# TAA-GCN: A Temporally Aware Adaptive Graph Convolutional Network for Age Estimation

Matthew Korban<sup>a</sup>, Peter Youngs<sup>b</sup>, Scott T. Acton<sup>a,\*</sup>

<sup>a</sup>Department of Electrical and Computer Engineering, University of Virginia, Charlottesville, VA 22904
<sup>b</sup>Department of Curriculum, Instruction and Special Education, University of Virginia, VA 22904

#### Abstract

This paper proposes a novel age estimation algorithm, the Temporally-Aware Adaptive Graph Convolutional Network (TAA-GCN). Using a new representation based on graphs, the TAA-GCN utilizes skeletal, posture, clothing, and facial information to enrich the feature set associated with various ages. Such a novel graph representation has several advantages: First, reduced sensitivity to facial expression and other appearance variances; Second, robustness to partial occlusion and non-frontal-planar viewpoint, which is commonplace in real-world applications such as video surveillance. The TAA-GCN employs two novel components, (1) the Temporal Memory Module (TMM) to compute temporal dependencies in age; (2) Adaptive Graph Convolutional Layer (AGCL) to refine the graphs and accommodate the variance in appearance. The TAA-GCN outperforms the state-of-the-art methods on four public benchmarks, UTKFace, MORPHII, CACD, and FG-NET. Moreover, the TAA-GCN showed reliability in different camera viewpoints and reduced quality images. Keywords: Age Estimation, Graph Convolutional Network, Facial Graphs, Skeletal Graphs.

# 1. Introduction

Age estimation has evolved from a carnival curiosity to an established task in computer vision [1]. It has many applications such as human-computer interaction (HCI), biometrics, age-restricted security control, video surveillance, and teacher-student differentiation in the classroom. However, age estimation brings several major challenges, including the following:

<sup>\*</sup>Corresponding author

Email addresses: acw6ze@virginia.edu (Matthew Korban), pay2n@virginia.edu (Peter Youngs), acton@virginia.edu (Scott T. Acton)

(1) the variance of appearance and facial expression, (2) viewpoint variations, and (3) non-ordinal temporal dependencies between ages. These challenges hinder the current standard methods to estimate age effectively. Therefore, we propose an algorithm including several novel components to handle the challenges above, which we explain as follows:

# 1.1. Variance in Appearance

People of the same age have remarkable variance in their appearance [2], which makes the age estimation challenging. To resolve this issue, some researchers have suggested age, gender and racial grouping [2, 3]. Nevertheless, such approaches fail when the grouping strategy is erroneous or when the same age groups are highly varying.

Current age estimation methods use raw images commonly with Convolutional Neural Networks (CNN) [1]. While there have been significant advances in developing effective CNN architectures such as AlexNet and VGG, CNNs models are commonly used for object/subject classification. So, they might be less effective for age estimation because the features obtained from CNN models differ from those in the face. As opposed to objects/subjects, facial information are commonly centralized around specific facial keypoints. Moreover, in contrast to object/subject classes (such as a car, house, or pedestrian) that often differ in non-localized regions, the difference between age classes is mainly defined as local differences around specific facial keypoints. This fact is the same for the ages of similar classes.

In contrast to raw images, facial keypoints provide a more potent representation of the face, eliminating unnecessary data [4] and yielding critical information. However, while some approaches used facial keypoints for age estimation [5], facial keypoints have not yet been effectively exploited in age estimation. It is because the current methods are based on 2D convolutional networks, which cannot effectively model the connectivity information in facial keypoints. Specifically, 2D convolutional operators are applied on a fixed image grid where there might not be any explicit connections between facial keypoints on the 2D image space.

An appropriate way to model facial keypoints and their connectivity is by using graphs. The Graph Convolutional Network (GCN) has shown to be efficacious in solving graph-based problems such as action recognition using skeletons [6]. Here, we propose a novel graph representation and a new GCN to model facial keypoints for age estimation effectively. To

better accommodate the variance in appearance in facial graphs, we introduce an Adaptive Graph Convolutional Layer (AGCL) that adaptively refines the graphs.

Another source of variance in the facial analysis is facial expressions that alter the structure of the face and facial keypoints. There have been a few recent studies investigating age estimation under facial expressions. For example, [7] learned both the expressions and the ages jointly. Their method, however, requires complex learning and prior knowledge regarding facial expressions. We put forth a simple yet effective algorithm to make the GCN less sensitive to facial expressions, unlike this joint approach. An example of facial expression-insensitive graph can be seen in Fig. 1 where  $f_2$ ,  $f_3$ , and  $f_4$  almost remain in the same positions despite varying facial expressions that alters  $f_1$  (face images are collected from the RaFD dataset [8]).

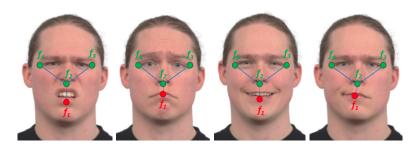


Figure 1: Four examples of simple facial graphs under different facial expressions. The appropriate graph nodes (green keypoints) are selected such that the features are less sensitive to the facial expression.

# 1.2. Viewpoint Variations

In typical real-life scenarios, videos are captured from different camera angles and viewpoints. Most of the reports in the literature are heavily dependent on the single-view frontal
face in which the face is frontal-planar with the imaging plane. A few existing strategies
accommodate various camera angles in age estimation, including the estimation of various
camera parameters [9] and geometric parameters [10]. However, these methods can handle
only limited variations in viewpoints. In real-world applications such as video surveillance,
the viewpoints are highly varying. As a result, a given face might be partially visible from
certain viewpoints. An example is shown in Fig. 2 in which a teacher's face and a student's
face are partially visible in given viewpoints (Fig. 2 is collected from a classroom video
dataset belonging to the University of Virginia). In such examples, distinguishing teachers

and students based on their ages has essential applications in some ongoing research such as teacher tracking and classroom activity recognition [11]. A standard age estimation algorithm encounters difficulty distinguishing their ages based on only facial information in these cases. Therefore, additional information is required to differentiate various ages when the viewpoint changes.



Figure 2: An example of a real-life scenario where skeleton and clothing information is more helpful than the face. In this frame sample from a classroom video, the teacher and student are captured from different viewpoints. As a result, their faces are only partially visible. However, their skeletons and clothing provide more critical information about their ages. To include this information, we propose to use Skeletal-Cosmetic (SC) keypoints (right image) which are obtained from detected skeletons (left image).

People belonging to different age groups have different skeleton structures, adopt specific postures, and wear particular clothes; an example is shown in Fig. 3 (we collected the people's images from the Relative Human dataset [12]) and removed the background for more clarity). Although these cues are "soft biometrics" rather than definitive features, such cues can be exploited in age estimation. The skeleton, posture, and clothing provide additional beneficial information in the attempt to distinguish different age groups, especially when the face is partially visible. As an example, in Fig. 2, the student and the teacher are now more distinguishable with their skeletal, posture and clothing information. (Fig 2 is collected from a dataset belonging to the University of Virginia [13]) The combination of skeleton and clothing provides a unique and strong feature space distinguishing different age groups. In this paper, for the first time, we propose the exploitation of the aforementioned additional soft biometrics, which we call *Skeletal-Cosmetic* (SC) information, to improve the age estimation performance in varying viewpoints. To better handle the variance in SC graphs, our AGCL adaptively refines SC graphs separately from facial graphs



Figure 3: Information beyond face is helpful for age estimation as people of different age groups have different skeleton structures, postures, and clothing.

# 1.3. Non-ordinal Temporal Dependencies

Aging is a temporal process, and as a result, neighboring ages are temporally dependent. An example of the aging process is shown in Fig. 4. Temporal dependencies help distinguish various age groups. However, these temporal dependencies are non-ordinal since age data samples are independent images from different individuals with no explicit temporal connections. In contrast, ordinal temporal sequences such as action video samples are typically continuous temporal frames showing the same individual in action. Therefore, standard temporal networks cannot capture such non-ordinal temporal dependencies among different ages.

Some strategies such as multi-stage classification [14], ranking [15], and grouping [16] use multiple classifiers to implicitly exploit the temporal properties of ages at a high computational cost. Nevertheless, there does not exist a computationally efficient approach that takes advantage of temporal dependencies among different ages explicitly. [17] proposes a kernel-based bi-directional PCA to find the kinship relationship between family members. However, this kinship relationship is limited to certain age groups (parent and child). Moreover, the proposed learning process is not end-to-end and heavily depends on multiple stages including a pre-processing feature extraction using PCA. To fill this gap, we put forth a new Temporal Memory Module (TMM) that captures non-ordinal temporal dependencies among a wide range of people with different ages. With a single-stage classification and end-to-end

network, our proposed method is also computationally efficient.

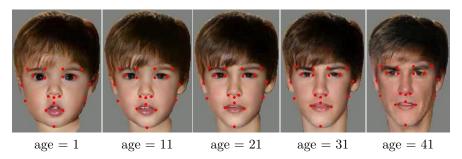


Figure 4: An image sequence exemplifying the temporal dependencies between neighboring ages.

# The main **contributions** of this paper are:

- We are the first to introduce a graph representation for age estimation. Our new graph representation has some benefits over the previous 2D image-based approaches: First, it more effectively models facial keypoints and their connectivity; Second, it allows to exploit our new Adaptive Graph Convolutional Layer (AGCL) and facial keypoint selection algorithm to make the age estimation less sensitive to facial expressions and other appearance variances.
- We propose a new Temporal-Aware Adaptive Graph Convolutional Network (TAA-GCNN) that includes a new Temporal Memory Module (TMM) to capture non-ordinal temporal dependencies among different ages. However, standard temporal networks cannot capture such non-ordinal temporal dependencies. We are the first to explicitly exploit non-ordinal temporal dependencies among different ages with a computationally efficient single-stage classification.
- Our TAA-GCN also includes another new module, Adaptive Graph Convolutional Layer (AGCL), that refines the facial and SC graphs adaptively to improve the age estimation performance under the variance of appearance.
- We are the first to include several soft biometrics including skeleton structure, posture and clothing (Skeletal-Cosmetic) information to improve the age features representation, especially in real-world applications where facial information is only partially available from certain viewpoints.
- We conducted thorough experiments to estimate the age in-the-wild.
- Our proposed method outperforms the state-of-the-art strategies as demonstrated on four public datasets, MORPHII, CACD, UTK-Face, and FG-NET.

# 2. Related Work

# 2.1. Age Estimation

Classical approaches utilized hand-crafted features such as wrinkles, landmark-based features for age estimation. [18] used natural wrinkles defined a search region on the face where facial wrinkles are more common. This early method could only estimate three age groups, baby and adult, and senior. [19] overcame this limitation by estimation a variety of ages using Multiscale Wrinkle Patterns (MWP) features. Similar to natural wrinkles [18], MWP also are defined on multiple search regions. However, MWP included several other attributes such as shape and textures to enrich the age feature representation. Some few research have been based on facial landmark features. [20] extracted Component extracted Bio-Inspired Feature (BIF) from facial landmarks using pyramid of convolution filters. [21] combines facial landmark points and gravity moment and builds a matrix that represents the the juvenile age range. Other features have been based on geometry, active shape, appearance [1], and relative-order information in different ages [22]. Different classifiers have been used with hand-crafted features such as Relevance Vector Machine (RVM) [23], ratio matching [18], and Support Vector Regression [19].

With the recent advancement of deep learning, deep networks have been used for age estimation. Some researchers suggested improving the training stage. For instance, [24] propose a CNN architecture to exploit age differences and reduce the number of training labels. [25] proposed a CNN architecture with a cumulative hidden layer and extracts discriminative aging features to resolve the issue of imbalance data. [26] suggested a label refinery network (LRN) to refine the age labels for a more effective training. [27] suggested a Progressive Margin Loss (PML) to include the dependencies between the intra-class and inter-class variance in various age groups. Some researchers proposed to improve the age feature representation by revising the deep network architecture. [28] suggested to use multi-scale output connections from different CNN layers to include diverse face features. [29] suggested using multiple features extracted from local and global regions. [30] proposed to weight important facial patches using Attention-based Dynamic Patch Fusion (ADPF).

Some researchers recommended assigning classifiers to various age groups. For example, [14] suggested a multi-stage classification approach for different age groups. [31] enhanced the multi-group classification using Ordinal Ensemble Learning. Another strategy for age

estimation has been using additional human attributes or features. For instance, [2] introduced an age grouping strategy including genders and sub-groups to facilitate the age estimation task. [3] proposed a deep conditional distribution learning which is conditioned to several attributes such as gender and age.

There have been a few approaches such as [32] that tried to estimate age with partial information. To accomplish this goal, they model different face regions, such as eyes and nose separately. However, [32] is still dependent on the high-quality frontal face and might not work in-the-wild scenarios when images are captured in reduced quality and from different camera angles. Our method, however, is reliable under various camera angles and for reduced quality images, as the skeleton, posture and clothing information are less affected than the face by camera viewpoints and image quality.

# 2.2. Graph Convolutional Network

Graph Convolutional Network (GCN) has been used in several computer vision tasks such as action recognition [33], brain disorder prediction [34], image retrieval [35], person re-identification [36], and recommendation systems [37]. There are different types of GCN which have been introduced based on various applications. An Spatial GCN can encode spatial properties of data such as image pixels [38]. A Temporal GCN computes the temporal dependencies of input like sequential traffic data [39]. A spatial-temporal GCN captures the information in both spatial and temporal domains such as pixels and sequential frames in activity videos [40].

For different tasks, the researchers designed various GCN architectures. For example, [33] designed a human pose-aware GCN to model the dependencies among human skeleton joints and body parts. [34] proposed a Hierarchical GCN to learn from different ROIs in fMRI data of the brain. [35] introduced a Siamese GCN to improve the discriminative property of image representations in image retrieval. [36] proposed a part-guided GCN to model the structural relationship in the learned features for person re-identification. [37] suggested an adversarial GCN to overcome the incomplete and noisy social network information for recommendation systems. According to our age estimation task, we designed a GCN which is (1) temporally aware of different age groups and (2) is adaptive to spatial-temporal variance in facial and SC graphs for different human faces, poses and clothing.

# 3. Methodology

#### 3.1. Overview and Terminology

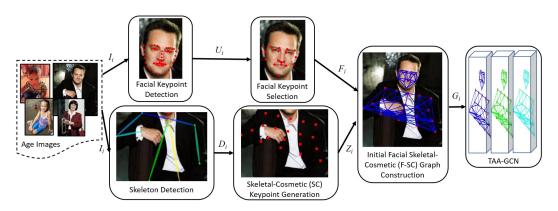


Figure 5: Overview of our proposed pipeline for age estimation. From a set of age images including  $I_i$ , we detect initial facial keypoints  $(U_i)$  and skeleton joints  $(D_i)$  using the methods of [41] and [42], respectively. Subsequently, a facial keypoint selection algorithm picks the most suitable subset of keypoints,  $F_i$ , that are more informative and less sensitive to facial expressions. Simultaneously, from  $D_i$  we generate the Skeletal-Cosmetic (SC) keypoints,  $Z_i$ , to include skeleton, posture, and clothing information. Then, given the previously obtained keypoints (and their corresponding patches of pixels), we construct an initial Facial Skeletal Cosmetic (F-SC) graph,  $G_i$ , by defining the initial connectivity (edges) among keypoints (nodes). Finally, our proposed Temporal-Aware Adaptive Graph Convolutional Network (TAA-GCN) estimates  $G_i$ 's age.

An overview of our proposed age estimation pipeline is shown in Fig. 5. Given a set of images of people of varying age  $I^S = \{I_i, i = 1, 2, \dots, \Lambda\}$ , for each  $I_i$ , we calculate a group of initial facial keypoints  $U_i = \{u_k, k = 1, 2, \dots N\}$  using [41], where B is the number of samples in the dataset, and N is the number of initial facial keypoints. Concurrently, for each  $I_i$ , we calculate a collection of skeleton joints  $D_i = \{d_k, k = 1, 2, \dots, M\}$  utilizing OpenPose [42], where M is the number detected joints. Next, we use our Facial Keypoint Selection algorithm to select the most informative, yet least expression sensitive facial keypoint indices as  $R = \{r_k, k = 1, 2, \dots, N'\}$ . We obtain R offline and during the pre-training phase and use it as a fixed parameter to select more effective facial keypoints as  $F_i = U_i(R)$  in the training and testing phases. On the other side, our Skeletal-Cosmetic Keypoint Generation algorithm processes  $D_i$  to generate SC keypoints  $Z_i = \{z_k, k = 1, 2, \dots, M'\}$  which represents both the skeleton and clothing (M' here is the number of SC keypoints). For  $F_i$  and  $Z_i$  we create two sets of initial feature vectors of  $X_i^F$  and  $X_i^Z$ , respectively. The full set of graph nodes is formed by concatenating the two sets of initial feature vectors as  $V_i = \{X_i^F, X_i^Z\}$ .

For facial and SC graph nodes (keypoints), we predefine initial adjacency matrices as  $A^F$ 

and  $A^Z$ , respectively.  $A^F$  and  $A^Z$  indicate the connectivity information among the graph nodes in the face and body. Using the full adjacency matrix  $A = \{A^F, A^Z\}$  (corresponding to both facial and SC graphs) and full set of nodes  $V_i$ , we construct an initial Facial Skeletal-Cosmetic (F-SC) graph as  $G_i = \{V_i; E_i\}$ , where  $E_i$  is the set of graph edges and obtained from A. Finally, our TAA-GCN estimates  $G_i$ 's age  $q_i$ . Our TAA-GCN includes two new modules to improve the age estimation performance. First, AGCL to refine  $G_i$  adaptively. Second, TMM to capture non-ordinal temporal dependencies among various ages. We explain all the terminologies and definitions with the corresponding reference sections of this manuscript in Table 7

Table 1: The terminologies used in this paper with descriptions and corresponding sections.

Terminology symbol	Description	Section
$I^S$	set of images $I_i$ in the dataset, $0 \le i \le \Lambda$	3.1
$F_i, R$	selected facial keypoints and their indices	3.2 , 3.1
$U_i$	initial set of facial keypoints $u_k$ before selection	3.2
K,eta	number of neighboring keypoints, and neighboring nodes set	3.2
$\zeta_k$	relative distance in a facial neighboring keypoints	3.2
H	facial neighboring keypoints normalization term	3.2
$v_k^E$ . $v_k^A$	facial expression, and age variance	3.2
$\eta$	keypoint selection weighting parameter	3.2
N,M	number of initial facial keypoints and skeleton joints	3.2,  3.3
N'. M'	number of selected facial, and generated SC keypoints	3.2,  3.3
$D_i$	set of skeleton joints $d_k$	3.3
O, J	number of hierarchical levels, and graph nodes	3.3
$X_i^F$	set of initial facial feature vectors $x_i^F$	3.4
$X_i^F \ X_i^Z \ A^F, A^Z$	set of initial SC feature vectors $x_i^Z$	3.4
$A^F, A^Z$	facial, and SC adjacency matrix	3.4
$G_i$ , $V_i$ , $E_i$	F-SC graph, nodes, and edges	3.4
$e\;, ho$	nodes edge, and correlation value	3.4
$P_{spt}, P_{temp}$	Spatial, and Temporal age probability	3.5
Q, a	number of ages, and age label	3.5
$W^F, W^Z,$	adaptive facial, and SC graph edge weights	3.5
$\phi, psi$	adaptive activation function, and age weighting parameter	3.5

# 3.2. Facial Keypoint Selection

Our new graph representation of the face allows us to select the most informative facial keypoints. Given the initial detected facial keypoints  $(U_i)$ , our facial keypoint selection algorithm picks the keypoints  $(F_i)$  that are less sensitive to facial expressions while more sensitive to aging. We observed that relative distances between neighboring keypoints often change noticeably under different facial expressions. An example is shown in Fig. 1,

where the relative Euclidean distances between two keypoints  $d(f_1, f_2)$  change during the expression "smiling". In the aging process, however, the global positions of keypoints shift more dramatically. We exploit this fact to select the most effective facial keypoints based on the facial expression and aging variances. We selected the most effective facial keypoints based on the images captured in-the-wild such as those for the UTKFace dataset [43]. We also further evaluated our method based on classical facial expressions on the PAL dataset [44]. Our facial keypoint selection algorithm aims to find the facial keypoints that are less sensitive to facial expressions while being informative enough to improve the age estimation performance. We designed such a trade-off in a data-driven way based on the datasets we used. The datasets include a variety of different individuals from different cultures and various ages.

Facial Expression Variance. First, for each facial keypoint  $f_k$ , we define a set of neighboring keypoints as  $\beta(f_k) = \{f_n, n = 0, 1 \cdots K\}$ , where K is the number of neighboring keypoints. Then, we calculate the sum of the relative Euclidean distance (d) of each keypoint  $f_k$  to its neighbors as:

$$\zeta_k = \frac{1}{H} \sum_{n=1}^{K} d(f_k, f_n)$$
 (1)

In the above,  $H = \max_{1 \le i \le N} (\zeta_i)$  is a normalization term,  $f_n \in \beta(f_k)$ , where N is the number of keypoints. For each  $f_k$ , we compute the facial expression variance  $v_k^E$  across all the samples in the dataset as:

$$v_k^E = \frac{1}{N} \sum_{k=0}^{\Lambda} (\zeta_k - \frac{1}{N} \sum_{k=0}^{\Lambda} \zeta_k)^2$$
 (2)

Age Variance. We define the global distance of  $f_k$  with respect to the root keypoint  $f_R$  as  $\gamma_k = d(f_k, f_R)$ . For each  $f_k$ , we calculate the age variance  $v_k^A$  across all the images in the dataset:

$$v_k^A = \frac{1}{N} \sum_{k=0}^{\Lambda} (\gamma_k - \frac{1}{N} \sum_{k=0}^{\Lambda} \gamma_k)^2$$
 (3)

Keypoint Selection. We select the facial keypoints with the lowest facial expression variances and highest age variances. Specifically, among the keypoints with the indices of

 $k \in N$ , we select top-N' keypoints with the highest  $v_k^T$ :

$$v_k^T = \eta \cdot v_k^A + (1 - \eta) \cdot (1 - v_k^E) \tag{4}$$

In the above,  $\eta$  is a weighting parameter. We store the selected keypoint indices  $R = \{r_k, k = 1, 2, \dots, N'\}$  to pick the selected facial keypoints  $F_i = U_i(R)$  during the training and testing phases. The value for  $\eta$  is obtained experimentally. Specifically, we first select uniform sampling values from the interval [0, 1] and narrow them down to the smaller interval that maximizes the overall age estimation performance.

# 3.3. Skeletal-Cosmetic (SC) Keypoint Generation

Pose estimation algorithms such as OpenPose [42] extract spatially consistent and stable keypoints which have been used in many applications such as action recognition using skeletal graphs [6]. Hence, we use the detected joints by the OpenPose as a reliable backbone to generate SC keypoints that represent skeletons, postures, and clothing. Given the initially detected skeleton joints  $D_i$ , we generate the SC keypoints,  $Z_i$ , by interpolating between  $d_k \in D_i$ . The main challenge here is that the number of detected joints varies based on the visibility of persons in different images. Therefore, different Hierarchical Levels (HL) can be detected for each image. We define a HL with O levels, as a set of joints that have a similar parent-child ranking in a human body skeleton. For example two shoulders or arms are in the same HL. Within the same HL, the number of detected joints can be also vary for different image samples according to the camera position. Consequently, we start our SC generation algorithm by interpolating SC keypoints among the same HL (parent joints) and then continue to the next HL (child joints). Our SC Keypoint Generation algorithm is illustrated in Algorithm 1. Two examples of our SC keypoint generation outputs are in Fig. 2 (right) and Fig. 6 (right).

The skeleton pose and facial landmark extraction algorithms that we used, [42] and [41] respectively, can relatively interpolate the missing parts when partial occlusion happens. To avoid the missing regions in the human body further, we only used the upper body of the human with HL O=4 (as described in Section 3.6). We observed that the upper body of the human provides sufficient skeleton information to model different age groups. Additionally, our SC keypoint generation algorithm can interpolate the missing keypoints of a side of the human body given the opposite side. For example, it can interpolate the right arm joint

given the available left arm joint. In the extreme and less common cases when the facial landmark and skeleton pose extraction algorithms fail and also the upper body part also is not available, we set the missing values to zeros.

# Algorithm 1 SC Keypoint Generation

```
Require: The detected skeleton joints, D
Ensure: SC keypoints, Z
  i = 0
Add D to Z
                                                                  \triangleright all the detected joints are added to Z
  while i \leq O do j = 0
                                                                 \triangleright O is the number of HL in the skeleton
                                                                                         \triangleright J is in the same HL
       while j \leq J do
           if d^{j}(o_{i}) exists then
                                                                      \triangleright d^j and d^{j+1} are neighboring nodes
               if d^{j+1}(o_i) exists then
                    Add Interpolation (d^{j}(o_{i}), d^{j+1}(o_{i})) to Z
                                                                                               ▷ in the same HL
           \begin{array}{c} \text{end if} \\ \text{end if} \end{array}
           j \leftarrow j + 1
           if d^{j}(o_{i+1}) exists then
                                                                 ▷ check neighboring nodes from next HL
                Add Interpolation(d^{j}(o_{i}), d^{j}(o_{i+1})) to Z
                                                                                     ▷ in two neighboring HL
           else if D^{j}(o_{i+1}) does not exist then
                if D^{j}(o_{i+1}) is not end-effector then
                                                                                          ▷ like wrists and feet
               Add 0 to Z end if
           end if
       end while i \leftarrow i + 1
  end while
```

# 3.4. Facial Skeletal-Cosmetic Graph Construction

After obtaining  $Z_i$  (SC keypoints) and  $F_i$  (facial keypoints) in the previous steps, we construct a graph for each image sample  $I_i$  to provide input to our TAA-GCN. We create the initial sets of feature vectors for  $F_i$  and  $Z_i$  as  $X_i^F$  and  $X_i^Z$ , respectively. To create the sets of feature vectors,  $\forall z_k \in Z_i$  and  $\forall f_k \in F_i$ , we assign an initial feature vector of  $x_k^f \in X_i^F$  and  $x_k^z \in X_i^Z$ , respectively. Specifically,  $x_k^f$  and  $x_k^z$  are created by concatenating the patch of pixels around the keypoints  $f_k$  and  $f_k$  and their 2D coordinates. We tile the 2D coordinates of keypoints to match the patch of pixels size. In this hybrid feature representation, the patches of pixels provide cosmetic (clothing) information, while 2D coordinates give information about skeletal structures and postures.

An example of the aforementioned feature representation for each keypoint can be seen in Fig. 6 (the image is collected from the UTKFace dataset). Each  $x_k^f$  and  $x_k^z$  then are converted to 1D vectors. Subsequently, the sets of feature vectors,  $X_i^F$  and  $X_i^Z$ , are fed

to the TAA-GCN. For two facial and SC graph nodes, we construct the initial adjacency matrices of  $A^F$  and  $A^Z$  by calculating the most correlated nodes. The correlation between each pair of nodes,  $x_i$  and  $x_j$ , is calculated across all the dataset samples as:

$$\rho_{ij} = \frac{cov(x_i, x_j)}{\sigma x_i \cdot \sigma x_j}.$$
 (5)

So, Aij = 1 if  $\rho_{ij}$  is among the top-n correlated values, otherwise Aij = 0. Finally, we create the initial graph as  $G_i = \{V_i; E_i\}$ , where  $V_i = \{X_i^F, X_i^Z\}$  and  $E_i$  is obtained from  $A = \{A^F, A^Z\}$ .



Figure 6: The left image shows the detected skeleton joints and facial keypoints; the right image indicates the final keypoints after using facial keypoint selection and SC keypoint generation algorithms. For each facial and SC keypoint  $f_k$  and  $z_k$ , we assign an initial feature vector  $x_k^f$  and  $x_k^z$  by concatenating the patch of pixels (around each keypoint) and the 2D coordinate of the keypoint, (X, Y). In this feature representation, the 2D joint coordinates serve as the skeleton/posture structure, and the patches of pixels represent the clothing.

# 3.5. Temporal-Aware Adaptive Graph Convolutional Network (TAA-GCN)

The pipleline of our proposed TAA-GCN is shown in Fig. 7. The input of the TAA-GCN is a F-SC graph,  $G_i = \{V_i; E_i\}$ , and the output is an age label  $\hat{q}_i \in q^S = \{q_t, t = 0, 1, \dots, Q\}$ , where  $V_i = \{v_k \in \mathbb{R}^2, k = 0, 1 \dots J\}$ ,  $E_i = \{E_i^F, E_i^Z\}$ , and Q is the number of age labels (maximum age). Here, J = M' + N' is the total number of nodes,  $E_i^F$  and  $E_i^Z$  are facial and SC graph edges, respectively. Our TAA-GCN includes several Adaptive Graph Convolutional Layers (AGCL) to refine  $E_i$  so that the updated graph edges,  $\hat{E}_i$ , become a more effective representation of connectivity among  $V_i$ . Specifically, by updating the graph edges, the TAA-GCN accommodates the variance in  $E_i$ , which is caused by variance in facial and SC graphs according to different faces, skeletons, postures, and clothing.

The AGCN learns the spatial information in facial and SC graphs and outputs  $L_0$ , a  $C \times J$  feature vector, where C is the number of channels.  $L_0$  is processed through Average Pooling (AP), 1D convolutional (Conv) and Softmax layers to output the spatial age probability  $P_{spt}(q^S|G_i)$ . The above spatial dataflow is also illustrated in Eq. 6, where  $L_0 = AGCN(G_i)$ . The architecture of the AGCN which include several AGCLs is shown in Fig. 8.

Simultaneously, the TMM processes a  $Q \times J$  feature vector to learn the non-ordinal temporal dependencies among different ages. The TMM outputs the temporal features,  $L'_1$ , a  $Q \times J$  feature vector.  $L'_1$  then passes through convolutional and Softmax layers to compute  $L'_3$ , which is the temporal age probability  $P_{temp}(q^S|G_i)$ . The aforementioned temporal dataflow is also shown in Eq. 7. The architecture of the TMM can be seen in Fig. 9.

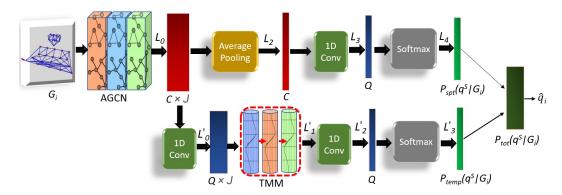


Figure 7: Our proposed TAA-GCN pipeline which includes Adaptive GCN (AGCN) and a new Temporal Memory Module (TMM). Given the initial graph,  $G_i$ , the AGCN extract features,  $L_0$  by refining the edges in  $G_i$ .  $L_0$  is then processed to compute the spatial age probability  $P_{spt}$ . Simultaneously, the TMM computes the non-ordinal temporal dependencies in  $L_0$  to find the temporal age probability,  $P_{temp}$ . The final age prediction is obtained by a weighted summation of two spatial and temporal probabilities.

$$P_{spt}(q^S|G_i) = Softmax(Conv(AP(L_0)))$$
(6)

$$P_{temp}(q^S|G_i) = Softmax(Conv(TMM(Conv(L_0))))$$
(7)

Lastly, we compute the final age prediction as is shown in Eq. 8. The final loss  $Loss_F$  can be seen in Eq. 9, where MSE is Mean Square Error,  $q_g$  is the ground truth value for age, and  $\omega$  is an adjustment weight parameter. Additionally,  $\hat{q}_{spt}$  and  $\hat{q}_{temp}$  are spatially

and temporally predicted ages, respectively.

$$P_{tot}(q^S|G_i) = \omega \cdot P_{spt}(q^S|G_i) + (1 - \omega) \cdot P_{temp}(q^S|G_i)$$
(8)

$$Loss_F = \omega \cdot MSE(\hat{q}_{spt} - q_q) + (1 - \omega) \cdot MSE(\hat{q}_{temp} - q_q)$$
(9)

We will explain both TMM and AGCL in the following:

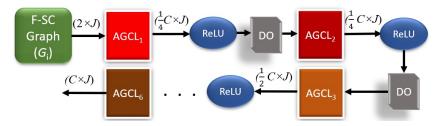


Figure 8: The architecture of the proposed Adaptive Graph Convolutional Network (AGCN). Here, DO signifies a Dropout layer, C=256, and the activation function is the Rectified Linear Unit (ReLU).

# Temporal Memory Modules.

Standard temporal networks follow an ordinal temporal structure to capture temporal dependencies in sequences. Such a temporal structure is irrelevant in age estimation since there is no explicit temporal connection between consecutive age samples. In contrast, our TMM can capture temporal dependencies among temporally non-ordinal age samples. As shown in Fig. 7, the input of the TMM is a  $Q \times J$  feature vector, where each raw represents different ages as  $\{q_t, t = 0, 1, \dots, Q\}$ . The TMM captures the temporal dependencies among different  $q_t$  in a recurrent manner which is shown in Fig. 9. In this figure, x is the input, x is the output, and x is the hidden state.

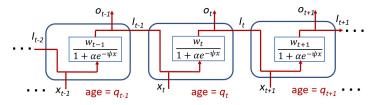


Figure 9: Our proposed Temporal Memory Module. With using our new adaptive activation function including the trainable parameter  $W_i$  and  $\psi$ , we extract non-ordinal temporal dependencies in age features,  $x_t$ , for various ages,  $q_t$ .

We exploit the fact that people of similar ages share more similarities than different ages. For example, people aged 20 and 21 look more similar, wear more similar clothing and have more similar poses than those aged 20 and 60. Our proposed TMM can capture the dependencies between both close age groups such as 20 and 21 as well as distant age groups like 20 and 60. By considering such temporal dependencies among age groups, we assist the age estimation network to (1) decrease the intra-class variations by finding similarities among the same ages, and (2) increase the inter-class variations by learning the differences among different ages. Especially, the latter is crucial to avoid significant errors such as confusion between ages 20 and 60 that lead to a remarkable decrease in the overall network performance. To make our TMM aware of such age group dependencies, we propose an adaptive activation function,  $\phi$ , illustrated in Eq. 10.

$$\phi = \frac{W}{1 + \alpha e^{-\psi x}}, \quad \psi = \frac{1}{|q_t - q_q|} \tag{10}$$

In the above,  $q_g$  is a ground truth age,  $\alpha > 1$  is an adjustment parameter, and  $W = \{w_t, t = 0, 1 \cdots, Q\}$  are adaptive weight parameters.  $\psi$  weights different age groups and ensures that during the training, the ages closer to the ground truth are assigned higher weights than distant ages. In the extreme cases, when two ages are the same,  $|q_t - q_g| \longrightarrow 0$ , then  $\psi \longrightarrow \infty$ , and so  $\phi \longrightarrow w_t$ . On the other hand, when  $|q_t - q_g| \longrightarrow Q$ , then  $\psi \longrightarrow 0$ , and so  $\phi \longrightarrow \frac{w_t}{1+\alpha}$ . The adaptive weight parameters memorize this weighting procedure for different  $q_t$ . In the testing phase,  $\psi$  is set to 1, and we use W to weight different  $q_t$ .

#### Adaptive Graph Convolutional Layer (AGCL).

The graph convolutional operator is applied over each  $v_k$ 's adjacent nodes  $\beta(v_k)$ . The set of graph edges,  $E_i$ , includes all  $\beta(v_k), v_k \in V_i$ . Hence, finding appropriate  $E_i$  is crucial for applying effective convolutions over  $\beta(v_k)$  that maximizes the network performance. Due to the variance in facial and SC graphs for different faces, skeletons, postures, and clothing, finding appropriate  $E_i$  is challenging. To solve this challenge, some researchers have suggested using dynamic temporal graphs for temporal problems such as skeleton-based action recognition [6]. Our case, however, is different because: (1) age estimation is a spatial problem, where many node candidates are equally important. For example, in a temporal sequence such as "running" few joints are involved in the action. In different ages, however many facial and SC keypoints have significant equal roles; (2) in contrast to the skeletal graphs [6], our F-SC graph consists of two separate facial and SC graphs, with

no explicit connection between them. So, instead of eliminating the nodes, we propose a weighting strategy that prioritizes the importance of node connectivity in  $E_i$ . Moreover, we utilize two separate weighting functions to refine the edges in facial and SC graphs.

Our AGCN consists of several AGCLs that refine  $E_i$  separately for facial and SC graphs via updating the weight of each edge  $e_{mn} \in E_i$  which connects  $(v_m, v_n) \in V_i$ . We define the relationship between the input  $(f_{in})$  and the output  $(f_{out})$  of each AGCL for each  $v_m \in V_i$  as:

$$f_{out}(v_m) = \sum_{v_n \in \beta(v_m)} f_{in}(v_n) \cdot (f_c(e_{mn}))$$
(11)

In the above,  $f_c$  is our proposed correlation weighting function and  $e_{mn} = \{e_{mn}^F, e_{mn}^Z\}$ , where F and Z represent for facial and SC graphs, respectively. The correlation weighting function,  $f_c$ , ensures that highly correlated nodes have larger weights on their connecting edges. To find the correlations among nodes we compute the cosine similarities  $(C_S)$  between the pair of nodes  $v_m$  and  $v_n$  as:

$$C_{S(v_m, v_n)} = \frac{v_m \cdot v_n}{\|v_m\| \|v_n\|} \tag{12}$$

We selected the cosine distance since compared to Euclidean distance is not sensitive to node vectors' magnitudes. Specifically, the Euclidean distance fails to capture the correlation between highly correlated nodes that are located distantly in the Euclidean space. For a similar reason, we could not use the city block distance.

To include the correlation information between each nodes  $(v_m, v_n) \in V_i$ , we update each edge  $e_{mn} \in E_i$  using  $f_c$ . For SC and facial graphs, the facial and SC correlation weighting functions,  $f_c^Z$  and  $f_c^F$ , can be seen in Eq. 13 and Eq. 14, respectively.

$$f_c^Z(e_{mn}^Z) = W^Z \cdot C_{S(v_m, v_n)} \cdot e_{mn}^Z$$
 (13)

$$f_c^F(e_{mn}^F) = \frac{W^F}{1 + C_{S(v_m, v_n)} \cdot e_{mn}^{F^{-\gamma_n}}}$$
(14)

In the above equations,  $W^Z$  and  $W^F$  are trainable parameters. Since facial graphs might not change dramatically similar to SC graphs, for facial graphs, we magnify the correlation change in  $e_{mn}^F$  by a logistic function with the power of  $\gamma_n$ . Specifically,  $\gamma_n \in 1 \leq \mathbb{Z} \leq K$  is the order of  $e_{mn}^F \in E_i(\beta(v_m))$ , where K is the number adjacent nodes in  $\beta(v_m)$ . Namely, for each  $v_n \in \beta(v_m)$ ,  $e_{mn}^F$  is ordered based on the correlation value between  $v_n$  and  $v_m$ . So, the optimal values of  $\gamma_n$  for  $v_m$  and its neighboring nodes  $v_n$  is calculated based on order of correlation values between neighboring nodes. We used this optimal weighting because it is consistent for all nodes  $V_i$  since K is a fixed parameter.

We separately update  $e_{mn}^F$  and  $e_{mn}^Z$  by defining two separate correlation weighting functions  $f_c^F$  and  $f_c^Z$  to ensure that the original connectivity importance in the facial and SC graphs are preserved. Each  $e_{mn}$  is adaptively updated after each convolutional operation between  $\beta(v_m) * v_m \cdot f_c(e_{mn})$  which also updates  $W^Z$  and  $W^F$  in each gradient descent iteration.

### 3.6. Implementation Details

Table 2 summarizes the implementation details of our proposed pipeline referenced in relevant sections of this paper.

Table 2: Implementation details of our proposed pipeline with associated paper section references.

Parameter	Value	Section
Number of Initial Facial Keypoints $(N)$	68	3.1
Number of Skeleton Joints $(M)$	18	3.1
Number of Selected Facial Keypoints $(N')$	19	3.2
Number of Neighboring Facial Keypoints $(K)$	5	3.2
Facial Keypoints Selection Weighting Parameter $(\eta)$	0.4	3.2
Number of SC Keypoints $(M')$	20	3.3
Number of Hierarchical Levels $(O)$	4	3.3
Size of Patch of Pixels	32x32 pixels	3.4
Number of AGCLs	6	3.5
Number of Channels in AGCL (each layer)	(64, 64, 128, 128, 256, 256)	3.5
Number of Layers in TMM	5	3.5
Number of Channels in TMM (for all layers)	64	3.5
Learning Rate	$1e^{-5}$	3.5
Number of Training Epochs	300	3.5
Optimizer	ADAM	3.5
Weight Decay	$1e^{-6}$	3.5
Loss Adjustment Weight $(\omega)$	0.65	3.5

# 4. Experimental Results

# 4.1. Datasets

For our major comparative experiments, we used four age datasets, CACD [45], MOR-PHII [45], UTKFace [43], and FG-NET [46], in our experiments. For the FG-NET, MOR-PHII, and CACD datasets we only used our facial graphs since they do not include the human body.

**UTKFace.** Consists of 24,000 images of 116 ages. We used 20,000 samples for training and 4,000 samples for testing. For now, it is the only major publicly-available age dataset that includes uncropped images with sufficient views of clothing. So, the dataset is well matched with our network when both facial and SC graphs are used.

**MORPHII.** Includes 55,000 samples of 13,000 subjects ranging from age 16 to 77. We used 44,000 samples for training and 11,000 samples for testing.

**CACD.** Includes 160,000 sample images of 2,000 celebrities from 49 age categories. In our experiments, the training, testing, and validation sizes are 127,000, 31,000 and 12,000 samples, respectively. As the CACD dataset only includes faces, we tested this dataset only on our facial graphs.

**FG-NET.** Consists of about 1,000 samples from 69 different ages, and 82 individuals. We used 900 samples for training and 100 samples for testing. This dataset also includes only face images.

#### 4.2. Comparative Results

We compared our method with the state-of-the-art approaches on four datasets, UTK-Face, CACD, MORPHII, and FG-NET, with the results shown in Table 4 and Table 6, respectively. We used the Mean Absolute Error (MAE), an standard evaluation metric in age estimation. The description of MAE is shown in Equation 15, where  $y_i$  is the estimated age for each test sample i,  $y_i^G$  is the ground truth, and N is the number of test samples.

$$MAE = \sum_{i=0}^{N} |y_i - y_i^G| \tag{15}$$

Our proposed approach outperformed the state-of-the-art methods on all four benchmarks. Specifically, for the UTK-Face dataset, with an MAE of 3.48, our TAA + F-SC graph outperformed the best previous record by approximately one year in age accuracy.

For the MORPHII dataset, we obtained an MAE of 1.69, which outperformed the other methods. For the CACD dataset, which is the largest existing age dataset, we obtained an MAE of 4.09, which improves upon the currently best results by 0.54 years in age accuracy. Our proposed method outperforms the existing approaches that did not use the IMDB-WIKI dataset pre-trained weights/features on the FG-NET dataset. We improved the current benchmark by 0.34 years. Among the methods that used IMDB-WIKI dataset for pre-training however, DAG-CNN [28] achieved the best performance. We outperformed the other works on two large MORPHII and CACD datasets, though many of them used IMDB-WIKI pre-trained weights/features.

Table 3: Comparison of our method, TAA-GCN + Facial graph and the state-of-the-art methods on the MORPHII dataset.

Team	Method	MAE
Lu et al. [47]	ORMO	3.27
Liu et al. [2]	$\operatorname{SAF}$	2.97
Zhang et al. [48]	C3AE	2.75
Cao et al. [49]	RCOR	2.64
Yang et al. [14]	SSR-Net	2.52
Pan et al. [50]	MVL	2.51
Shen et al. [51]	LDIR	2.24
Shen et al. [52]	DRF	2.17
Ours	TAA-GCN+F	1.69

Table 4: Comparison of our proposed method, TAA-GCN + F-SC graph and the state-of-the-art methods on the UTKFace dataset.

Team	Method	MAE
Rothe et al. [53]	DEX	4.31
Yoshimura et al. [54]	FOSS	4.49
Cao et al. [49]	RCOR	5.39
Al et al. [55]	SDTL	4.86
Li et al. [56]	SAM	4.77
Berg et al. [57]	DOR	4.55
Sun et al. [3]	DCD	4.47
Ours	TAA-GCN+F-SC	3.48

# 4.3. Ablation Study

We conducted an ablation study to evaluate the impact of the constituent components of our proposed pipeline on the overall age estimation performance. For our study, we define some abbreviations that are shown in Table 7. We carried out our ablation study on the

Table 5: Comparison of our method, TAA-GCN + Facial graph and the state-of-the-art methods on the CACD dataset.

Team	Method	MAE
Lu et al. [47]	ORMO	5.36
Cao et al. [49]	RCDR	5.24
Yang et al. [14]	SSR-Net	4.96
Zhang et al. [48]	C3AE	4.88
Rothe et al. [53]	DEX	4.78
Shen et al. [51]	LDIR	4.73
Shen et al. [52]	DRF	4.63
Ours	TAA-GCN+F	4.09

Table 6: Comparison of our method and the state-of-the-art methods on the FG-NET dataset. The results for this small dataset are reported for two categories: (1) methods that employed training from scratch rather than pre-trained weights and features; methods that used IMDB-WIKI pre-trained weights/features.

Methods NOT pre-trained	DE [58]	CS-LBF [59]	GADF [60]	SAF [2]	Ours
MAE	4.80	4.36	3.93	3.92	3.58
Methods pre-trained	DEX [53]	MVL [50]	C3AE [48]	DAG-CNN [28]	
MAE	4.63	4.10	4.09	3.05	

UTKFace dataset because it includes uncropped images with visible clothing, allowing us to test various types of graphs. We compare two facial and SC graphs (keypoints and patches of pixels) separately in Table 8. The overall age estimation performance changed when different graphs (facial or skeletal-cosmetic) were used.

We encode the skeletal and posture information with our Skeletal-Cosmetic Keypoints (SCK), while we obtain the clothing information from our Skeletal-Cosmetic Image Patches (SCIP). As can be seen in Table 8, using the skeleton/posture (SCK) information alone led to slightly better performance (MAE=5.50) compared to using clothing information alone (MAE=5.56). The combination of skeleton/posture and clothing information (SCK+CSIP), however, yields the best performance (MAE=5.45). Table 9 illustrates the impact of the combination of facial and SC graphs (skeleton/posture and clothing information) on the overall age estimation performance. Overall, combining all facial, skeleton/posture, and clothing features (FK+SCK+FPP+SCIP) led to superior performance.

Without facial information, and by using only SC graphs and SCK+SCIP features, still our proposed approach achieved competitive performance (with an MAE of 5.45 according to Table 8). Our age estimation algorithm works well without facial information, using only skeleton, posture and clothing information. So, our proposed algorithm is practical in real-world scenarios such as surveillance systems, where facial information is partially

available.

We also analyzed the impact of the constituent components of our proposed network, which is shown in Table 10 (with facial keypoint selection) and Table 11 (without facial keypoint selection). We achieved the best performance (with an MAE of 3.48) when we jointly used all our proposed modules (TAA-GCN) with our facial keypoint selection algorithm. Moreover, we evaluated the impact of the trainable facial and SC graph edge parameters,  $W^F$  and  $W^Z$ , on the overall age estimation performance illustrated in Table 12. To evaluate the generalization of our proposed approach, we conducted a cross-dataset evaluation on the MORPHII and CACD datasets which have sufficient numbers of samples for appropriate training. The results are shown in Table 13.

Table 7: The abbreviations used in the ablation study.

Abbreviation	Definition
FK	Facial Keypoints (2D coordinates)
FPP	Facial Patches of Pixels
SCK	Skeletal-Cosmetic Keypoints (2D coordinates)
SCIP	Skeletal-Cosmetic Image Patches
TAA-GCN	Temporal-Aware Adaptive GCN
AGCN	Adaptive GCN (no temporal awareness)
TA-GCN	Temporal-Aware GCN (no adaptivity)
GCN	Baseline GCN (no TA and adaptivity)

Table 8: The impact of facial/SC graphs and their keypoints/patches of pixels, individually, on the overall age estimation performance. The skeletal and posture information are encoded using the Skeletal-Cosmetic Keypoints (SCK). The clothing information are encoded using Skeletal-Cosmetic Image Patches (SCIP).

Modules	FK	FPP	FK+FPP	SCIP	SCK	SCK+SCIP
Modality	face	face	face	clothing	skeleton/posture	clothing+skeleton/posture
MAE	5.51	5.05	4.18	5.56	5.50	5.45

Table 9: The impact of the skeletal and posture (SCK) and clothing (SCIP) information, combined with facial information on the overall age estimation performance.

Modules	FK+FPP+SCK	FK+FPP+SCIP	FK+FPP+SCK+SCIP
Modality	face+skeleton/posture	face+clothing	face+skeleton/posture+clothing
MAE	3.64	3.75	3.46

Table 10: The impact of the constituent components of our proposed network on the overall age estimation performance when the facial keypoint selection algorithm is used.

Modules	TAA-GCN	AGCN	TA-GCN	GCN
MAE	3.48	4.83	4.02	5.33

Table 11: The impact of the constituent components of our proposed network on the overall age estimation performance, when the facial keypoint selection algorithm is **not** used.

Modules	TAA-GCN	AGCN	TA-GCN	GCN
MAE	4.01	5.60	4.78	5.98

Table 12: The impact of the trainable facial and SC graph edge parameters,  $W^F$  and  $W^Z$ , on the overall age estimation performance.

trainable parameter $W^F$	adaptive	adaptive	non-adaptive	non-adaptive
trainable parameter $W^Z$	adaptive	non-adaptive	adaptive	non-adaptive
MAE	3.48	4.43	4.20	4.78

Table 13: Cross-dataset analysis between the MORPHII and CACD datasets.

Scenario	Trained on MORPHII	Trained on CACD
	tested on CACD	tested on MORPHII
MAE	5.56	3.02

# 4.4. Images in-the-Wild

The images from the UTKFace [43] that we used in our experiments are captured inthe-wild. To further evaluate our proposed approach in-the-wild condition, we also used
the Relative Human dataset [12] which includes a variety of images captured in-the-wild.
The Relative Human dataset includes 24800 image samples with a variety of partial face
views commonly due to the impact of different camera angles or subjects/objects overlap
in-the-wild. Such the in-the-wild condition also shares many similarities with surveillance
conditions, where people are captured from different camera viewpoints. Some image examples are shown in Fig. 10. The dataset also includes both human face and body which
is useful in our experiments. The Relative Human dataset provided the label set consisting of four age groups "baby", "kid", teenager", and "adult". For evaluation, we used the
"top-1 score", a common metric in classification problems, that matches the top class with
the highest probability and the target label. We reported the age estimation results for
different combinations of facial and SC graphs (face, skeleton/posture and clothing information) in Table 14. We also compared our method to two other approaches, DEX [53] and
SSR [14] in Table 15. As can be seen, our approach with the combination of all features

(FK+FPP+SCIP+SCK) achieved a top-1 score of 97.80% which is significantly higher than those scores for other methods.

Table 14: The impact of facial/SC graphs and their keypoints/patches of pixels on the age estimation performance on the Relative Human dataset.

Modules	FK+FPP	SCIP	SCK	SCK+SCIP
Modality	face	clothing	skeleton/posture	clothing+skeleton/posture
Top-1 score	94.35%	91.85%	91.10%	91.98%

Table 15: Comparison of our proposed method to other approaches on the Relative Human dataset.

Method	Ours (FK+FPP+SCIP+SCK)	DEX [53]	SSR [14]
Top-1 score	<b>97.80</b> %	90.03%	92.31%



Figure 10: Some image samples form the Relative Human dataset with a variety of partial facial views captured in-the-wild that also resembles surveillance conditions.

# 4.5. Blurring Effects and Facial Expressions

To simulate the reduced-quality image capture in-the-wild or under surveillance conditions, following [61] and [62], we evaluated our proposed method under several blurring effects. We tested our algorithm on the PAL dataset [44] which includes 3000 image samples with a variety of facial expressions. We synthetically added Gaussian and motion blurring effects to the image samples of the dataset to emulate those effects in real-life scenarios. Some image examples from the PAL dataset with three facial expressions, "sad", "surprise", and "happy" and synthetically added blurring effects are shown in Fig. 11. We compared our method with two other strategies and showed the results in Table 16. As can be seen, our method remarkably outperforms the other approaches which shows the reliability of our proposed approach under blurring effects which is a common condition in-the-wild or in surveillance scenarios.

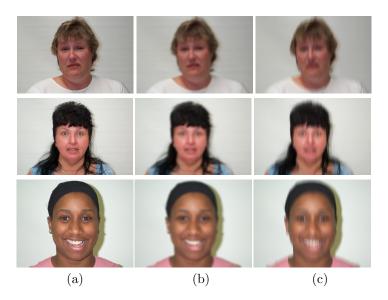


Figure 11: Some image samples form the PAL dataset with two blurring effects under different facial expressions. (a): Original image, (b): added Gaussian blurring effect, and (c): added motion blurring effect. The facial expressions are: (up): sad, (middle): surprise, and (down): happy.

Table 16: Comparison of our proposed method to other approaches on the PAL dataset under different blurring effects that simulates those effects in-the-wild or under surveillance conditions.

Method	MAE (Gaussian Blur)	MAE (Motion Blur)
[61]	6.42	6.48
[62]	6.0	6.0
Ours	3.49	3.70

#### 5. Conclusions

We proposed a new graph representation for age estimation using skeleton joints and facial keypoints. This new representation yields more relevant information than raw images and is more reliable under different viewpoints and facial expressions. We also suggested a new Temporally-Aware Adaptive Graph Convolutional Network with two improvements (1) it captures non-ordinal temporal dependencies in different ages which is not possible with standard temporal networks, and (2) it adaptively refines facial and Skeletal-cosmetic graphs edges to accommodate the variance in appearance. Furthermore, we proposed to use skeleton structure, posture, and clothing information in the age estimation solution. This rich set of features accommodates significant performance improvements when the face is only partially visible in real-life scenarios. Our method outperformed the state-of-the-art approaches on four public benchmarks, including the UTKFace dataset, whose images are captured in-the-wild. We also further tested the reliability of our proposed age estimation algorithm in uncontrolled environments in two more scenarios: (1) the images captured in-the-wild from the Relative Human dataset and (2) the synthetically blurred images from the PAL dataset under a variety of facial expressions.

Our new graph representation of soft biometrics, including the skeleton, posture, clothing, and face, can be used as a backbone by other researchers in the field. Such a graph structure is a powerful representation of the skeleton and face in tandem, since graph nodes and edges can effectively describe facial landmarks, and skeleton joints and encode their connectivity information. We also are the first to introduce a complete experimental setup for age estimation in-the-wild. Such an experimental setup can be used as a standard benchmark in the future. Moreover, our new Temporal Memory Module (TMM) can be exploited in any research problem, such as age estimation that involves computing non-ordinal temporal dependencies.

# 5.1. Limitations

Here is our work's main limitations:

• Severe occlusion: While our method is effective when the human face and body are partially occluded, it can fail when both the face and body are severely occluded. However, such a severe occlusion is unlikely.

- Non-conventional facial expressions: Our facial keypoint selection algorithm can accommodate a variety of facial expressions. Nevertheless, our proposed algorithm might be less effective in handling non-conventional facial expressions.
- Complex human poses: Although our Adaptive Graph Convolutional Layer (AGCL) can accommodate diverse standard and non-standard human poses, it might be less effective in handling extremely complex human poses in sports activities such as gymnastics or martial arts. It is because, in many of these activities, the skeleton configuration of human changes significantly. So, the neighboring skeleton joints can highly vary in such sport activities.

#### 5.2. Future Work

We suggest some future work to solve the limitations explained above:

- Modeling human sub-parts: Although our graph-based model can model several human face and body sub-parts, such as eyes and noise, the network's loss function does not depend independently on human sub-parts. We suggest modeling human sub-parts independently with separate loss functions to handle severe face and human body occlusion. Such independent human sub-parts modeling is more effective when only a sub-part of the human face or body is visible. This can be achieved, i.e., by designing multiple graph convolutional streams for several human sub-parts.
- Adaptive facial keypoint selection: Although our facial keypoint selection algorithm is data-driven, it is not based on a learning process. To accommodate a variety of cross-cultural and non-conventional facial expressions, we recommend a more adaptive model, preferably using a separate deep learning module to select more effective facial keypoints.
- Adaptive skeletal-cosmetic keypoint generation: Our skeletal-cosmetic keypoint generation algorithm selects fixed and spatially consistent keypoints based on human skeleton structure. We suggest an adaptive way to generate such keypoints to handle various complex human poses.

### Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 2000487 and the Robertson Foundation under Grant No. 9909875. Any opinions,

findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funders.

#### References

- [1] N. Mansouri, Automatic age estimation: A survey, Computación y Sistemas 24 (2) (2020) 883–895.
- [2] K.-H. Liu, T.-J. Liu, A structure-based human facial age estimation framework under a constrained condition, IEEE Transactions on Image Processing 28 (10) (2019) 5187– 5200.
- [3] H. Sun, H. Pan, H. Han, S. Shan, Deep conditional distribution learning for age estimation, IEEE Transactions on Information Forensics and Security 16 (2021) 4679–4690.
- [4] A. Singh, D. Patil, M. Reddy, S. Omkar, Disguised face identification (dfi) with facial keypoints using spatial fusion convolutional network, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2017, pp. 1648–1655.
- [5] Y. H. Kwon, N. da Vitoria Lobo, Locating facial features for age classification, in: Intelligent robots and computer vision XII: Algorithms and techniques, Vol. 2055, SPIE, 1993, pp. 62–72.
- [6] M. Korban, X. Li, Ddgcn: A dynamic directed graph convolutional network for action recognition, in: European Conference on Computer Vision, Springer, 2020, pp. 761–776.
- [7] Z. Lou, F. Alnajar, J. M. Alvarez, N. Hu, T. Gevers, Expression-invariant age estimation using structured learning, IEEE transactions on pattern analysis and machine intelligence 40 (2) (2017) 365–375.
- [8] O. Langner, R. Dotsch, G. Bijlstra, D. H. Wigboldus, S. T. Hawk, A. Van Knippenberg, Presentation and validation of the radboud faces database, Cognition and emotion 24 (8) (2010) 1377–1388.
- [9] D. T. Nguyen, S. R. Cho, T. D. Pham, K. R. Park, Human age estimation method robust to camera sensor and/or face movement, Sensors 15 (9) (2015) 21898–21930.

- [10] T. Wu, P. Turaga, R. Chellappa, Age estimation and face verification across aging using landmarks, IEEE Transactions on Information Forensics and Security 7 (6) (2012) 1780–1788.
- [11] M. Korban, S. Singh, P. Youngs, G. S. Watson, S. T. Acton, Ai-assisted activity detection in k-6 classroom environments: A preliminary framework to assist in pedagogical performance evaluation, in: 2021 55th Asilomar Conference on Signals, Systems, and Computers, IEEE, 2021, pp. 1136–1140.
- [12] Y. Sun, W. Liu, Q. Bao, Y. Fu, T. Mei, M. J. Black, Putting people in their place: Monocular regression of 3d people in depth, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 13243–13252.
- [13] U. of Virginia, AIAI Project, https://aiaiproject.weebly.com/, [Online; accessed 16-August-2022] (2022).
- [14] T.-Y. Yang, Y.-H. Huang, Y.-Y. Lin, P.-C. Hsiu, Y.-Y. Chuang, Ssr-net: A compact soft stagewise regression network for age estimation., in: IJCAI, Vol. 5, 2018, p. 7.
- [15] S. Chen, C. Zhang, M. Dong, J. Le, M. Rao, Using ranking-cnn for age estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5183–5192.
- [16] Z. Tan, J. Wan, Z. Lei, R. Zhi, G. Guo, S. Z. Li, Efficient group-n encoding and decoding for facial age estimation, IEEE transactions on pattern analysis and machine intelligence 40 (11) (2017) 2610–2623.
- [17] M. M. Dehshibi, J. Shanbehzadeh, Cubic norm and kernel-based bi-directional pca: toward age-aware facial kinship verification, The Visual Computer 35 (1) (2019) 23–40.
- [18] Y. H. Kwon, da Vitoria Lobo, Age classification from facial images, in: 1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 1994, pp. 762–767.
- [19] C.-C. Ng, M. H. Yap, Y.-T. Cheng, G.-S. Hsu, Hybrid ageing patterns for face age estimation, Image and Vision Computing 69 (2018) 92–102.

- [20] G.-S. J. Hsu, Y.-T. Cheng, C. C. Ng, M. H. Yap, Component biologically inspired features with moving segmentation for age estimation, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, 2017, pp. 540–547.
- [21] E. N. A. Hammond, S. Zhou, H. Cheng, Q. Liu, Improving juvenile age estimation based on facial landmark points and gravity moment, Applied Sciences 10 (18) (2020) 6227.
- [22] K.-Y. Chang, C.-S. Chen, A learning framework for age rank estimation based on face images with scattering transform, IEEE Transactions on Image Processing 24 (3) (2015) 785–798.
- [23] P. Thukral, K. Mitra, R. Chellappa, A hierarchical approach for human age estimation, in: 2012 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, 2012, pp. 1529–1532.
- [24] Z. Hu, Y. Wen, J. Wang, M. Wang, R. Hong, S. Yan, Facial age estimation with age difference, IEEE Transactions on Image Processing 26 (7) (2016) 3087–3097.
- [25] K. Li, J. Xing, W. Hu, S. J. Maybank, D2c: Deep cumulatively and comparatively learning for human age estimation, Pattern Recognition 66 (2017) 95–105.
- [26] P. Li, Y. Hu, X. Wu, R. He, Z. Sun, Deep label refinement for age estimation, Pattern Recognition 100 (2020) 107178.
- [27] Z. Deng, H. Liu, Y. Wang, C. Wang, Z. Yu, X. Sun, Pml: Progressive margin loss for long-tailed age classification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 10503–10512.
- [28] S. Taheri, Ö. Toygar, On the use of dag-cnn architecture for age estimation with multistage features fusion, Neurocomputing 329 (2019) 300–310.
- [29] Z. Tan, Y. Yang, J. Wan, G. Guo, S. Z. Li, Deeply-learned hybrid representations for facial age estimation., in: IJCAI, 2019, pp. 3548–3554.
- [30] H. Wang, V. Sanchez, C.-T. Li, Improving face-based age estimation with attention-based dynamic patch fusion, IEEE Transactions on Image Processing 31 (2022) 1084–1096.

- [31] J.-C. Xie, C.-M. Pun, Deep and ordinal ensemble learning for human age estimation from facial images, IEEE Transactions on Information Forensics and Security 15 (2020) 2361–2374.
- [32] F. Alonso-Fernandez, K. H. Diaz, S. Ramis, F. J. Perales, J. Bigun, Soft-biometrics estimation in the era of facial masks, in: 2020 International Conference of the Biometrics Special Interest Group (BIOSIG), IEEE, 2020, pp. 1–6.
- [33] L. Shi, Y. Zhang, J. Cheng, H. Lu, Action recognition via pose-based graph convolutional networks with intermediate dense supervision, Pattern Recognition 121 (2022) 108170.
- [34] H. Jiang, P. Cao, M. Xu, J. Yang, O. Zaiane, Hi-gcn: A hierarchical graph convolution network for graph embedding learning of brain network and brain disorders prediction, Computers in Biology and Medicine 127 (2020) 104096.
- [35] U. Chaudhuri, B. Banerjee, A. Bhattacharya, Siamese graph convolutional network for content based remote sensing image retrieval, Computer vision and image understanding 184 (2019) 22–30.
- [36] Y. Wu, G.-D. He, L.-H. Wen, X. Qin, C.-A. Yuan, V. Gribova, V. F. Filaretov, D.-S. Huang, Discriminative local representation learning for cross-modality visible-thermal person re-identification, IEEE Transactions on Biometrics, Behavior, and Identity Science.
- [37] J. Yu, H. Yin, J. Li, M. Gao, Z. Huang, L. Cui, Enhance social recommendation with adversarial graph convolutional networks, IEEE Transactions on Knowledge and Data Engineering.
- [38] A. Qin, Z. Shang, J. Tian, Y. Wang, T. Zhang, Y. Y. Tang, Spectral–spatial graph convolutional networks for semisupervised hyperspectral image classification, IEEE Geoscience and Remote Sensing Letters 16 (2) (2018) 241–245.
- [39] L. Zhao, Y. Song, C. Zhang, Y. Liu, P. Wang, T. Lin, M. Deng, H. Li, T-gcn: A temporal graph convolutional network for traffic prediction, IEEE Transactions on Intelligent Transportation Systems 21 (9) (2019) 3848–3858.

- [40] C. Ding, S. Wen, W. Ding, K. Liu, E. Belyaev, Temporal segment graph convolutional networks for skeleton-based action recognition, Engineering Applications of Artificial Intelligence 110 (2022) 104675.
- [41] A. Bulat, G. Tzimiropoulos, How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks), in: International Conference on Computer Vision, 2017.
- [42] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, Y. Sheikh, Openpose: realtime multi-person 2d pose estimation using part affinity fields, IEEE transactions on pattern analysis and machine intelligence 43 (1) (2019) 172–186.
- [43] Z. Zhang, Y. Song, H. Qi, Age progression/regression by conditional adversarial autoencoder, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 5810–5818.
- [44] M. Minear, D. Park, A lifespan database of adult facial stimuli. behaviour research methods, instruments, & computers, 36, 630-633 (2004).
- [45] B.-C. Chen, C.-S. Chen, W. H. Hsu, Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset, IEEE Transactions on Multimedia 17 (6) (2015) 804–815.
- [46] Y. Fu, T. M. Hospedales, T. Xiang, Y. Yao, S. Gong, Interestingness prediction by robust learning to rank, in: ECCV, 2014.
- [47] Z. Niu, M. Zhou, L. Wang, X. Gao, G. Hua, Ordinal regression with multiple output cnn for age estimation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 4920–4928.
- [48] C. Zhang, S. Liu, X. Xu, C. Zhu, C3ae: Exploring the limits of compact model for age estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 12587–12596.
- [49] W. Cao, V. Mirjalili, S. Raschka, Rank consistent ordinal regression for neural networks with application to age estimation, Pattern Recognition Letters 140 (2020) 325–331.

- [50] H. Pan, H. Han, S. Shan, X. Chen, Mean-variance loss for deep age estimation from a face, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 5285–5294.
- [51] W. Shen, K. Zhao, Y. Guo, A. L. Yuille, Label distribution learning forests, Advances in neural information processing systems 30.
- [52] W. Shen, Y. Guo, Y. Wang, K. Zhao, B. Wang, A. L. Yuille, Deep regression forests for age estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2304–2313.
- [53] R. Rothe, R. Timofte, L. Van Gool, Deep expectation of real and apparent age from a single image without facial landmarks, International Journal of Computer Vision 126 (2) (2018) 144–157.
- [54] M. Yoshimura, S. Ogata, Foss: Multi-person age estimation with focusing on objects and still seeing surroundings, arXiv preprint arXiv:2010.07544.
- [55] A. Al-Shannaq, L. Elrefaei, Age estimation using specific domain transfer learning, Jordanian Journal of Computers and Information Technology (JJCIT) 6 (2) (2020) 122–139.
- [56] Q. Li, Y. Liu, Z. Sun, Age progression and regression with spatial attention modules, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 11378–11385.
- [57] A. Berg, M. Oskarsson, M. O'Connor, Deep ordinal regression with label diversity, in: 2020 25th International Conference on Pattern Recognition (ICPR), IEEE, 2021, pp. 2740–2747.
- [58] H. Han, C. Otto, X. Liu, A. K. Jain, Demographic estimation from face images: Human vs. machine performance, IEEE transactions on pattern analysis and machine intelligence 37 (6) (2014) 1148–1161.
- [59] J. Lu, V. E. Liong, J. Zhou, Cost-sensitive local binary feature learning for facial age estimation, IEEE Transactions on Image Processing 24 (12) (2015) 5356–5368.

- [60] H. Liu, J. Lu, J. Feng, J. Zhou, Group-aware deep feature learning for facial age estimation, Pattern Recognition 66 (2017) 82–94.
- [61] D. T. Nguyen, S. R. Cho, K. R. Park, Age estimation-based soft biometrics considering optical blurring based on symmetrical sub-blocks for mlbp, Symmetry 7 (4) (2015) 1882–1913.
- [62] J. S. Kang, C. S. Kim, Y. W. Lee, S. W. Cho, K. R. Park, Age estimation robust to optical and motion blurring by deep residual cnn, Symmetry 10 (4) (2018) 108.



Matthew Korban received his BSc and MSc degree in Electrical Engineering in 2013 from the University of Guilan, where he worked on sign language recognition in video. He received his PhD in Computer Engineering from Louisiana State University. He is currently a

Postdoc Research Associate at the University of Virginia, working with Prof. Scott T. Acton. His research interest includes Human Action Recognition, Early Action Recognition, Motion Synthesis, and Human Geometric Modeling in Virtual Reality environments.



Peter Youngs is a professor in the Department of Curriculum, Instruction and Special Education at the University of Virginia. He studies how neural networks can be used to automatically classify instructional activities in videos of elementary mathematics and

reading instruction. He currently serves as co-editor of American Educational Research Journal.



Scott T. Acton received his M.S. and Ph.D. degrees at the University of Texas at Austin. He received his B.S. degree at Virginia Tech. Professor Acton is a Fellow of the IEEE "for contributions to biomedical image analysis." Currently, Acton is a program director in the Computer and Information Science and Engineering at

the U.S. National Science Foundation. He is also professor of Electrical and Computer Engineering and of Biomedical Engineering at the University of Virginia. Professor Acton's laboratory at UVA is called VIVA - Virginia Image and Video Analysis. They specialize in bioimage analysis problems. Professor Acton has over 300 publications in the image analysis area including the books Biomedical Image Analysis: Tracking and Biomedical Image Analysis: Segmentation. He was the 2018 Co-Chair of the IEEE International Symposium on Biomedical Imaging. Professor Acton was recently Editor-in-Chief of the IEEE Transactions on Image Processing (2014-2018).