning approach is illustrated for localizing anatomical landmarks in clinical face images to facilitate early recognition of fetal alcohol syndrome [1], where the source and target tasks are face recognition and landmark detection. Intuitively, the proposed representation learning is interpreted as preserving feature representations of a source classification model on the target task data, which serves as a regularization constraint for learning the landmark detector. Two approaches for the proposed representation learning are developed by constraining either final or intermediate network features on the target model.

Related Work. Current state-of-art methods formulate the landmark detection as a CNN based regression problem, including two main frameworks: direct coordinate regression [6, 24] and heatmap regression [16, 22]. Heatmap regression usually outperforms its counterpart as it preserves high spatial resolution during regression. In medical imaging, several CNN architectures have been developed based on attention mechanisms [3, 26], and cascaded processing [23] for the enhancement of anatomical landmark detection. However, the proposed learning approach in this paper focuses on internally enriching the feature representations for the keypoint localization without complicating the network design.

Among existing knowledge transfer approaches, fine-tuning [22], as a standard practice initializes from a pre-trained model and shifts its original capability towards a target task, where a small learning rate is often applied and some model parameters may need to be frozen to avoid overfitting. However, empirically modifying the existing parameters may not generalize well over the small training dataset. Knowledge distillation originally proposed for model compression [9] is also related to knowledge transfer. This technique has been successfully extended and applied to various applications, including hint learning [20], incremental learning [5, 15], privileged learning [4], domain adaptation [7] and human expert knowledge distillation [19]. These distillation methods focused on training a compact model by operating the knowledge transfer across the same

tasks [9, 19, 20]. However, our proposed learning approach aims to regularize the transfer learning across different tasks.

Contributions. We propose a new deep learning framework for anatomical landmark detection under limited training samples. The main contributions are: (1) we propose a cross-task representation learning approach whereby the feature representations of a pre-trained classification model are leveraged for regularizing the optimization of landmark detection. (2) We present two approaches for the proposed representation learning by constraining either final or intermediate network features on the target task data. In addition, a cosine similarity inspired by metric learning is adopted as a regularization loss to transfer the relational knowledge between tasks. (3) We experimentally show that the proposed learning approach performs well in anatomical landmark detection with limited training samples and is superior to standard transfer learning approaches.

2   Method

In this section, we first present the problem formulation of anatomical landmark detection, and then describe the design of the proposed cross-task representation learning to address this task.

2.1   Problem Formulation

In this paper, our target task is anatomical landmark detection, which aims to localize a set of pre-defined anatomical landmarks given a facial image. Let $\mathcal{D}^t = \{\mathbf{I}_i^t, \mathbf{p}_i^t\}_{i=1}^{N_t}$ be the training dataset with N_t pairs of training samples in the target domain. $\mathbf{I}_i^t \in \mathbb{R}^{H \times W \times 3}$ represents a 2D RGB image with height H and width W, $\mathbf{p}_i^t = [(x_1, y_1), (x_2, y_2), ..., (x_K, y_K)] \in \mathbb{R}^{2 \times K}$ denotes the corresponding labeled landmark coordinates, and K is the number of anatomical landmarks ($K = 14$). We formulate this task using heatmap regression, inspired by its recent success in keypoint localization [16, 22]. Following prior work [16], we downscale the labeled coordinates to 1/4 of the input size ($\mathbf{p}_i^t = \mathbf{p}_i^t/4$), and then transform them to a set of heatmaps $\mathbf{G}_i^t \in \mathbb{R}^{(H/4) \times (W/4) \times K}$. Each heatmap $\mathbf{g}_k^t \in \mathbb{R}^{(H/4) \times (W/4)}, k \in \{1, ..., K\}$ is defined as a 2D Gaussian kernel centered on the k-th landmark coordinate (x_k, y_k). The (a, b) entry of $\mathbf{g}_k^t$ is computed as $\mathbf{g}_k^t(a, b) = \exp(-\frac{(a-x_k)^2 + (b-y_k)^2}{2\sigma^2})$, where σ denotes the kernel width ($\sigma = 1.5$ pixels). Consequently, the goal is to learn a network which regresses each input image to a set of heatmaps, based on the updated dataset $\mathcal{D}^t = \{\mathbf{I}_i^t, \mathbf{G}_i^t\}_{i=1}^{N_t}$.

For this regression problem, most state-the-of-the-art methods [22, 25] follow the encoder-decoder design, in which a pre-trained network (e.g. ResNet50 [8]) is usually utilized in the encoder for feature extraction, and then the entire network or only the decoder is fine-tuned during training. However, due to the limited number of training samples in our case, merely relying on standard fine-tuning may not always provide a good localization accuracy. Therefore, we present the proposed solution to address this problem in the next section.
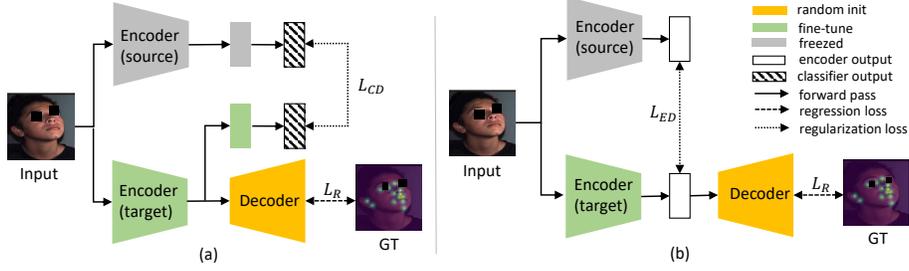
Fig. 1. Illustration of proposed approaches for learning the anatomical landmark detection models, where (a) presents the regularization constraint on the final layer output (L_{CD}), and (b) is to constrain the predictions on the encoder output (L_{ED}).

2.2 Cross-Task Representation Learning

Overview. Fig. 1 depicts the overall design of the proposed cross-task representation learning approach. Firstly, the source model pre-trained on a face classification task is operated in the inference mode to predict rich feature representations from either classification or intermediate layers for the target task data. The target model is then initialized from the source model and extended with a task-specific decoder for the task of landmark detection (L_R). Obtained feature representations are then transferred by regularization losses (L_{CD} or L_{ED}) for regularizing the target model learning.

Source Model. We consider a pre-trained face classification network as our source model, since generic facial representations generated from this domain-similar task have been demonstrated to be helpful for other facial analysis [21]. Formally, let $\mathcal{S}_{\theta_1,\theta_2} : \mathbb{R}^{H \times W \times 3} \to \mathbb{R}^C$ be the source network for a face classification task with C classes, where θ_1 and θ_2 are the learnable parameters. The network consists of a feature extractor (encoder) $f_{\theta_1}^s : \mathbb{R}^{H \times W \times 3} \to \mathbb{R}^d$ and a classifier $g_{\theta_2}^s : \mathbb{R}^d \to \mathbb{R}^C$, where d denotes the dimensionality of the encoder output. A cross-entropy loss is typically used to train the network $\mathcal{S}_{\theta_1,\theta_2} := g_{\theta_2}^s(f_{\theta_1}^s(\mathbf{I}))$ which maps a facial image to classification scores based on a rich labeled dataset $\mathcal{D}^s$. In practice, we adopt a pre-trained ResNet-50 [8] model from VGGFace2 [2] for the source network. Other available deep network architectures could also be utilized for this purpose.

Target Model. For the task of heatmap regression, the target network $\mathcal{T}_{\theta_1,\theta_2}$ is firstly initialized from the pre-trained source network. We then follow the design of [22], employing three deconvolutional layers after the encoder output $f_{\theta_1}^t(\mathbf{I})$ to recover the desired spatial resolution, where each layer has the dimension of 256 and 4×4 kernel with the stride of 2. Finally, a 1×1 convolutional layer is added to complete this task-specific decoder $h_{\theta_3}^t(f_{\theta_1}^t(\mathbf{I})) : \mathbb{R}^d \to \mathbb{R}^{(H/4) \times (W/4) \times K}$. The
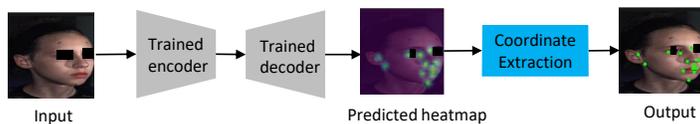
Fig. 2. Illustration of proposed framework for testing landmark detection models.

primary learning objective is to minimize the following loss between the decoder outputs and the labeled heatmaps,

$$L_R = \frac{1}{N_t} \sum_{i=1}^{N_t} \left\| \mathbf{G}_i^t - h_{\theta_3}^t(f_{\theta_1}^t(\mathbf{I}_i^t)) \right\|_F^2 \tag{1}$$

where F denotes the Frobenius norm.

Regularized Knowledge Transfer. Motivated by knowledge distillation, we aim to regularize the network training by directly acquiring the source model's predictions for the target task data $\mathcal{D}^t$, which are further transferred through a regularization loss L_D. Hence, the total loss is defined as,

$$L = L_R + \lambda L_D \tag{2}$$

where λ is a weighting parameter. If $\lambda = 0$, the knowledge transfer becomes standard fine-tuning, as no regularization is included.

For the design of L_D, we firstly consider constraining the distance between the final layer outputs of the two networks, as shown in Fig. 1 (a). Similar to the distillation loss in [9], we use a temperature parameter μ with $softmax$ function to smooth the predictions, but the original cross-entropy function is replaced by the following term,

$$L_{CD} = \frac{1}{N_t} \sum_{i=1}^{N_t} \left\| softmax \left(\frac{g_{\theta_2}^s(f_{\theta_1}^s(\mathbf{I}_i^t))}{\mu} \right) - softmax \left(\frac{g_{\theta_2}^t(f_{\theta_1}^t(\mathbf{I}_i^t))}{\mu} \right) \right\|_2^2 . \tag{3}$$

The purpose of this design of L_{CD} is to directly align the facial embeddings between instances, instead of preserving the original classification ability.

Moreover, we consider matching the features maps produced from both encoders as another choice, as shown in Fig. 1 (b). Motivated by the work in [18], we adopt the cosine similarity for the feature alignment as described below,

$$L_{ED} = 1 - \sum_{i=1}^{N_t} \cos(f_{\theta_1}^s(\mathbf{I}_i^t), f_{\theta_1}^t(\mathbf{I}_i^t)). \tag{4}$$

We conjecture that penalizing higher-order angular differences in this context would help transfer the relational information across different tasks, and also give

more flexibility for the target model learning. Besides, both regularization terms can be combined together to regularize the learning process. Different approaches of the proposed learning strategy will be evaluated in the experimental section.

During inference, as shown in Fig. 2, only the trained target model is used to infer the heatmaps, and each of them is further processed via an *argmax* function to obtain final landmark locations.

3   Experiments

3.1   Dataset and Implementation Details

We evaluate the proposed approach for extracting facial anatomical landmarks. Images used for training and test datasets were collected by the Collaborative Initiative on Fetal Alcohol Spectrum Disorders (CIFASD)[3], a global multidisciplinary consortium focused on furthering the understanding of FASD. It contains subjects from 4 sites across the USA, aged between 4 and 18 years. Each subject was imaged using a commercially available static 3D photogrammetry system from 3dMD[4]. For this study, we utilize the high-resolution 2D images captured during 3D acquisition, which are used as UV mapped textures for the 3D surfaces.

Specifically, we acquired in total 1549 facial images annotated by an expert, and randomly split them into training/validation set (80%), and test set (20%). All the images were cropped and resized to 256×256 for the network training and evaluation. Standard data augmentation was performed with randomly horizontal flip (50%) and scaling (0.8). During training, the Adam optimizer [12] was used for the optimization with the mini-batch size of 2 for 150 epochs. A polynomial decay learning rate was used with the initial value of 0.001. Parameters of λ and μ used in (2) and (3) were set to 0.002 and 2, respectively.

3.2   Evaluation Metrics

For the evaluation, we firstly employ the Mean Error (ME), which is a commonly-used evaluation metric in the task of facial landmark detection. It is defined as, $ME = \frac{1}{N_e} \sum_{i=1}^{N_e} \frac{1}{K} \|\mathbf{p}_i - \hat{\mathbf{p}}_i\|_2$, where N_e is the number of images in the test set, and $\mathbf{p}_i$ and $\hat{\mathbf{p}}_i$ denote the manual annotations and predictions, respectively. Note that the original normalization factor measured by inter-ocular distance (Euclidean distance between outer eye corners) is not included in this evaluation, due to the unavailable annotations for the other eye, as illustrated in Fig. 3. In addition, we use the Cumulative Errors Distribution (CED) curve with the metrics of Area-Under-the-Curve (AUC) and Failure Rate (FR), where a failure case is considered if the point-to-point Euclidean error is greater than 1.2. Higher scores of AUC or lower scores of FR demonstrate the larger proportion of the test set is well predicted.

---

[3] https://cifasd.org/
[4] http://www.3dmd.com/
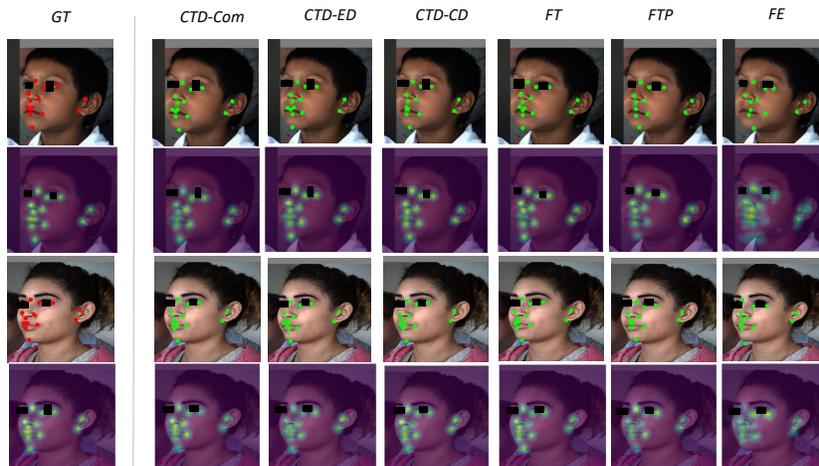
Fig. 3. Qualitative performance of landmark prediction and heatmap regression on the test set. Subjects' eyes are masked for privacy preservation. Better viewed in color.
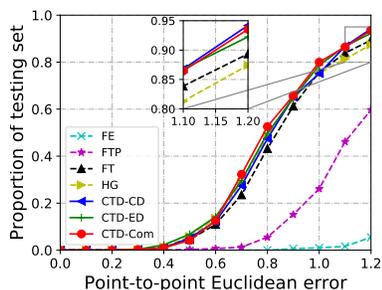
Fig. 4. Evaluation of CED curve on the test set. Better viewed in color.

Method	ME ± SD	FR	AUC
FE [22]	1.822±0.501	94.52%	0.01
FTP [22]	1.161±0.261	40.32%	0.10
FT [22]	0.858±**0.238**	10.65%	0.29
HG [16]	0.879±0.386	12.58%	0.30
CTD-CD	0.842±0.246	**5.81%**	0.31
CTD-ED	0.830±0.245	7.74%	**0.32**
CTD-Com	**0.829**±0.253	6.45%	**0.32**

Table 1. Quantitative evaluation on the test set.

3.3   Results and Discussions

To verify the effectiveness of the proposed cross-task representation learning (CTD) approach, we compare to a widely-used CNN model: stacked Hourglass (HG) [16] and three variants of fine-tuning [22] without regularization ($\lambda = 0$): Feature Extraction (FE) with freezing the encoder, Fine Tuning Parts (FTP) without freezing the final convolutional layer of the encoder, and Fine Tuning (FT) without freezing any layer. In addition, we present an ablation study to examine the significance of each approach in our proposed CTD, including the regularization on the classifier output (CTD-CD), the regularization on the encoder output (CTD-ED), and the regularization on both outputs (CTD-Com).

Fig. 3 shows the qualitative comparisons between different models on the test set. As we can see, the predicted landmarks from the proposed methods generally achieve the better alignment with the ground truth (the first left column) than the others, and seem to be more robust to difficult pixels especially when

landmarks are in close proximity (upper lip). One possible reason is that feature representations generated from the source model encode richer facial semantics, which make landmark spatial locations more discriminative. Furthermore, the visualization of predicted heatmaps explains how each compared model responds to the desired task. We observe that our cross-task representation learning can effectively suppress spurious responses and improve the feature confidence in related regions, so that more accurate predictions can be achieved.

On the other hand, Table 1 summarizes the quantitative evaluation by reporting the statistics for each model. Fig. 4 depicts the CED curve which provides an intuitive understanding of the overall performance of the compared models. These evaluations above demonstrate that the proposed methods consistently outperform standard fine-tuning solutions. Moreover, CTD-ED performs slightly better than CTD-CD considering the scores of ME and AUC. This may be explained by the fact that features from intermediate layers are not only semantic, and also contain to some extent structural information which is beneficial for localization [7]. Interestingly, CTD-Com using both regularization losses achieves similar results in CTD-ED, as a result, CTD-ED may be considered as a better choice for the regularization of transfer learning.

4   Conclusions

In this paper, we presented a new cross-task representation learning approach to address the problem of anatomical landmark detection where labeled training data is limited. The proposed learning approach considered reusing the knowledge from a domain-similar source task as a regularization constraint for learning the target landmark detector. Moreover, several regularization constraints for the proposed learning approach were considered. Experimental results suggested that the proposed learning approach works well with limited training samples and outperforms other compared solutions. The proposed approach can be potentially applied to other related applications in the clinical domain where the target task has small training set and the source task data is not accessible.

References

1. Astley, S.J.: Palpebral fissure length measurement: accuracy of the FAS facial photographic analysis software and inaccuracy of the ruler. Journal of Population Therapeutics and Clinical Pharmacology **22**(1), e9–e26 (2015)

2. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. In: IEEE International Conference on Automatic Face Gesture Recognition. pp. 67–74 (2018)

3. Chen, R., Ma, Y., Chen, N., Lee, D., Wang, W.: Cephalometric landmark detection by attentive feature pyramid fusion and regression-voting. In: Medical Image Computing and Computer Assisted Intervention (MICCAI). pp. 873–881 (2019)

4. David Lopez-Paz, Lon Bottou, B.S.V.V.: Unifying distillation and privileged information pp. 1–10 (2016)

5. Dhar, P., Singh, R.V., Peng, K.C., Wu, Z., Chellappa, R.: Learning without memorizing. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)

6. Feng, Z.H., Kittler, J., Awais, M., Huber, P., Wu, X.J.: Wing loss for robust facial landmark localisation with convolutional neural networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)

7. Gupta, S., Hoffman, J., Malik, J.: Cross modal distillation for supervision transfer. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)

8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)

9. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. In: Conference on Neural Information Processing Systems (NeurIPS) Workshops (2015)

10. Hoyme, H.E., May, P.A., Kalberg, W.O., et al.: A practical clinical approach to diagnosis of fetal alcohol spectrum disorders: Clarification of the 1996 institute of medicine criteria. Pediatrics **115**(1), 39–47 (2006)

11. Huang, R., Suttie, M., Noble, J.A.: An automated CNN-based 3D anatomical landmark detection method to facilitate surface-based 3D facial shape analysis. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI) Workshops. pp. 163–171 (2019)

12. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: Proc. of International Conference on Learning Representations (ICLR). pp. 1–15 (2015)

13. Li, X., Xiong, H., Wang, H., Rao, Y., Liu, L., Chen, Z., Huan, J.: DELTA: DEep learning transfer using feature map with attention for convolutional networks. In: Proc. of International Conference on Learning Representations (ICLR). pp. 1–13 (2019)

14. Li, X., Grandvalet, Y., Davoine, F.: Explicit inductive bias for transfer learning with convolutional networks. In: International Conference on Machine Learning (ICML). vol. 80, pp. 2830–2839 (2018)

15. Li, Z., Hoiem, D.: Learning without forgetting. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(12), 2935–2947 (2018)

16. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European Conference on Computer Vision (ECCV). pp. 483–499 (2016)

17. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering **22**(10), 1345–1359 (2010)

18. Park, W., Kim, D., Lu, Y., Cho, M.: Relational knowledge distillation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)

19. Patra, A., Cai, Y., Chatelain, P., Sharma, H., Drukker, L., Papageorghiou, A.T., Noble, J.A.: Efficient ultrasound image analysis models with sonographer gaze assisted distillation. In: Proc. of Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 394–402 (2019)

20. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. In: Proc. of International Conference on Learning Representations (ICLR). pp. 1–13 (2015)
21. Wiles, O., Koepke, A., Zisserman, A.: Self-supervised learning of a facial attribute embedding from video. In: British Machine Vision Conference (BMVC) (2018)
22. Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. In: European Conference on Computer Vision (ECCV). pp. 472–487 (2018)
23. Zhang, J., Liu, M., Shen, D.: Detecting anatomical landmarks from limited medical imaging data using two-stage task-oriented deep neural networks. IEEE Transactions on Image Processing **26**(10), 4753–4764 (2017)
24. Zhang, Z., Luo, P., Loy, C.C., Tang, X.: Learning deep representation for face alignment with auxiliary attributes. IEEE Transactions on Pattern Analysis and Machine Intelligence **38**(5), 918–930 (2016)
25. Zhao, Y., Liu, Y., Shen, C., Gao, Y., Xiong, S.: MobileFAN: Transferring deep hidden representation for face alignment. Pattern Recognition **100**, 107–114 (2020)
26. Zhong, Z., Li, J., Zhang, Z., Jiao, Z., Gao, X.: An attention-guided deep regression model for landmark detection in cephalograms. In: Proc. of Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 540–548 (2019)