ABSTRACT  Age estimation of facial images is very challenging because of the complexity of face aging process and the difficulty of collecting and labeling data. A holistic regression model is subject to imbalanced training data, while a divide-and-conquer method highly depends on the effect of the age classification, which usually has boundary effect due to cross-age correlations. This paper proposes a simple but effective multi-task learning (MTL) network combining classification and regression for age estimation called CR-MT net, where classification acts as an auxiliary task to regression. MTL can boost the generalization performance of the age regression task by shared information representation learning of the two tasks. Compared to divide-and-conquer methods, our method performs MTL for two tasks with an end-to-end training, and has no error propagation from classification to regression. The holistic regression model in CR-MT net does not meet the boundary effect, and can fit the heterogeneous or unbalanced age data more accurately with the aid of a good age data partition in classification. We evaluate two age grouping techniques to find a good data partition, and diagnose various factors which can influence the performance of the CR-MT net by extensive experiments. The CR-MT net is verified in three public datasets, and achieves state-of-art results.

INDEX TERMS  Age estimation, multi-task learning, deep learning.
as well as appropriate learning methods like fuzzy LDA [8], support vector machine [5], neural network [6] to predict age. However, the design of features as well as the selection of learning methods needs strong experiences and many efforts which deeply affect the performance of age estimation. Recently, deep learning has achieved great success in many computer vision tasks due to its strong capability of feature learning and representation. Some work [13]–[17] attempt to use deep learning models for age classification or regression. Compared to traditional age estimation methods, both the feature representation and prediction tasks can be integrated into a deep learning framework by end-to-end network training, thus the prediction accuracy is improved. However, most of these methods don’t take into account the relationship between different age categories and the heterogeneous data distribution. In order to model the cross-age correlations, several works [6], [18], [19] associate each facial image with a label distribution that covers a certain number of age labels, and predict age by label distribution learning. It has obtained promising results, but the label distribution models used cannot represent various cross-age correlations between overall age range.

Some recent works [7], [20]–[24] combine classification and regression in a sequential way, and use multiple local regression models established in several age groups to model the heterogeneous age data. Compared to a global model, local models can fit the heterogeneous data better. However, the performance of these methods highly depends on the effect of age classification, and an improper age data partition cannot well deal with the issues of the boundary effect on classification and the heterogeneous data distribution.

Multi-task learning (MTL) [25] is an inductive transfer learning paradigm inspired by human learning, in which the knowledge acquired by learning related tasks can be applied to the learning of new task. The objective of multi-task learning is to improve the generalization performance of all tasks by using shared information relevant to multiple related tasks. Moreover, owing to the shared representations, the number of data sources needed and the size of the overall model parameters are reduced. Deep MTL has been applied successfully to many computer vision applications. Wang et al. [26] proposes a multi-task CNN network combining classification and regression for object pose estimation, and the performance has a significant improvement compared to single-task CNN network. Up to date, there is only few attempts [17], [27], [28] for age estimation using Deep MTL. The nature of the age estimation task is very similar to that of pose estimation task [26] since both tasks intend to estimate a continuous variable, but the age estimation is more challenging due to the particularity of human face aging and the heterogeneous data distribution.

Inspired by the work [26] introduced above, we design a multi-task learning network called CR-MT for age estimation. The contributions are as follows:

1) A simple deep multi-task network combining classification and regression is proposed. The shared information representation of the multi-task learning helps to alleviate the influence of the imbalanced training age data onto age regression, and thus boosts the generalization performance of the age regression task. To our knowledge, it is the first time to combine classification and regression for MTL of age estimation.

2) We try two different techniques to find a good data partition for classification. One is based on modified adjacent age grouping, the other focuses on K-means clustering. A good data partition can help the regression model fit the heterogeneous age data well.

3) We diagnose various factors which affect the performance of our net by extensive experiments, and find the optimal weight setting and shared layers of two tasks.

4) Our CR-MT net achieves the state-of-the-art results on multiple public datasets, including Morph Album II [29], Webface [30] and CACD [31].

II. RELATED WORK

A holistic model for age estimation is often subject to the issues of heterogeneous data distribution and imbalanced training data. So recent work [7], [12], [21], [22], [32]–[36] tend to learn multiple local models in a divide-and-conquer strategy. For example, Chen et al. [35] proposes a Ranking-CNN, which contains a series of binary classification CNNs trained with ordinal age labels, and the binary outputs are aggregated for the final age estimation. More generally, these work adopt a hierarchical-based technique with a classification stage followed by a local regression stage. Local regression models can fit the heterogeneous age data than a holistic model. For example, Sawant and Bhurchandi [36] proposes a hierarchical Gaussian process framework for coarse-to-fine age estimation, which applies multi-class Gaussian process classifier to classify the input images into different age groups followed by a warped Gaussian process regression to model group specific aging patterns. However, most of these work [21], [22], [32], [33] usually adopt a heuristic data partition in age grouping, and do not consider the issue of the correlation between adjacent ages and the heterogeneous data distribution. Liu et al. [37] exploits the smoothness for adjacent age groups by a group-aware deep feature learning (GA-DFL) approach which employs an overlapped coupled learning method. However, this method increases intra-group variances and subtracts inter-group variances, which is adverse for age classification. Yang et al. [38] proposes a compact model called SSR-Net. The method first performs multi-class classification with multiple stages in a coarse-to-fine strategy, then uses the classification results of age regression. Considering class imbalance and the correlation between adjacent ages, they assign some sliding strides to each age class. To deal with the issues of heterogeneous data distribution, Huang et al. [12] proposes Soft-Margin Mixture of Regression (SMMR) method, which simultaneously finds homogeneous partitions in the joint input-output space using max-margin classification and learns a linear local regressor for each partition. Shen et al. [7] proposes a deep regression
forests (DRFs) model combining CNN and random forests, where the local input-output correlation at each leaf node is homogeneous, and CNN is used for local regression. Li et al. [20] proposes a continuity-aware probabilistic network consisting of local regressors and gating networks. Local regressors model the heterogeneous data, while gating networks learn continuity-aware gating functions to weight the results of local regressors.

Compared to these divide-and-conquer methods, our method trains classification and regression simultaneously by MTL with an end-to-end manner. The regression model in our work is a holistic one, which does not meet the boundary effect on classification. Most importantly, owing to the shared information representation learning of the two tasks, the regression model can finely fit the heterogeneous age data or the unbalanced training data with the aid of a good data partition in age grouping of the classification. Moreover, since classification is only an auxiliary task for regression, there is no error propagation from classification to regression in our method, which is unavoidable in those divide-and-conquer methods.

The most related work to ours is those deep MTL methods. Both Niu et al. [14] and Tan et al. [39] transform the age ordinal regression problem into a series of binary classification sub-problems, and use a multiple output CNN to solve these sub-problems, where each output layer corresponds to a binary classification task, and all the tasks share the same intermediate layers. The difference is that Tan et al. [39] encodes the relationship among adjacent ages in age grouping, i.e. adjacent ages are grouped into the same group and each age corresponds to $n$ groups. Since the CNN network used in these methods is derived by only modifying the output layer from a single output CNN network to a multi-output network, where each output judges whether the input face image belongs to the corresponding age group or not, the MTL model is more like a holistic regression model. Therefore, these methods are still subjected to the issues of the imbalanced training data and heterogeneous data distribution. Both Xing et al. [27] and Yi et al. [17] uses CNN for simultaneously performing age estimation, gender and race classification. According to the reported results, the improvement on the age estimation accuracy is not significant compared to single-task deep networks. Only when the multiple tasks are closely related to each other, deep MTL can bring about the performance improvement of multiple tasks. Classification is the most relevant task to age regression due to the fact that the aim at both two tasks is to predict age, so the two tasks in our method can mutually reinforce each other by MTL.

III. METHOD

Generally for prediction of continuous variables, regression methods suffer from the issue of imbalanced training data and can get superior accuracy on homogeneously distributed datasets [35], [37], while classification methods are difficult to discriminate adjacent variable values, i.e. boundary effect in adjacent age groups, but they are effective for a coarse-grained prediction even if the dataset is imbalanced. So classification and regression methods are complementary in the case of imbalanced training data. Thus, the two tasks can play a mutually reinforcing role in age estimation through multi-task learning.

A. CR-MT NETWORK

The basic net of our CR-MT net is Alexnet, which consists of five convolution layers and three fully connected layers. Alexnet has lower accuracy compared with some deeper networks such as VGG and Res-Net, but it has the advantage of tiny structure and flexibility, also takes shorter time for training.

As shown in Fig.1, the architecture of our network consists of three parts, i.e. shared layers, classification branch and regression branch. Images are aligned to $256 \times 256$ and fed into network as input. The shared layers are to learn the shared feature representation of the two tasks. The two branches are parallelly connected to the shared layers to learn the task-specific features. Modified adjacent ages clustering and K-means clustering represent two methods of age grouping, adjacent ages clustering finds an optimal age grouping by adding sliding strides, and K-means clustering combines the age features and age label to obtain the homogeneous data partition. Specific details are introduced in section C.

Fig. 2 shows some feature maps from five convolution layers of three contrastive networks which are a single-classification network, a single-regression network and our CR-MT network respectively for a randomly selected image in Morph dataset. The two single-task networks also adopt Alexnet as the basic network. Here the age grouping for the classification branch in the CR-MT network adopts the adjacent ages clustering as an example. For the convenience to display, we select the first nine feature maps of each convolution layer. It can be illustrated from Fig.2 that the features extracted by our CR-MT network are sharper than those in single-task networks. The differences of feature maps in first three convolution layers are not obvious, it is because the low-level representation learns textures, contours, and shapes features, while high-level representation learns semantic features more suitable for specific tasks (e.g. eyes and mouth area). Red rectangles in Fig.2 highlight some contrasting features. Classification net extracts a small number of effective higher-dimensional features. The features extracted from the regression network are more abundant, but there are redundant noises, which affect the final result. However, the features extracted from our CR-MT network are more refined and more in line with the features of human faces, which indicate that our CR-MT network can learn more detailed features. This contrast demonstrates that the classification in our CR-MT method as an auxiliary task is indeed effective for boosting age estimation.

B. OBJECTIVE

In CR-MT net, we use the cross-entropy loss function [40] for the classification branch and ordinal regression loss
FIGURE 1. The main architecture of our network with basic network Alexnet. It is a parallel MTL network composed by classification branch and regression branch. The dotted and solid lines of blue, red and black represent the shared layer, the classification branch and the regression branch respectively. $l_i$ denotes the age groups and $y_i$ is precise age label. The two boxes in top of figure show two methods of age grouping respectively.

function [41] for the regression branch. Joint multi-task loss function is designed by weighting the losses of two branches.

1) CROSS-ENTROPY LOSS FUNCTION
The cross entropy characterizes the distance between the actual output probability and the expected output probability, and is defined as follows:

$$H(p, \hat{p}) = -\log \hat{p}$$

where $p$ and $\hat{p}$ represent the probabilities of true age classification and predicted age classification respectively for one image. Since the implementation of the cross-entropy loss uses one-hot encoding, i.e. $p_i = 1$ and 0 otherwise, where $i$ represents the i-th sample. classification loss can be simplified to:

$$L_{cls} = -\sum_i \log \hat{p}_i$$
2) ORDINAL REGRESSION LOSS FUNCTION
Many age classification methods divide the overall age range into independent age groups, and treat each age label independently, but in fact there is a strong correlation among adjacent ages. Ordinal regression [14] takes the relationship among adjacent ages into account, and transforms the age estimation problem into a series of simple binary sub-problems. It can be regarded as a compromise between classification and regression. The predicted age $\hat{y}_i$ can be calculated as follows:

$$\hat{y}_i = 1 + \sum_{k=1}^{C-1} f_k (X_i)$$

where $C$ represents the maximum of age, $f_k$ is the $k$-th binary classifier, $X_i$ is $i$-th test face. The final ordinal regression loss function is represented using Mean Absolute Error (MAE).

$$L_{\text{reg}} = \frac{1}{N} \sum_{i=1}^{N} |\hat{y}_i - y_i|$$

where $N$ is the total number of training facial images, $\hat{y}_i$ and $y_i$ is the estimated age and the true age respectively. We use the method in [14] to address ordinal regression problems.

3) JOINT MULTI-TASK LOSS FUNCTION
We use a multi-task loss $L$ to joint train the classification and regression networks as follows:

$$L_{\text{MTL}} = L_{\text{cls}} + \lambda L_{\text{reg}}$$

which consists of a classification error $L_{\text{cls}}$ and a regression error $L_{\text{reg}}$. $\lambda$ is the weight to balance the two tasks.

C. AGE GROUPING
Since human aging is a continuous and gradual process and faces of adjacent ages usually look alike, there is a strong correlation among adjacent ages. Many existing divide-and-conquer methods divide the age range into independent groups for classification and leave out of consideration the relationship between adjacent age groups. An arbitrary age grouping will lead to the boundary effect on classification, i.e. one sample near the boundary of two adjacent age groups can be classified as either one. Moreover, as mentioned above, the age feature space is heterogeneous according to age. A casual data partition can not deal with the issue of heterogeneous data distribution. Only with a homogeneous data partition, local regressions in individual homogeneous age groups can well fit the heterogeneous data.

However, it is hard to measure what data parathon is ‘good’ or not. In our CR-MT net, classification is an auxiliary task of playing a reinforce role on regression. Therefore, a good data partition in age grouping can help the regression model fit the heterogeneous data well. In order to obtain a homogeneous data partition and alleviate the boundary effect in classification, we try the following two strategies for age data grouping, and we analyse proposed scheme (the Adjacent ages clustering and the K-means clustering) both from insight analysis and experimental results analysis.

Adjacent ages clustering is a strategy which makes adjacent ages be grouped into the same group and then be regarded as an independent class in training stage. To explore the relationship between the real age and its adjacent ages, we conduct adjacent ages clustering procedure, which is the most obvious data grouping method since face aging is a slow, continuous and extremely non-stationary process with much randomness, and simply exploits the range of the age labels for clustering. However, we do not have exact knowledge of face aging, i.e. which age range would make the face aging homogeneous? So we determine the partition of the age intervals by trial-and-error.

Firstly, we determine an optimal group number and an initial age grouping by experiments. Then in order to alleviate the boundary effect, we slide the group intervals along the overall age range with a set of strides respectively, and perform training and testing for the CR-MT net repeatedly with the new age groupings generated with the strides. Finally, we find the optimal age grouping according to the test results. This trial-and-error method is simple and easy to implement, but it is cumbersome.

K-means clustering is a simple and effective non-supervised learning method. It groups the data according to the similarity of features. K-means clustering mainly contains two steps, i.e. computing clustering centroids and assigning the samples to the close centroid. Given initial clustering centroids, each sample is assigned to a clustering according to the distance from it to the centroid, and then the two steps above are repeated until convergence. Considering adjacent ages clustering only exploits the range of the age labels for clustering, we tend to concatenate feature vectors of images and corresponding age labels to improve clustering performance.

To improve the age features extracting performance of the single-regression network in our CR-MT net, whose basic net is Alexnet, we expect a data partition that the changing of the age feature is homogeneous by age in each group. We combine the output of the last fully-connected layer of the regression network and the age label as the feature space of the clustering. In fact, compared to adjacent ages clustering, the class label assigned by K-means clustering is more like a virtual label for supervised age classification in CR-MT net, and the grouping more directly reflects the character of the age data distribution.

D. IMPLEMENTATION DETAILS
1) TRAINING PARAMETER SETTING
The training of the CR-MT net is based on GPU mode in Caffe [42] framework. The network was trained with a weight decay of 0.0005 and a momentum of 0.9. The learning rate starts from 0.001 and is reduced by a factor of 10 along with the iterations. For all experiments, the network was initialized with the weights from the basic net trained on ImageNet.
2) NETWORK PARAMETER SETTING
Model-related hyper-parameters are set as follows: the shared layers involve 5 convolution layers, the weight $\lambda$ of the classification and regression task is set to 1.6, the age interval of the classification branches is set to 10 ages with the composition of 5 groups, and the sliding stride is set to 5, the number of k-means clustering is also 5.

3) ALTERNATE TRAINING STRATEGY
Through a large number of experiments, we find an optimal training strategy in multiple steps. We first train the classification net for 50k iterations by keeping regression branch fixed. Then, we train the regression net for 50k iterations by keeping classification branch fixed. Finally, the fixed layers of each branch are released and train 50k iterations jointly. Experimental results show that such alternate training is effective in improving the accuracy of age estimation.

IV. EXPERIMENTS
A. DATASET
We use three public datasets, Morph Album II [29], Webface [30] and CACD [31] in our experiments. Examples of the three datasets are illustrated in Fig.3. Besides, IMDB-WIKI [22] is used for pre-training.

![Table 1. The number of images in the three splits S1, S2 and S3 of Morph Setting II.](image)

| Gender | Black | Race | White | Others |
|--------|-------|------|-------|--------|
| Female | $1,1285$ | $2,1285$ | $3,3187$ | $3,1285$ |
| Male   | $1,3900$ | $2,3900$ | $3,2872$ | $3,1900$ |

FIGURE 3. Some examples of Morph, Webface and CACD. The left three columns are the original images selected randomly from three public datasets, and the right three columns are the corresponding images after face alignment.

Morph Album II contains 55134 images of 13618 people with different age, gender, and race. The age ranges from 16 to 77, and the average age is 33. Setting I is 5-fold cross-validation in whole dataset, where four folds are for training and the remaining one is for testing. Setting II adopts the segmentation method in [43]. In order to get a balanced training set in gender and race, [43] splits the dataset into three subsets S1, S2 and S3 (Table 1). S1 and S2 are evenly distributed for gender and race, and each has about 10,000 images while S3 has about 34000 images. Setting II is 2-fold cross-validation: 1) training on S1, testing on S2 + S3. 2) training on S2, testing on S1 + S3. Thus, there are about 10,000 training images and 40000 testing images. Setting III uses a subset followed the methods [44]–[46]. This setting selects 5,492 facial images of Caucasian descent to avoid ethnic variations and is randomly divided into two non-overlapped parts: 80% for training and 20% for testing. In order to reduce the randomness of the experimental results, we conduct five experiments and take the average value.

Webface dataset contains 62, 203 images. It involves large expressions and poses and is captured in the wild environment. The dataset with age ranging from 1 to 80 is very challenging since it contains incomplete images and unreal facial images. We conduct experiments on Webface with a 4-fold cross-validation following [47]. Segmentation steps can refer to Setting I in Morph Album II.

CACD is a large dataset containing around 160,000 images of 2,000 celebrities collected through internet. The age ranges from 14 to 62. It is divided into three subsets for training, validating and testing according to celebrities, which include 1800, 80, and 120 celebrities respectively. According to the segmentation settings of CACD in [31], Testing and validating sets are clean data with noisy images removed manually, and the experiment setting is only 1-fold.

IMDB-WIKI is the largest facial dataset with age labels, which consists of 523,051 images. Age range is from 0 to 100. It contains much more noise, so we utilize IMDB-WIKI only for pre-training.

B. DATA PREPROCESSING
In order to mitigate the influence of inconsistent size of different datasets and uneven data distribution, data preprocessing is introduced in the following.

1) ALIGNMENT
Face alignment is necessary for learning discriminative facial features. We firstly select five landmarks in face, i.e. left and right eye corners, tip of the nose and two mouth corners as key points by DCNN [48], and then apply an affine transformation for face registration. All the aligned images are of the size 256 $\times$ 256 and then fed into the network. Fig.3 shows the original images and aligned images. Since wrinkles are vital in face aging, unlike other work such as DRFs [7], we preserve the forehead in this alignment step.

2) PURIFICATION
This step is primarily applied for the Webface dataset, since it is collected from wild environments and full of complexity. The data purification process can be roughly shown in Fig.4. We utilize face++ to remove images including multiple faces in processing 1 and reject images of different modalities such as face sketches and non-facial images in processing 2.
3) HOMOGENIZATION

This step intends to keep a uniform data distribution for each fold in dataset segmentation. A data distribution map of three datasets is presented in Fig. 5, which shows the distribution of Morph and Webface is uneven. Most of the Morph data is concentrated on the age less than 40, and there is a significant difference between peaks and troughs in Webface data, while age distribution of CACD is relatively uniform. The uneven distribution of datasets illustrated in Fig. 5 at various ages underscores the importance of our data homogenization procedure. We first divide each dataset into subgroups according to age and keep the age distribution of each age group as uniform as possible, and then the data in each age group is segmented evenly and assigned to each fold.

C. AGE DATA GROUPING

In CR-MT net, the generalization performance of the regression task can be reinforced with a proper age data grouping. Too few groups cannot alleviate the bad effect of the data imbalance in the regression task, while too many groups can bring a drop of classification accuracy and are not helpful for learning the reliable classification information [37], [38], [49]. Therefore, a good data partition is essential for our CR-MT net. In order to obtain reliable and consistent results from different datasets, we select CACD, Webface Fold-1 and Morph Fold-1 in Setting I for analysis. Hereafter, the CR-MT net using adjacent age clustering and k-means clustering are denoted by CR-MTa and CR-MTk respectively.

1) ANALYSIS OF ADJACENT AGE CLUSTERING

Original data are grouped into the age interval of 5, 10, 15, 20, 25 and 30 respectively according to the range of age in different datasets. Table 2 shows the results with different age interval on three datasets. The first row represents the age interval, and the next rows show the classification and regression results of the three datasets. As shown in the Table 2 that the larger age interval, i.e. less groups, leads to higher accuracy of classification. But high accuracy of classification branch does not mean a small MAE in regression branch, it is because that the classification on the age groups with a large age interval would be easier than those with a small age interval. The results of all three datasets indicate that using 10 as the age interval is the optimal.

We have obtained the optimal group interval, but how to group the ages by the interval still needs to be considered. An arbitrary age grouping can bring a serious boundary effect. Given an initial age grouping by the optimal interval, i.e. 10-year interval, we slide the group intervals along the age range with a set of strides varying from –9 to 9 respectively, and perform train and test for the CR-MT net repeatedly with the new age groupings generated with the strides. This experiment is also performed on the three datasets. The results are shown in Table 3. From the table we can see clearly that the result reaches best consistently in all datasets when the sliding stride is 5. Therefore the ages can be grouped by sliding the initial grouping interval 10 with a stride 5.

2) ANALYSIS OF K-MEANS CLUSTERING

The k-means clustering is performed in the feature space described in Section III C, where each facial image is
represented as a 201-dimensional feature vector including 200-dimensional fully-connected features of the regression network and a 1-dimensional age label. Adjacent age clustering can be considered as a heuristic method for determination of the clustering number. According to the analysis of the adjacent age clustering, Clustering to 5 groups is the best choice. So we also set $k = 5$ for k-means clustering. Each clustering centroid is initialized by the mean of the image features in each group. Fig. 6 shows the relationship between k-means clustering results and age distribution. It can be seen from the distribution of curves that each group approximates to a normal distribution by age, and the ages in different groups overlap each other. Hence, the results of clustering reflect the cross-age correlations.

In order to validate our selection of the clustering number $k$, we vary the clustering number from 4 to 6. Table 4 shows the results obtained using different clustering number, and $k = 5$ is the optimal result.

**D. RESULTS AND ANALYSIS**

Table 5 shows experimental results of our CR-MT net using two different grouping methods and baseline derived on Morph and Webface dataset respectively, where the baseline represents a single-regression network. MAE of our CR-MTa and CR-MTk in Morph Setting I and Setting III stays stable among all five folds with an average below 2.40. The average MAE of our CR-MT net on Morph Setting II are 3.02 and 2.97, they are worse than the results of the Morph Setting I and Setting III. It is mainly because Setting II owns 4 times testing images than training. In addition, Setting II is mainly based on balance of gender and ethnicity, without taking the balance of ages into account. For Webface data, results in all four folds in our CR-MT net are not too uniform. The range of MAE in CR-MTk varies from 5.58 to 5.76 with an average of 5.67, which has a significant variation for age estimation. The same condition exists on CR-MTa with average 5.69. A possible reason is that there remains a degree of unevenness among different folds after data preprocessing.

CACD dataset has more images than the other two datasets. Most of the recent methods use deeper basic networks such as VGG and ResNet. Here, both VGG and Alexnet are adopted respectively as the basic network, because CACD is large enough to fit parameters in deeper network. When using VGG, we regard all convolution layers as shared layers. Also, we train on training set and manually cleaned validation set respectively following [22]. Table 6 exhibits the above experimental results. The optimal result is 4.48 when the basic architecture is VGG by CR-MTk. The accuracy of the network models trained with training subset outperforms than the accuracy of the network models trained with validation subset. The reason is that the size of the train subset is far larger than that of the validation subset. Moreover, the results using VGG as basic network outperform the results using...
Alexnet due to its far deeper architecture. For our CR-MT net, when the size of the training dataset is large enough, VGG is better than Alexnet as the basic network. Besides, it is obvious that the performance drops to a MAE of 6.34 when training on carefully annotated validation dataset. It shows that large and slightly rough datasets perform better than small and precise samples. In addition, deep networks are superior to shallow networks but with a higher space and time complexity, even with higher possibility of overfitting.

The results of both Table 5 and Table 6 show that our CR-MT net beats the baseline on three datasets. It indicates the MTL combining both classification and regression tasks is effective for age estimation. Also, K-means clustering proved to be slightly superior to adjacent age clustering for our CR-MT, the reason could be that K-means clustering concatenate both feature vectors of images and corresponding age labels. Results of Webface are far worse than those of the other datasets. The possible reason is that the Webface data is captured in the wild environment involving large poses and various diversity of expressions which indicates that age estimation from images in the natural environment is still challenging.

E. ABLATIVE ANALYSIS

In this section, we design a series of contrastive experiments to analyse the weight setting of two tasks and selection of the shared layers. Our ablative experiments are performed in CACD dataset with sliding stride setting to 0 in CR-MTa net.

1) ANALYSIS OF WEIGHT SETTING OF TWO TASKS

In CR-MT net, the two tasks play a different role, i.e. regression branch is the main task, and the classification branch is used to reinforce the regression branch to get more accurate age estimation results. Refer to [50], adjusting the weight to balance the importance of the two tasks and get more reliable results is necessary.

Following the method [50], we sample a set of values of \( \lambda \) and train the CR-MT net on CACD dataset with each \( \lambda \) respectively. The validation and test results are exhibited in Fig. 7, where horizontal axis denotes \( \lambda \), which represents the weight ratio of classification to regression, vertical axis denotes MAE. Validation and test MAE are drawn by blue and red lines respectively in the figure. \( \lambda = \infty \) means that we only use a single regression network for age estimation, while \( \lambda = 0 \) represents we only use classification net which does not make sense. From Fig. 7 we can see that multi-task network is superior to a single-task network for age estimation, and \( \lambda = 1.6 \) is an appropriate weighting which produces the optimal results.

2) ANALYSIS OF SHARED LAYER SETTING

Shared presentation layer in a multi-task network enables several related tasks share relevant information, and reduce computational complexity. But the choice of the shared layers has always been a tricky issue because the optimal shared layers are often various in different tasks. Fig.8 shows the experimental results on CACD dataset by setting the shared layers from the first to seventh layer based on Alexnet. The MAEs dramatically descend with shared layers varying from the second convolution layer to the fifth convolution layer in both validation and test net. After that, MAEs increase gradually until the last fully connected layer. This indicates that the shared low level representation is very important to age estimation. The MAE reaches the minimum with 5 shared convolution layers.

F. COMPARISON WITH THE STATE-OF-THE-ART METHODS

With the optimal age grouping in classification branch, weight setting of two tasks and shared layers of two tasks, we compare our CR-MT net with some representative age estimation methods of recent years, where NET-Hybrid [27] and AGEn [39] are deep MTL methods, the others are divide-and-conquer methods. The comparative results from three datasets are illustrated in Table7, Table 8, and Table 9 respectively.

For Morph Setting I, Setting II and Setting III in Table 7 and Table 8, our CR-MT net outperforms the previous state-of-the-art methods on all three protocols, with achieving the best MAE of 2.15, 2.61, 2.31 respectively. The mark * in the table represents pre-training in IMDB. The MTL based methods such as NET-Hybrid [27] and AGEn [39] are not optimal because the auxiliary tasks are not sufficiently relevant to age regression. On the other hand, the divide-and-conquer based
 strategies for age data grouping, i.e. adjacent ages clustering and K-means clustering and demonstrate that the K-means clustering is better. We also diagnose some hyper-parameters that affect the performance of the CR-MT net such as the weight setting of two tasks and the selection of shared layers to find the optimal hyper-parameters. Finally we verify our CR-MT net in three public datasets, and the results show that the proposed approach is fairly competitive compared with the state-of-the-art methods.

In the future, there are also some possible improvements in our work. For example, the basic network can be replaced by a deeper network like our experiments on CACD. Also, We should consider the balance of performance and complexity, because a deeper network can lead to a larger network architecture, which requires higher hardware configuration and larger training datasets. In addition, local features also can be considered to be used to construct multiple MTL nets.

### Table 7. The comparisons between the proposed method and other state-of-the-art methods on MORPH Setting I and Setting III dataset.

| Method       | MORPH Setting I | MORPH Setting III |
|--------------|-----------------|-------------------|
| Ranking-CNN  | 2.96            | -                 |
| AGEnet       | -               | 2.52*             |
| DEX          | -               | 2.68*             |
| CasCNN       | 3.30            | 3.16*             |
| SSR-Net      | -               | 2.17              |
| DFRs [7]     | 2.17            | 2.91              |
| MV [19]      | 2.16* / 2.41    | -                |
| PGR-NET [21] | 2.33            | 2.75*             |
| CCAE [16]    | -               | -                |
| Hierarchical [36] | -   | 3.7             |
| BridgeNet [20] | -         | 2.38*             |
| CR-MTa       | 2.16* / 2.32    | 2.32* / 2.36      |
| CR-MTk       | 2.15* / 2.30    | 2.31* / 2.34      |

### Table 8. The results on MORPH Setting II and CACD.

| Method          | MORPH Setting II | CACD          |
|-----------------|------------------|---------------|
| CasCNN [23]     | 2.93*            | 5.22          |
| DEX [22]        | -                | 4.785         |
| AGEnet [39]     | 2.70* / 2.86     | 4.68          |
| NET HYbrid [27] | 2.96             | -             |
| DFRs [7]        | 2.98             | 4.657         |
| BridgeNet [20]  | 2.63*            | -             |
| CR-MTa          | 2.63* / 3.02     | 4.52          |
| CR-MTk          | 2.61* / 2.97     | 4.48          |

### Table 9. The results on Webface. Our method achieves the state-of-the-art performance.

| Method          | Webface          |
|-----------------|------------------|
| Tree-a-CNN [34] | 7.72             |
| D2C [15]        | 6.04             |
| NET HYbrid [27] | 5.75             |
| PGR-NET [21]    | 5.98             |
| CR-MTa          | 5.69             |
| CR-MTk          | 5.67             |

V. CONCLUSION

Age estimation of facial images is very challenging and is still not solved well. This paper proposes an end-to-end multi-task learning network called CR-MT net for age estimation, which combines age classification and regression. In CR-MT net, classification is an auxiliary task to boost the generalization performance of the regression task, and a good age grouping for classification is also important for the regression task. In order to obtain a homogeneous data partition and alleviate the boundary effect in classification, we present two
[16] C. Zhang, S. Liu, X. Xu, and C. Zhu, “C3AE: Exploring the limits of compact model for age estimation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 12587–12596.

[17] D. Yi et al., “Age estimation by multi-scale convolutional network,” in Proc. Asian Conf. Comput. Vis., 2014, pp. 144–158.

[18] S.-B. Guo, H.-Y. Zhou, J. Wu, and X. Geng, “Age estimation using expectation of label distribution learning,” in Proc. IJICAR, 2018, pp. 712–718.

[19] H. Han, S. Han, and X. Chen, “Mean-variance loss for deep age estimation from a face,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 5285–5294.

[20] W. Li, J. Lu, J. Feng, C. Xu, J. Zhou, and Q. Tian, “BridgeNet: A continuity-aware probabilistic network for age estimation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 1145–1154.

[21] N. Liu, L. Chang, and F. Duan, “PGR-Net: A parallel network based on group and regression for age estimation,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2019, pp. 2377–2381.

[22] R. Rothe, R. Timofte, and L. Van Gool, “Deep expectation of real and apparent age from a single image without facial landmarks,” Int. J. Comput. Vis., vol. 126, nos. 2–4, pp. 144–157, Apr. 2018.

[23] J. Han, Z. Tan, Z. Lei, G. Guo, and S. Z. Li, “Auxiliary demographic information assisted age estimation with cascaded structure,” IEEE Trans. Cybern., vol. 48, no. 9, pp. 2531–2541, Sep. 2018.

[24] P. Li, Y. Hu, X. Wu, R. He, and Z. Sun, “Deep label refinement for age estimation,” Pattern Recognit., vol. 100, Apr. 2020, Art. no. 107178.

[25] M. Ruder, “An overview of multi-task learning in deep neural networks,” 2017, arXiv:1706.05098. [Online]. Available: http://arxiv.org/abs/1706.05098

[26] Z. Wang, W. Li, Y. Kao, D. Zou, Q. Wang, M. Ahn, and S. Hong, “HCR-Net: A hybrid of classification and regression network for age estimation,” in Proc. IJCAI, 2018, pp. 1014–1020.

[27] J. Xing, K. Li, W. Hu, C. Yuan, and H. Ling, “Diagnosing deep learning models for high accuracy age estimation from a single image,” Pattern Recognit., vol. 66, pp. 106–116, Jun. 2017.

[28] X. Liu, S. Li, M. Kan, J. Zhang, S. Wu, W. Liu, H. Han, S. Shan, and X. Chen, “AgeNet: Deeply learned regressor and classifier for robust apparent age estimation,” in Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW), Dec. 2015, pp. 258–266.

[29] K. Ristic and T. Tesafaye, “MORPH: A longitudinal image database of normal adult age-progression,” in Proc. 7th Int. Conf. Autom. Face Gesture Recognit. (FG’R), Apr. 2006, pp. 341–345.

[30] B. Ni, Z. Song, and S. Yan, “Web image mining towards universal age estimator,” in Proc. 17th ACM Int. Conf. Multimedia (MM), 2009, pp. 85–94.

[31] B.-C. Chen, C.-S. Chen, and W. H. Hsu, “Face recognition and retrieval based on an entropy cost function to retrieve documents represented in journal articles on related topics.”

[32] FAN ZHANG received the B.E. degree from the Beijing University of Posts and Telecommunications, China, in 2016. She is currently pursuing the Ph.D. degree with Beijing Normal University, China. Her research interests include image processing, face attribute estimation, and deep learning.

[33] FUQING DUAN (Member, IEEE) received the B.S. and M.S. degrees in mathematics from North-west University, Xi’an, China, in 1995 and 1998, respectively, and the Ph.D. degree in pattern recognition from the National Laboratory of Pattern Recognition, Beijing, China, in 2006. He is currently a Professor with the College of Artificial Intelligence, Beijing Normal University, Beijing. His current research interests include 3D face reconstruction, skull identification, and machine learning and applications. He has authored more than 80 conference and journal articles on related topics.

[34] NA LIU received the B.S. degree from Xidian University, Xi’an, China, in 2016. She is currently pursuing the Ph.D. degree with Beijing Normal University, China. Her research interests include image processing, and deep learning.