Learning Triangular Distribution in Visual World

Ping Chen†[1], Xingpeng Zhang†[4], Chengtao Zhou[1], Dichao Fan[1], Peng Tu[3], Le Zhang[1], Yanlin Qian[1,2*]
[1]MicroBT Inc. [2]Waseda University, IPS. [3]RuqiMobility Inc.
[4]School of Computer Science and Software Engineering, Southwest Petroleum University, Chengdu, China

Abstract

Convolution neural network is successful in pervasive vision tasks, including label distribution learning, which usually takes the form of learning an injection from the nonlinear visual features to the well-defined labels. However, how the discrepancy between features is mapped to the label discrepancy is ambient, and its correctness is not guaranteed.To address these problems, we study the mathematical connection between feature and its label, presenting a general and simple framework for label distribution learning. We propose a so-called Triangular Distribution Transform (TDT) to build an injective function between feature and label, guaranteeing that any symmetric feature discrepancy linearly reflects the difference between labels. The proposed TDT can be used as a plug-in in mainstream backbone networks to address different label distribution learning tasks. Experiments on Facial Age Recognition, Illumination Chromaticity Estimation, and Aesthetics assessment show that TDT achieves on-par or better results than the prior arts. Code is available at https://github.com/redcping/TDT.

1. Introduction

Label distribution learning (LDL) utilizes the advantages of distribution to solve the task of quasi-continuous label or inner-correlation between labels [22, 58]. LDL assigns a distribution over label value to an instance, which can be obtained by fitting a Gaussian or Triangle distribution whose peak indicates the label and represents the relative importance of each label describing an instance [58]. Hence, LDL has an impact on many real-world applications, such as facial age estimation [22], head-pose estimation [20, 23], crowd counting [76], zero-shot learning [35], facial beauty prediction [51], hierarchical classification [68], partial multi-label learning [69] and so on.

When LDL was first proposed, it mainly used maximum entropy and Kullback-Leibler divergence to learn label distribution [22]. Then, a more efficient optimization method BFGS is proposed to replace the IIS method [19]. The LDL is also widely combined with other algorithms, such as random forest [57], deep convolution neural network [71], hashing method [77], Bayesian [79], metric learning [80] and so on. Although LDL has wide applications, it encounters some challenges, i.e. the feature space and label (solution) space are inhomogeneous [57]. To be more specific, the feature space generated by the model is nonlinear, while the label information is gradually changing. In addition, the existing label distribution is mainly based on label information to construct the distribution, placing the label information for one instance over the whole label distribution.

Therefore, this article proposes a Triangular Distribution Transform (TDT) method, aiming to linearly map the transformed feature information to its corresponding label information concisely and efficiently. Inspired by label distribution learning [57], we propose using symmetric triangular distributions to represent this symmetric linear transformation. Specifically, the high-dimensional features obtained by feature extraction networks will reflect the relationship between labels through symmetric triangular distributions. To better achieve this goal, we draw inspiration from the paradigm of comparative learning and supply two sets of images to the feature extraction network each time. One set is the prior knowledge that needs to be compared, and the other set is the training sample. Our method is more suitable for visual tasks with linearly continuous changing label information, such as age, aesthetics, lighting intensity, etc. Our TDT can be used as a plug-in in mainstream backbone networks. Therefore, our method has achieved excellent results on multiple visual tasks, such as facial age estimation, image aesthetics estimation, and illumination estimation.

The contribution of this article is summarized as follows:

- We analyze to lay the theoretical foundation for the Triangular Distribution Transform, enabling feature discrepancy to explain label difference.
- We show with the proposed symmetry-related loss and commutativity-related loss, TDT can be learnt by mainstream backbone networks.
- TDT outperforms other methods on age estimation, aesthetics estimation, and illumination estimation.

*Corresponding author. †These authors contributed equally.

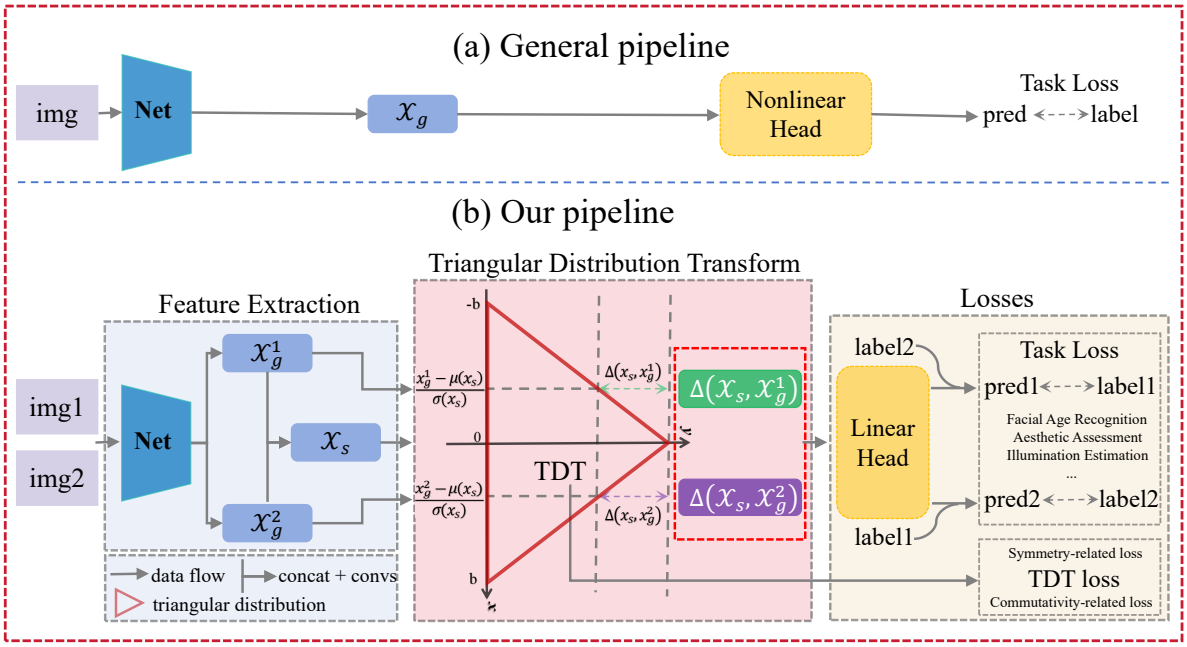

Figure 1. The overall structure of our TDT. Our pipeline (b) diverges from the general pipeline (a) by incorporating a parameter-free TDT (Triangular Distribution Transform), enabling converting the nonlinear feature to the one which vary "linearly" as per Eq.4. Consequently, a linear head module alone suffices to establish the mapping between image features and their respective labels, with clear explanation which is missing for a conventional network head. For detailed information on the TDT loss, please refer to Figure.2.

2. Related works

2.1. Label distribution learning

The label distribution method learns the relevance between labels to reflect the relative importance of different labels [21, 22, 29, 58], which can also be seen as a special facial age classification method. Label distribution learning methods learn a label distribution that represents the relative importance of each label when describing an instance[21]. The label distribution covers a certain number of labels. Each label has its description degree, representing the degree to which each label describes the instance[62]. The description degrees of all the labels sum up to 1 [62]. Due to the advantages of label distributed learning, it has achieved very excellent performance in tasks such as facial age estimation[22], head-pose estimation [20, 23], crowd counting [76], zero-shot learning[35], facial beauty prediction[51], hierarchical classification[68], and partial multi-label learning [69]. The label distribution learning method is very consistent with the potential law of big data. Nevertheless, acquiring distributional labels for thousands of face images itself is a non-trivial task[58].

2.2. Facial Age estimation

The regression methods [45], classification methods [53], and ranking methods [7, 9] for age estimation pay more attention to putting forward different research methods according to label information. The age regression methods consider labels as continuous numerical values. To handle the heterogeneous data, researchers also proposed hierarchical models [26] and the soft-margin mixture of regression [33]. And age classification regards labels as independent values[53]. It regards each age as a separate category and ignores the similarity of the same person between different ages. While the ranking approaches treat labels as rank-order data and use multiple binary classifiers to determine the age rank in a facial image[7, 9]. Besides, some scholars also focus on the objective optimization function [12, 46]. ML-loss [46] proposed mean-variance loss for robust age estimation via distribution learning. Deng et al. [12] proposed progressive margin loss (PML) for long-tailed age classification. These methods gradually consider that aging is a slow and continuous process, which also means that the processing of label information is significant.

2.3. Aesthetic Assessment

An aesthetic assessment task refers to evaluating the visual beauty and artistic value of given image. Early studies rely on handcrafted features [13, 43, 44], which ignore the spatial features and semantics in assessing aesthetics. In data-driven learning-based methods, NIMA[63] uses Earth Mover's Distance to optimize aesthetic distribution predic-

tion. A-lamp[42] and MP$_{ada}$[59] both achieve good results with a multi-patch approach. Hierarchical Layout-aware Graph Convolution Network (HGCN)[56] captures layout information. TANet [27] adaptively learns aesthetic prediction rules based on identified themes, using Mean Squared Error as the metric. Transformer [28] assign attention levels to color spaces, enabling segmentation learning. Yi *et al*. [72] effectively combine feature style and general aesthetic information using AdaIN [34] and self-supervised pre-training for accurate aesthetic assessment.

2.4. Illumination Estimation

Illumination Estimation is often dubbed as Auto White Balance verbally and aims at measuring the normalized illumination vector given at least one single image. There exists a bunch of traditional methods that are easy to implement but easily fail due to the over-optimistic assumption, for example, White Patch [4], General Gray World [1], Gray Edge [65], Shades-of-Gray [16], LSRS [17], PCA [10] and Grayness Index [49], *etc*. Relying on labeled data, a bunch of learning-based methods with tunable inner weight *e.g*. [2, 5, 6, 15, 18, 24, 25, 32, 36, 48, 61] leads the leaderboard by a large gap generally. All aforementioned methods output a single illumination vector in a deterministic way. A few works deal with label distribution learning; Egor *et al*. [14] proposed an efficient illumination distribution estimation method; FFCC [3] from Barron outputs a unique illumination vector which can be modified as a distribution.

3. Proposed Method

The pipeline of proposed Triangular distribution learning is illustrated in Fig.1. Fig.1(a) depicts the general pipeline for CNN tasks, where non-linear head modules are employed to map non-linear features to labels due to the linear independence of image features. These non-linear head modules can include modules that combine convolution with non-linear activation functions, Gaussian Mixture Model, and others. In Fig.1(b), we propose a novel pipeline that differs from the general pipeline. We introduce a parameter-free TDT (Triangular Distribution Transform) after the non-linear image features, allowing for the possibility of transforming linearly independent image features into linearly correlated ones. As a result, we only need to utilize a linear head module, such as a fully connected (FC) module, in combination with the relevant loss function to achieve linear correlation learning among the features. Additionally, we employ a contrastive learning-like approach to guide the predictions of unknown samples based on known samples.

3.1. Triangular Distribution Transform on Latent Feature

With some ambiguity, vision regression problems can be classified into linear-label problems and nonlinear-label ones. The mapping from the original information modality (for example 2D images, text, voice, video *etc.*) to the corresponding labels is usually non-linear and hard to model in a fixed rule. Take facial age recognition as an example, age label varies linearly while the visual feature of the facial image usually does not work in the same way. The majority of works on distribution learning mainly focus on learning a mapping from visual features to the label in a nonlinear way, such as the deep net, with its interpretability remaining in suspense. In this article, we discuss the relationship between the feature and its label and present a so-called Triangular Distribution Transform (TDT) to rigidly connect the feature and its label, so that the connection is injective.

Starting with an input $X \in \mathbb{R}^{N\times 3\times H\times W}$ (N is the batch size, $3, H, W$ denotes the channel, height, and width respectively) and it label Y, we extract feature by some net module g(parameterized by Θ) as follows:

$$\mathcal{X}_g = g(\mathcal{X}, \Theta) \tag{1}$$

where $\mathcal{X}_g \in \mathbb{R}^{N\times C\times h\times w}$ is a C-channel feature extracted from some backbone nets (*e.g.*, ResNet18). The mean and standard deviation are calculated along the h and w dimensions, referred to as $\mu(\mathcal{X}_g) \in \mathbb{R}^{N\times C\times 1\times 1}$ and $\sigma(\mathcal{X}_g) \in \mathbb{R}^{N\times C\times 1\times 1}$ respectively.

Based on the Central Limit Theorem, we assume $\mathcal{X}_g$ follows Gaussian distribution and denote the probability density function (PDF) as $\phi(\mathcal{X}_g)$. Given features for two samples ($\mathcal{X}_g^1$ and $\mathcal{X}_g^2$), we define their "feature difference" based on the Gaussian distribution of $\mathcal{X}_g^1$ as:

$$\begin{aligned}\Delta(\mathcal{X}_g^1, \mathcal{X}_g^2) &= \phi(\mathcal{X}_g^1) - \phi(\mathcal{X}_g^2) \\ &\propto \mathcal{N}_{0,1}(\frac{\mathcal{X}_g^1 - \mu(\mathcal{X}_g^1)}{\sigma(\mathcal{X}_g^1)}) - \mathcal{N}_{0,1}(\frac{\mathcal{X}_g^2 - \mu(\mathcal{X}_g^1)}{\sigma(\mathcal{X}_g^1)})\end{aligned} \tag{2}$$

The last part of Eq.2 is the normalized feature difference using the standard Gaussian distribution $\mathcal{N}_{0,1}$. We give our first assumption:

Assumption 1: There exists a function transforming $\mathcal{X}_g$ to $\mathcal{X}'_g$, whose change linearly *w.r.t.* the label.

To better fit Assumption 1, considering $\Delta(\mathcal{X}_g^1, \mathcal{X}_g^2)$ does not correlate linearly with the label, we approximate the Gaussian distribution in Eq.2 using a symmetric Triangular distribution $\tau(\cdot|b)$ following [55], formulated as:

$$\begin{aligned}\Delta(\mathcal{X}_g^1, \mathcal{X}_g^2) &\approx \Delta_\tau(\mathcal{X}_g^1, \mathcal{X}_g^2) \\ &= \tau(\frac{\mathcal{X}_g^1 - \mu(\mathcal{X}_g^1)}{\sigma(\mathcal{X}_g^1)}|b) - \tau(\frac{\mathcal{X}_g^2 - \mu(\mathcal{X}_g^1)}{\sigma(\mathcal{X}_g^1)}|b),\end{aligned} \tag{3}$$

$$\tau(s|b) = \begin{cases}\frac{1}{b}\cdot(1-\frac{|s|}{b}), & -b \le s \le b \\ 0, & otherwise\end{cases} \tag{4}$$

where the scalar b controls the approximation error and is set to $\sqrt{6}$ (b is obtained via the method of moments in [55]).

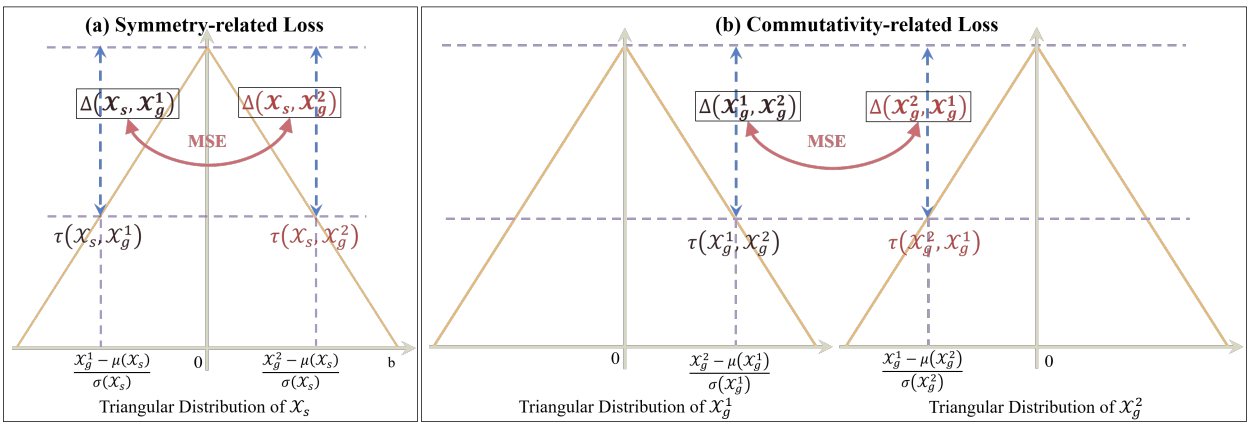

Figure 2. TDT is learned relying on the commutativity-related loss and symmetry-related loss, while the latter plays the primary role. The feature difference, associated with the symmetry-related loss, is used for result prediction. MSE is the mean square error.

In contrast to Eq.2, Eq.4 has the property of linearly describing the corresponding label difference. Then we can search a linear function $f(\Delta(\mathcal{X}_g^1, \mathcal{X}_g^2))$ that transforming the feature difference to the label difference $|Y^1 - Y^2|$, in a linear, symmetric and commutative way[1].

3.2. Optimization towards Triangular Distribution Transform and Vision Tasks

The learned feature does not follow Triangular distribution unless we optimize the net towards it. Then we state the optimization loss towards TDT and related task-specific loss. The optimization loss of TDT, shown in Fig.2, aims to gradually transform the non-linear correlation among image features into a linear correlation through label-guided learning.

Loss for Triangular Distribution

Given two sets of nonlinear features $\mathcal{X}_g^1$ and $\mathcal{X}_g^2$ extracted from backbone network, we use multiple convolution layers to learn to make a feature that follows symmetric Triangular distribution, formulated as $\mathcal{X}_s = Convs(Concat(\mathcal{X}_g^1, \mathcal{X}_g^2), \theta)$, where the concatenation happens in the batch dimension. We calculate the mean and standard deviation separately for $\mathcal{X}_s$, $\mathcal{X}_g^1$ and $\mathcal{X}_g^2$ along the spatial dimensions. Via Eq.4, we get $\Delta(\mathcal{X}_s, \mathcal{X}_g^1)$ and $\Delta(\mathcal{X}_s, \mathcal{X}_g^2)$.

To guide the net to reach a state that $\mathcal{X}_g^1$ and $\mathcal{X}_g^2$ are symmetrical around the "center" $\mathcal{X}_s$, a symmetry-related loss is introduced:

$$L_S = \|\Delta(\mathcal{X}_s, \mathcal{X}_g^1) - \Delta(\mathcal{X}_s, \mathcal{X}_g^2)\|_2, \tag{5}$$

where the operator $\|\cdot\|_2$ is a L2 loss. Analogously, the commutativity-related loss is designed as:

$$L_M = \|\Delta(\mathcal{X}_g^1, \mathcal{X}_g^2) - \Delta(\mathcal{X}_g^2, \mathcal{X}_g^1)\|_2, \tag{6}$$

Then, to facilitate the learning of the linear function $f(\Delta)$, we adopt a comparative approach and design a relevant supervised loss:

$$\begin{aligned} L_P = &\|\delta(Y) - 2 \cdot f(\Delta(\mathcal{X}_s, \mathcal{X}_g^1) \cdot sgn(\mathcal{X}_s, \mathcal{X}_g^1))\|_1 \\ &+ \|\delta(Y) + 2 \cdot f(\Delta(\mathcal{X}_s, \mathcal{X}_g^2) \cdot sgn(\mathcal{X}_s, \mathcal{X}_g^2))\|_1, \end{aligned} \tag{7}$$

Here, $\|\cdot\|_1$ refers to smooth L1 operator, $\delta(Y) = Y^2 - Y^1$, $sgn(\mathcal{X}_s, \mathcal{X}_g)$ represents the sign of $\Delta(\mathcal{X}_s, \mathcal{X}_g)$. When $\mathcal{X}_g$ in the distribution of $\mathcal{X}_s$ has a cumulative distribution function (CDF) less than 0.5, $sgn(\mathcal{X}_s, \mathcal{X}_g)$ is considered positive, and vice versa. The first part of Eq.7 represents using samples labeled as Y^2 to predict samples labeled as Y^1, and the other term follows the same principle.

Therefore, the final loss function for learning the triangular distribution can be formulated as follows:

$$L_T = L_S + L_M + L_P, \tag{8}$$

Loss for Vision Tasks

To quickly validate the effectiveness of TDT, we have selected three common visual tasks: facial age recognition, aesthetic assessment, and illumination estimation. Below are the descriptions and loss definitions for these three tasks.

Facial Age Recognition: Age variation can be considered linear, and we can utilize TDT to draw the relationship between visual feature differences among different images and the corresponding age variation. Therefore, we can directly employ Eq.8 for optimizing age estimation.

Aesthetic Assessment: The aesthetic assessment variation in scores can be considered linear, similar to the optimization for the age estimation task, we can also choose

[1]Linearity: given $\Delta_\tau(\mathcal{X}_g^1, \mathcal{X}_g^2) = \Delta_\tau(\mathcal{X}_g^2, \mathcal{X}_g^3)$, we have $Y^1 - Y^2 = Y^2 - Y^3$.
Symmetry: $\Delta_\tau(\mathcal{X}_g^1, \mathcal{X}_g^1 + \delta\mathcal{X}_g) = \Delta_\tau(\mathcal{X}_g^1, \mathcal{X}_g^1 - \delta\mathcal{X}_g)$.
Commutativity: $\Delta_\tau(\mathcal{X}_g^1, \mathcal{X}_g^2) = \Delta_\tau(\mathcal{X}_g^2, \mathcal{X}_g^1)$.

Table 1. Experiments setting for three tasks evaluated.

Task	Baseline	Dataset	Input size
Facial Age Recognition	DAA(Resnet18)[8]	MegaAge-Asian[75]	96×96
Aesthetic Assessment	SAAN(Resnet50+VGG19) [72]	TAD66K[27]	256×256
Illumination Estimation	FC^4(FC, SqueezeNet)[31]	Reprocessed Color Checker [18, 60]	512×512

Eq.8 as the overall optimization. In this case, the label corresponds to the aesthetic score.

Illumination Estimation: In illumination estimation, a key step of automatic white balance, we consider the variation of the illumination to be approximately linear. Additionally, to better capture the differences in illumination angles, we combine Eq.8 and the angle error [31] as the overall optimization function $L_{ill} = L_T + \frac{180}{\pi} \cdot arccos(p \cdot p^*)$, where the p and p^* represent the normalized estimation and ground truth of illumination color, respectively.

Algorithm 1 The training process of our method

Input: Sample set X and corresponding labels Y
Output: Loss
Prior Set Selection: Randomly select N samples X_2 from the training set with labels Y_2.
for each sample X_1 in forward pass **do**
  # Feature fusion and distribution generation
  $\mathcal{X}_g^1, \mathcal{X}_g^2 \leftarrow g(X_1, \Theta), g(X_2, \Theta)$ # Extract features
  $N, C, h, w \leftarrow \mathcal{X}_g^2$.shape # C is distribution number
  $\mathcal{X}_g^1 \leftarrow \mathcal{X}_g^1.repeat(N, 1, 1, 1)$
  $\mathcal{X}_s \leftarrow conv(concat(\mathcal{X}_g^1, \mathcal{X}_g^2, 1), C)$
  # Perform TDT operation on $\mathcal{X}_g^1$ based on $\mathcal{X}_s$
  $\mu(\mathcal{X}_s), \sigma(\mathcal{X}_s) \leftarrow calc\_mean\_std(\mathcal{X}_s)$
  $\Delta(\mathcal{X}_s, \mathcal{X}_g^1) \leftarrow \tau(\frac{\mathcal{X}_s - \mu(\mathcal{X}_s)}{\sigma(\mathcal{X}_s)}|b) - \tau(\frac{\mathcal{X}_g^1 - \mu(\mathcal{X}_s)}{\sigma(\mathcal{X}_s)}|b)$
  # Feature-to-Label Difference Mapping
  $sgn(\mathcal{X}_s, \mathcal{X}_g^1) \leftarrow (-1)^{cdf(\mathcal{X}_g^1) > 0.5}$
  $\delta Y_1 \leftarrow f(\Delta(\mathcal{X}_s, \mathcal{X}_g^1) \cdot sgn(\mathcal{X}_s, \mathcal{X}_g^1))$ # via Eq.7
    where $f(\cdot) = fc(gap(\cdot))$
  Optimize TDT, including L_S, L_M, and L_P
end for

3.3. TDT Algorithm Process

To be more clear, we provide pseudocode 1, in 5 parts.

Prior Samples Selection: Prior samples are several selected training samples, providing comparison with those samples used for training/testing. Features for the prior samples can be pre-computed and integrated into the model.

Distribution generation: We obtain $\mathcal{X}_s$ and form its high-dimensional Normal Distribution using $\mu(\mathcal{X}_s)$ and $\sigma(\mathcal{X}_s)$. After standardizing and approximating, we get a zero-symmetric high-dimensional Triangular Distribution to linearly represent feature differences.

TDT Operation: It calculates the PDF differences between $\mathcal{X}_s$ and $\mathcal{X}_g$ based on the distribution of $\mathcal{X}_s$ meet $|\Delta(\mathcal{X}_g^1, \mathcal{X}_g^2)| = 2 \cdot |\Delta(\mathcal{X}_s, \mathcal{X}_g^1)| = 2 \cdot |\Delta(\mathcal{X}_s, \mathcal{X}_g^2)|$. Remarkably, TDT is entirely formulaic, parameter-free, and fit any net outputting feature maps.

Difference Mapping: Base on distribution of $\mathcal{X}_s$, $\mathcal{X}_s$ generally has a higher PDF than $\mathcal{X}_g$, meaning $\Delta(\mathcal{X}_s, \mathcal{X}_g) = pdf(\mathcal{X}_s) - pdf(\mathcal{X}_g) \geq 0$. We infer feature difference signs from $cdf(\mathcal{X}_g) > 0.5$ and linearly map them to label difference via global average pooling and a linear layer.

Optimization: For commutativity loss, we utilize the mean and std of $\mathcal{X}_g^1$ and $\mathcal{X}_g^2$. For symmetry and supervisory losses, those of $\mathcal{X}_s$ are computed.

4. Experiments

4.1. Notes and Implementation Details

We perform TDT validation on the baseline in three tasks, whose configuration details are shown in Table 1.

Train and Test notes: As observed in pseudocode 1, our TDT aims to guide the learning of unknown samples using prior samples, resembling contrastive learning. Therefore, for each task experiment, we randomly select a prior set from the training dataset. During training and testing, each batch comprises both the prior set samples used to obtain $\mathcal{X}_g^2$ and the unknown samples used to obtain $\mathcal{X}_g^1$. Each unknown sample undergoes TDT operations with all prior samples to make predictions, and the average of these predictions is considered as the final prediction. To reduce feature extraction time during testing, we encapsulate all the features from the prior set into the model parameters, resulting in an optimized model with accelerated inference.

Common Setting: For all experiments, we used the Adam optimizer, where the weight decay and the momentum were set to 0.0005 and 0.9, respectively. The initial learning rate was set to 0.001 and changed according to cosine learning rate decay. We trained our model using PyTorch on a cluster of 8 RTX 3090 GPUs. In an online manner, we augment all images with random horizontal flipping, scaling, rotation, and translation.

Facial Age Recognition: We utilize the *FG-Net* [47] dataset, which comprises 1002 facial images from 82 subjects, spanning an age range from 0 to 69. We follow the setup described in the papers [8, 12, 40], employing leave-one-person-out (LOPO) cross-validation. We report the average performance over 82 splits using the Mean Absolute Error (MAE) as the evaluation metric.

Additionally, we also use the *MegaAge-Asian* [75]

Table 2. CA on MegaAge-Asian.

Methods	Pre-trained	CA(3)	CA(5)	CA(7)
Posterior [75]	IMDB-WIKI	62.08	80.43	90.42
MobileNet [54]	IMDB-WIKI	44.0	60.6	-
DenseNet [70]	IMDB-WIKI	51.7	69.4	-
SSR-Net [70]	IMDB-WIKI	54.9	74.1	-
UVA [38]	-	60.47	79.95	90.44
LRN(ResNet10) [39]	IMDB-WIKI	62.86	81.47	91.34
LRN(ResNet18) [39]	IMDB-WIKI	64.45	82.95	91.98
VGG16(norm) [78]	ImageNet, IMDB-WIKI, AFAD	65.58	83.01	89.17
PVP+VGG16 [78]	ImageNet, IMDB-WIKI, AFAD	**72.65**	**87.24**	93.16
DAA(single channel) [8]	-	67.97	84.06	92.40
DAA(multi-channel) [8]	-	68.29	84.84	92.47
DAA [8]	-	68.82	84.89	92.70
TDT (ours)	-	69.60	85.42	**93.26**

Table 3. MAEs on FG-Net dataset.

Methods	MAE	Year
DEX[52]	3.09	2015
MV[46]	2.68	2018
C3AE[74]	2.95	2019
DRFs[58]	3.85	2021
PML[12]	2.16	2021
DAA [8]	2.19	2023
TDT(ours)	**2.12**	-

Dataset, consisting of 40,000 age-labeled samples spanning 0 to 70 years and 3,945 images for testing. For this dataset, we choose cumulative accuracy (CA) [75] as the evaluation metric, defined as $CA(n) = \frac{K_n}{K} \times 100$, where K is total number of testing images and K_n is the number whose absolute errors are smaller than n.

We adopt the official implementation of DAA[8] as the baseline for this task while replacing its DAA mapping with our TDT. The output size of $\mathcal{X}_g$ is set to 3×3.

In this task, 256 images are randomly picked from the training set serves as the prior set for all experiments.

Aesthetic Assessment: We use *TAD66K (Theme and Aesthetics Dataset with 66K images)* Dataset [27], a large-scale aesthetic quality assessment database with 47 themes. Images belonging to each theme are annotated independently, with each image containing a minimum of 1200 valid annotations. The aesthetic score of each image is treated as the label. Following TANet[27], the evaluation metric is the Mean Squared Error (MSE). We adopt the same train-test split setting. For the prior set, we select samples by choosing 2 random samples for each score ranging from 0 to 10 with interval 0.1 for each theme. We employ the official implementation of SAAN[72] as the baseline for this task while removing the final average pooling and BN layers. Additionally, we incorporate extra convolutions to obtain $\mathcal{X}_g$ with size of 8×8.

Illumination Estimation: We use *reprocessed Color Checker* [18, 60] dataset, one of the most widely adopted datasets in illumination estimation. Following FC^4 [31], the evaluation metric is the Recovery Angular Error. To facilitate better comparisons, we use the linear fully connected version of FC^4-net as the baseline, instead of the non-linear weight pooling version. With additional convolutional layers, we obtain $\mathcal{X}_g \in \mathbb{R}^{N \times C \times 8 \times 8}$. When comparing with the FC^4 model, we randomly select 128 samples from the training set as the prior set. However, when comparing with the method proposed by Tang *et al*. [64] that utilizes the extended external sRGB datasets, we choose same images for scene classification as the prior set, but only 1280 samples.

Following previous AWB works[18, 31, 60], 3-fold cross-validation is used. Standard metrics (mean, median, tri-mean of all the errors, the mean of the lowest 25%, and the mean of the highest 25% of errors, the 95th percentile error) are reported in terms of angular error in degrees.

Table 4. Results on Aesthetic Benchmark TAD66K.

method	Pub.	Basic	MSE
RAPID [41]	ACMMM2014	incorporate heterogeneous	0.0200
PAM[50]	ICCV2017	residual-based, active learning	0.0200
ALamp[42]	CVPR2017	layout-aware, multi-patch	0.0190
NIMA [63]	TIP2018	predict distribution	0.0210
MPada [59]	ACMM2018	attention, multi-patch	0.0220
MLSP [30]	CVPR2019	staged training,multi-level features	0.0190
UIAA [73]	TIP2019	unified probabilistic formulation	0.0210
HGCN [56]	CVPR2021	graph convolution networks	0.0200
TANet [27]	IJCAI2022	attention, adaptive features	**0.0161**
SAAN [72]	CVPR2023	AdaIN,self-supervised Pretraining	0.0185
TDT(ours)	-	Triangular Distribution	0.0172

4.2. Results and Analysis

Facial Age Estimation

The quantitative comparison of a list of top-performing methods on the MegaAge-Asian Dataset is shown in Table 2. From the table, we can find that even stocked with pre-training on large-scale datasets, methods like [39, 78] are inferior to ours. Compared with SSR-Net, our improvement is over 8%. PVP [78] gets higher CA(3) and CA(5) indexes, due to the massive pretraining on ImageNet[11], IMDB-WIKI[53], and AFAD[45]. DAA[8], published in 2023, shares the same setting with our proposed method, while TDT obtains better scores in all CA indexes.

To better prove the effectiveness of TDT, we also profile TDT on FG-Net [47], reported in Table 3. Using MAE as the metric, TDT leads the board with the smallest MAE, surpassing the same-setting competitor DAA.

Aesthetic Assessment.

In Table 4, we list the MSE scores for the selected and close-related aesthetic works. In an overall manner, TDT ranks second with an MSE score of 0.0172. This improvement simply originates from the deployment of the TDT plugin over SAAN [72], whose MSE is 0.0185, again showing the advantage of the label discriminative ability of TDT.

Illumination Estimation.

In Table 5, we report the statistics of the predicted angular error for a set of AWB algorithms on the Color Checker Dataset. On metrics like mean, and median values, the TDT based on FC^4 outperforms other state-of-the-art methods,

Table 5. Results on AWB Dataset Color Checker Dataset.

models	mean	med.	tri.	best25%	worst 25%	95th pct.
FFCC [3]	1.80	0.95	1.18	0.27	4.65	-
FC^4(Weighted) [31]	1.65	1.18	1.27	0.38	3.78	4.73
FC^4[31]	1.84	1.27	1.39	0.46	4.20	5.46
AlexNet FC4 [31]	1.77	1.11	1.29	0.34	3.78	4.29
Multi-Hypothesis [37]	2.10	1.32	1.53	0.36	5.10	-
IGTN [67]	1.58	0.92	-	0.28	3.70	-
MDLCC [66]	1.58	0.95	1.11	0.37	3.77	-
TLCC+sRGB [64]	1.51	0.98	1.07	0.33	**3.52**	-
TDT+FC^4	1.64	1.12	1.25	0.41	3.80	**4.53**
TDT+FC^4+sRGB	**1.46**	**0.85**	**1.05**	**0.26**	3.61	4.61

Table 6. Loss ablation analysis on MegaAge-Asian. ✓ and – indicate whether to add this loss to the final loss.

L_S	L_M	L_P	CA(3)	CA(5)	CA(7)
-	✓	✓	67.74	84.33	92.51
✓	-	✓	68.43	84.62	92.60
✓	✓	✓	**69.60**	**85.42**	**93.26**

such as MDLCC [66] and TLCC [64]. On top of FC^4, TDT also boosts the performance by a noticeable gap.

4.3. Ablation Study

We conduct an ablation study on the loss choice, the number of distributions and prior samples.

Loss function

It is vital to study the impact of different losses in Eq.8 and the results are presented in Table 6. We observe that removing any loss incurs a decrement in the overall performance. And symmetry loss brings the most improvement.

We design a toy experiment to further analyze the loss function Eq.7. The age loss L_P is a supervisory loss due to the label is needed for computing L_P, while the symmetry-related loss L_S and the commutativity-related loss L_M are unsupervisory losses. Switching off all nonsupervisory loss, as seen in the top two rows of Table 7, we see a clear CA score drop, indicating the advantage brought by the nonsupervisory loss.

If we cut off the access to unsupervisory loss for part1 (in other words, for only 75% data, the unsupersiory loss is calculated), as reported in the bottom half of Table 7, we claim: with increasing unlabeled data, the performance is massively improved, which again verifies the functionality of the symmetry-related and commutativity-related loss.

Distribution and Prior Number

Besides the loss selection, the number of distributions is a vital factor. On the same MegaAge-Asian Dataset, we test the number of distributions from the set $32, 64, 128, 256$. Table 8 shows that in general, the number of distributions has a minor effect on the final results and when feature is set 128-dimensional the performance is optimal.

The selected prior samples work like expert voting, hence we exploit the factor of the prior sample number. Table 8 proves that as the number of samples increases, the credibility of the results improves, yet the enhancement in credibility gradually diminishes.

Table 7. Loss analysis on MegaAge-Asian. The whole dataset is divided into part1 and part2, with a ratio of 25% and 75%.

L_S, L_M	L_P	label accessed to	CA(3)	CA(5)
No	all	all	68.26	84.26
Yes	all	all	**69.60**	**85.42**
part2	part2	part2	60.23	78.54
all	part2	part2	**66.59**	**83.62**

Table 8. Experiments on feature distributions and prior samples.

types	numbers	CA(3)	CA(5)	CA(7)
feature distributions	256	68.77	84.82	92.70
	128	**69.60**	**85.42**	**93.26**
	64	68.39	84.59	92.85
	32	67.80	85.04	92.90
prior samples	256	**69.60**	**85.42**	**93.26**
	128	68.73	84.94	92.80
	64	67.34	83.42	91.38

4.4. Visualization of what TDT learns

Fig.3 shows that: once the training is completed, we see feature difference approximates a triangular distribution function on varying symmetric age. The same property can be found for the predicted delta age (Inset (b)) in Fig.3). This results from TDT learning. From the inset (c) in Fig.3, we observe that the symmetry loss range is almost the same, meeting the feature symmetry assumption.

Fig.4 shows some qualitative comparison on illumination estimation task. Starting from raw image, different prior sample gives a close-to-groundtruth predicted white point, averaged to form the final robust white point. For the 3rd row, the prediction slightly shifts away from the groundtruth, which we guess it is due to multi-illumination.

As shown in Fig.5, we study the performance of TDT learning the symmetry axis location. For any pair of facial images, for example an age-a image from test set and an age-b image from the prior set, thus the symmetry axis should locate at $(a+b)/2$. With $(a+b)/2$ fixed, we alter the testing image and prior image, and draw a histogram of the resulted $\mathcal{X}_s$, which in fact gather around the symmetry axis. For pairs of $\{$age-3, age-34$\}$, the mean of predicted symmetric age is 18.27, close to the symmetric age 18.50.

5. Conclusion

A learning framework based on the Triangular Distribution Transform is proposed in this paper. It connects the nonlinear feature difference and the corresponding label difference in a verified linear, symmetric, and commutativity manner. This transform can be used as a portable plug-in for vision regression tasks, *e.g.*, we verify its application on three vision tasks. On facial age estimation, aesthetic as-

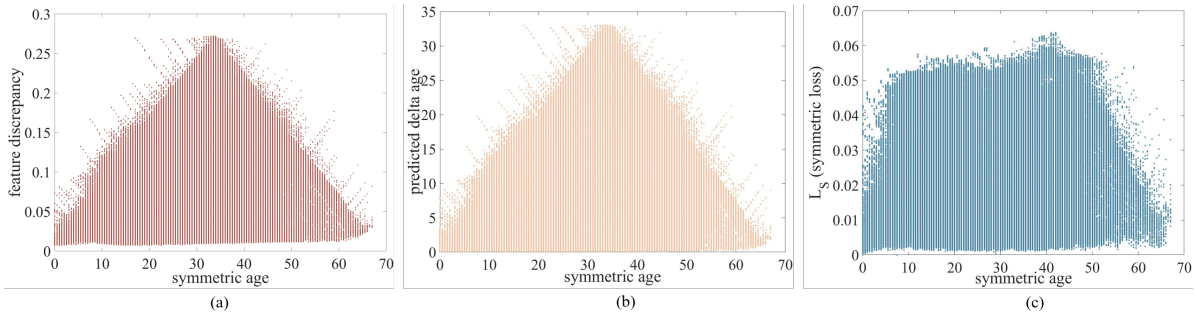

Figure 3. Relationship between symmetric age with the feature discrepancy, predicted delta age, and symmetric loss. (a) feature discrepancy; (b) predicted delta age; (c) symmetric loss

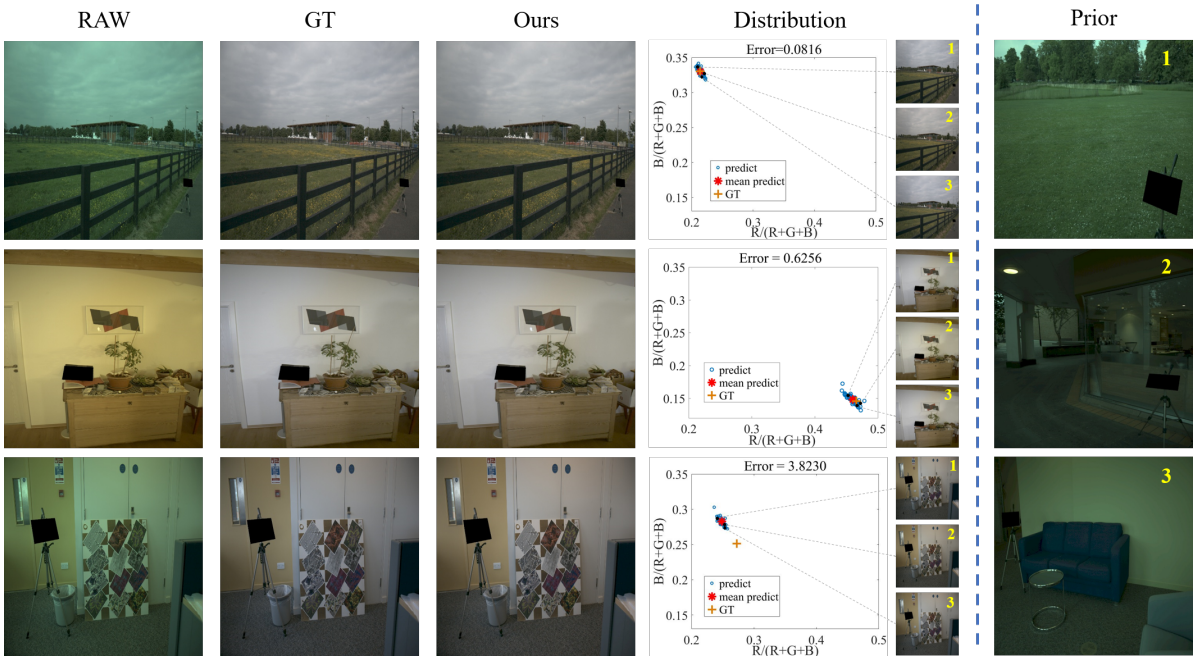

Figure 4. Qualitative comparison of the final color correction result and those from individual prior samples. In the fourth column, the predictions from different priors (their color corrected images are given in right insets) are shown clustered around the ground truth location, with limited variance. In the rightmost column, prior samples are given.

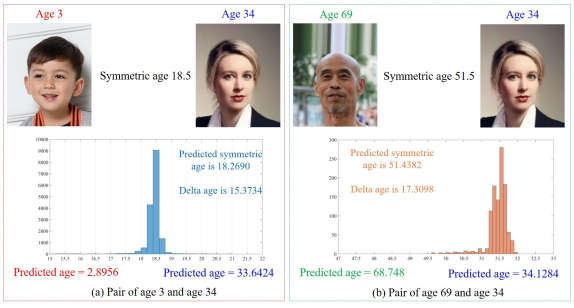

Figure 5. Study on the symmetry learnt in $\mathcal{X}_s$. For any pair of age-a image from test set and age-b image from the prior set, we draw a histogram of the resulted $\mathcal{X}_s$, which in fact gather around the symmetry axis $(a+b)/2$, w.r.t. the histogram figure.

sessment, and illumination estimation, the proposed TDT obtains on-par or even better performance than the prior arts, without much modification on the affiliated backbone. In the future, we will explore the application of TDT in a wider context, for example shape and pose estimation.

6. Acknowledgements

This work is supported by the Natural Science Starting Project of SWPU (No.2022QHZ023), the Sichuan Scientific Innovation Fund (No.2022JDRC0009), the Sichuan Provincial Department of Science and Technology Project (No.2022NSFSC0283) and the Key Research and Development Project of Sichuan Provincial Department of Science and Technology (No.2023YFG0129).

References

[1] Kobus Barnard, Vlad Cardei, and Brian Funt. A comparison of computational color constancy algorithms. i: Methodology and experiments with synthesized data. *IEEE TIP*, 11 (9):972–984, 2002. 3

[2] Jonathan T Barron. Convolutional color constancy. In *ICCV*, pages 379–387, 2015. 3

[3] Jonathan T. Barron and Yun-Ta Tsai. Fast fourier color constancy. In *CVPR*, pages 6950–6958, 2017. 3, 7

[4] David H Brainard and Brian A Wandell. Analysis of the retinex theory of color vision. *JOSA A*, 3(10), 1986. 3

[5] Ayan Chakrabarti. Color constancy by learning to predict chromaticity from luminance. In *NeurIPS*, pages 163–171, 2015. 3

[6] Ayan Chakrabarti, Keigo Hirakawa, and Todd Zickler. Color constancy with spatio-spectral statistics. *IEEE TPAMI*, 34 (8):1509–1519, 2012. 3

[7] Kuang-Yu Chang and Chu-Song Chen. A learning framework for age rank estimation based on face images with scattering transform. *IEEE TIP*, 24(3):785–798, 2015. 2

[8] Ping Chen, Xingpeng Zhang, Ye Li, Ju Tao, Bin Xiao, Bing Wang, and Zongjie Jiang. DAA: A delta age adain operation for age estimation via binary code transformer. In *CVPR*, pages 15836–15845, 2023. 5, 6

[9] Shixing Chen, Caojin Zhang, Ming Dong, Jialiang Le, and Mike Rao. Using ranking-cnn for age estimation. In *CVPR*, pages 742–751, 2017. 2

[10] Dongliang Cheng, Dilip K Prasad, and Michael S Brown. Illuminant estimation for color constancy: Why spatial-domain methods work and the role of the color distribution. *JOSA A*, 31(5):1049–1058, 2014. 3

[11] Jia Deng, Wei Dong, Richard Socher, LiJia Li, Kai Li, and FeiFei Li. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 6

[12] Zongyong Deng, Hao Liu, Yaoxing Wang, Chenyang Wang, Zekuan Yu, and Xuehong Sun. Pml: Progressive margin loss for long-tailed age classification. In *CVPR*, pages 10503–10512, 2021. 2, 5, 6

[13] Sagnik Dhar, Vicente Ordonez, and Tamara L Berg. High level describable attributes for predicting aesthetics and interestingness. In *CVPR*, pages 1657–1664, 2011. 2

[14] Egor Ershov, Vasily Tesalin, Ivan Ermakov, and Michael S Brown. Physically-plausible illumination distribution estimation. In *ICCV*, pages 12928–12936, 2023. 3

[15] Graham D Finlayson. Corrected-moment illuminant estimation. In *ICCV*, pages 1904–1911, 2013. 3

[16] Graham D Finlayson and Elisabetta Trezzi. Shades of gray and colour constancy. In *Color Imaging Conference (CIC)*, pages 37–41, 2004. 3

[17] S Gao, W Han, K Yang, C Li, and Y Li. Efficient color constancy with local surface reflectance statistics. In *ECCV*, pages 158–173, 2014. 3

[18] Peter V. Gehler, Carsten Rother, Andrew Blake, Thomas P. Minka, and Toby Sharp. Bayesian color constancy revisited. In *CVPR*, 2008. 3, 5, 6

[19] Xin Geng and Rongzi Ji. Label distribution learning. In *IEEE International Conference on Data Mining Workshops (ICDM)*, pages 377–383, 2013. 1

[20] Xin Geng and Yu Xia. Head pose estimation based on multivariate label distribution. In *CVPR*, pages 1837–1842, 2014. 1, 2

[21] Xin Geng, Kate Smith-Miles, and Zhi-Hua Zhou. Facial age estimation by learning from label distributions. In *AAAI*, pages 451–456, 2010. 2

[22] Xin Geng, Chao Yin, and Zhi-Hua Zhou. Deep differentiable random forests for age estimation. *IEEE TPAMI*, 35(10): 2401–2412, 2021. 1, 2

[23] Xin Geng, Xin Qian, Zeng-Wei Huo, and Yu Zhang. Head pose estimation based on multivariate label distribution. *IEEE TPAMI*, 44(4):1974–1991, 2022. 1, 2

[24] Arjan Gijsenij and Theo Gevers. Color constancy using natural image statistics and scene semantics. *IEEE TPAMI*, 33 (4), 2011. 3

[25] Arjan Gijsenij, Theo Gevers, and Joost van de Weijer. Generalized gamut mapping using image derivative structures for color constancy. *IJCV*, 86(2-3):127–139, 2010. 3

[26] Hu Han, Charles Otto, Xiaoming Liu, and Anil K. Jain. Demographic estimation from face images: Human vs. machine performance. *IEEE TPAMI*, 37(6):1148–1161, 2015. 2

[27] Shuai He, Yongchang Zhang, Rui Xie, Dongxiang Jiang, and Anlong Ming. Rethinking image aesthetics assessment: Models, datasets and benchmarks. In *IJCAI*, pages 942–948, 2022. 3, 5, 6

[28] Shuai He, Anlong Ming, Yaqi Li, Jinyuan Sun, ShunTian Zheng, and Huadong Ma. Thinking image color aesthetics assessment: Models, datasets and benchmarks. In *CVPR*, pages 21838–21847, 2023. 3

[29] Zhouzhou He, Xi Li, Zhongfei Zhang, Fei Wu, Xin Geng, Yaqing Zhang, Ming-Hsuan Yang, and Yueting Zhuang. Data-dependent label distribution learning for age estimation. *IEEE TIP*, 26(8):3846–3858, 2017. 2

[30] Vlad Hosu, Bastian Goldlucke, and Dietmar Saupe. Effective aesthetics prediction with multi-level spatially pooled features. In *CVPR*, pages 9375–9383, 2019. 6

[31] Yuanming Hu, Baoyuan Wang, and Stephen Lin. Fc^4 : Fully convolutional color constancy with confidence-weighted pooling. In *CVPR*, pages 330–339, 2017. 5, 6, 7

[32] YuanMing Hu, Baoyuan Wang, and Stephen Lin. FC4: Fully convolutional color constancy with confidence-weighted pooling. In *CVPR*, pages 330–339, 2017. 3

[33] Dong Huang, Longfei Han, and Fernando De la Torre. Soft-margin mixture of regressions. In *CVPR*, pages 4058–4066, 2017. 2

[34] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, pages 1501–1510, 2017. 3

[35] Zeng-Wei Huo and Xin Geng. Ordinal zero-shot learning. In *IJCAI*, pages 1916–1922, 2017. 1, 2

[36] Hamid Reza Vaezi Joze and Mark S Drew. Exemplar-based color constancy and multiple illumination. *IEEE TPAMI*, 36 (5):860–873, 2014. 3

[37] Daniel Hernández Juárez, Sarah Parisot, Benjamin Busam, Ales Leonardis, Gregory G. Slabaugh, and Steven McDonagh. A multi-hypothesis approach to color constancy. In *CVPR*, pages 2267–2277, 2020. 7

[38] Peipei Li, Huaibo Huang, Yibo Hu, Xiang Wu, Ran He, and Zhenan Sun. Uva: A universal variational framework for continuous age analysis. *arXiv preprint arXiv:1904.00158*, 2019. 6

[39] Peipei Li, Yibo Hu, Xiang Wu, Ran He, and Zhenan Sun. Deep label refinement for age estimation. *PR*, 100:107178, 2020. 6

[40] Wanhua Li, Jiwen Lu, Jianjiang Feng, Chunjing Xu, Jie Zhou, and Qi Tian. Bridgenet: A continuity-aware probabilistic network for age estimation. In *CVPR*, pages 1145–1154, 2019. 5

[41] Xin Lu, Zhe Lin, Hailin Jin, Jianchao Yang, and James Z Wang. Rapid: Rating pictorial aesthetics using deep learning. In *ACM MM*, pages 457–466, 2014. 6

[42] Shuang Ma, Jing Liu, and Chang Wen Chen. A-lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment. In *CVPR*, pages 4535–4544, 2017. 3, 6

[43] Luca Marchesotti, Florent Perronnin, Diane Larlus, and Gabriela Csurka. Assessing the aesthetic quality of photographs using generic image descriptors. In *ICCV*, pages 1784–1791, 2011. 2

[44] Masashi Nishiyama, Takahiro Okabe, Imari Sato, and Yoichi Sato. Aesthetic quality classification of photographs based on color harmony. In *CVPR*, pages 33–40, 2011. 2

[45] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Ordinal regression with multiple output cnn for age estimation. In *CVPR*, pages 4920–4928, 2016. 2, 6

[46] Hongyu Pan, Hu Han, Shiguang Shan, and Xilin Chen. Mean-variance loss for deep age estimation from a face. In *CVPR*, pages 5285–5294, 2018. 2, 6

[47] Gabriel Panis, Andreas Lanitis, Nicolas Tsapatsoulis, and Timothy F. Cootes. Overview of research on facial ageing using the fg-net ageing database. *IET Biometrics*, 5(2):37–46, 2016. 5, 6

[48] Yanlin Qian, Ke Chen, Joni Kämäräinen, Jarno Nikkanen, and Jiri Matas. Recurrent color constancy. In *ICCV*, pages 5459–5467, 2017. 3

[49] Yanlin Qian, Joni-Kristian Kamarainen, Jarno Nikkanen, and Jiri Matas. On finding gray pixels. In *CVPR*, pages 8062–8070, 2019. 3

[50] Jian Ren, Xiaohui Shen, Zhe Lin, Radomir Mech, and David J Foran. Personalized image aesthetics. In *ICCV*, pages 638–647, 2017. 6

[51] Yi Ren and Xin Geng. Sense beauty by label distribution learning. In *IJCAI*, pages 2648–2654, 2017. 1, 2

[52] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Dex: Deep expectation of apparent age from a single image. In *ICCV*, pages 10–15, 2015. 6

[53] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *IJCV*, 126(2):144–157, 2018. 2, 6

[54] Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, pages 4510–4520, 2018. 6

[55] William T. Scherer, Thomas A. Pomroy, and Douglas N.Fuller. The triangular density to approximate the normal density: decision rules-of-thumb. *Reliability Engineering & System Safety*, 82(3):331–341, 2003. 3

[56] Dongyu She, Yu-Kun Lai, Gaoxiong Yi, and Kun Xu. Hierarchical layout-aware graph convolutional network for unified aesthetics assessment. In *CVPR*, pages 8475–8484, 2021. 3, 6

[57] Liyue Shen, John Pauly, and Lei Xing. Nerp: implicit neural representation learning with prior embedding for sparsely sampled image reconstruction. *TNNLS*, 2022. 1

[58] Wei Shen, Yilu Guo, Yan Wang, Kai Zhao, Bo Wang, and Alan L. Yuille. Deep differentiable random forests for age estimation. *IEEE TPAMI*, 43(2):404–419, 2021. 1, 2, 6

[59] Kekai Sheng, Weiming Dong, Chongyang Ma, Xing Mei, Feiyue Huang, and Bao-Gang Hu. Attention-based multi-patch aggregation for image aesthetic assessment. In *ACM MM*, pages 879–886, 2018. 3, 6

[60] Lilong Shi and Brian Funt. Re-processed version of the gehler color constancy dataset of 568 images. *accessed from* `http://www.cs.sfu.ca/~colour/data/`, 2010. 5, 6

[61] Wu Shi, Chen Change Loy, and Xiaoou Tang. Deep specialized network for illumination estimation. In *ECCV*, page 371–387, 2016. 3

[62] Kai Su and Xin Geng. Soft facial landmark detection by label distribution learning. In *AAAI*, pages 5008–5015, 2019. 2

[63] Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE TIP*, 27(8):3998–4011, 2018. 2, 6

[64] Yuxiang Tang, Xuejing Kang, Chunxiao Li, Zhaowen Lin, and Anlong Ming. Transfer learning for color constancy via statistic perspective. In *AAAI*, pages 2361–2369, 2022. 6, 7

[65] Joost Van De Weijer, Theo Gevers, and Arjan Gijsenij. Edge-based color constancy. *IEEE TIP*, 16(9):2207–2214, 2007. 3

[66] Jin Xiao, Shuhang Gu, and Lei Zhang. Multi-domain learning for accurate and few-shot color constancy. In *CVPR*, pages 3255–3264, 2020. 7

[67] Bolei Xu, Jingxin Liu, Xianxu Hou, Bozhi Liu, and Guoping Qiu. End-to-end illuminant estimation based on deep metric learning. In *CVPR*, pages 3613–3622, 2020. 7

[68] Changdong Xu and Xin Geng. Hierarchical classification based on label distribution learning. In *AAAI*, pages 5533–5540, 2019. 1, 2

[69] Ning Xu, Yun-Peng Liu, and Xin Geng. Partial multi-label learning with label distribution. In *AAAI*, pages 6510–6517, 2020. 1, 2

[70] Tsun-Yi Yang, Yi-Hsuan Huang, Yen-Yu Lin, Pi-Cheng Hsiu, and Yung-Yu Chuang. Ssr-net: A compact soft stage-wise regression network for age estimation. In *IJCAI*, pages 1078–1084, 2018. 6

[71] Xu Yang, Bin-Bin Gao, Chao Xing, Zeng-Wei Huo, Xiu-Shen Wei, Ying Zhou, Jianxin Wu, and Xin Geng. Deep label distribution learning for apparent age estimation. In *ICCV*, pages 344–350, 2015. 1

[72] Ran Yi, Haoyuan Tian, Zhihao Gu, Yu-Kun Lai, and Paul L Rosin. Towards artistic image aesthetics assessment: a large-scale dataset and a new method. In *CVPR*, pages 22388–22397, 2023. 3, 5, 6

[73] Hui Zeng, Zisheng Cao, Lei Zhang, and Alan C Bovik. A unified probabilistic formulation of image aesthetic assessment. *IEEE TIP*, 29:1548–1561, 2019. 6

[74] Chao Zhang, Shuaicheng Liu, Xun Xu, and Ce Zhu. C3ae: Exploring the limits of compact model for age estimation. In *CVPR*, pages 12587–12596, 2019. 6

[75] Yunxuan Zhang, Li Liu, Cheng Li, and Chen Change Loy. Quantifying facial age by posterior of age comparisons. In *BMVC*, 2017. 5, 6

[76] Zhaoxiang Zhang, Mo Wang, and Xin Geng. Crowd counting in public video surveillance by label distribution learning. *Neurocomputing*, 166:151–163, 2015. 1, 2

[77] Zhen Zhang, Lei Zhu, Yaping Li, and Yang Xu. Deep discrete hashing for label distribution learning. *IEEE Signal Process. Lett.*, 29:832–836, 2022. 1

[78] Qilu Zhao, Junyu Dong, Hui Yu, and Sheng Chen. Distilling ordinal relation and dark knowledge for facial age estimation. *TNNLS*, 32(7):3108–3121, 2021. 6

[79] Rui Zheng, Shulin Zhang, Lei Liu, Yuhao Luo, and Mingzhai Sun. Uncertainty in bayesian deep label distribution learning. *Appl. Soft Comput.*, 101:107046, 2021. 1

[80] Xiuzhuang Zhou, Zeqiang Wei, Min Xu, Shan Qu, and Guodong Guo. Facial depression recognition by deep joint label distribution and metric learning. *IEEE Trans. Affect. Comput.*, 13(3):1605–1618, 2022. 1