Abstract—Can we construct an explainable face recognition network able to learn a facial part-based feature like eyes, nose, mouth and so forth, without any manual annotation or additional datasets? In this paper, we propose a generic Explainable Channel Loss (ECLoss) to construct an explainable face recognition network. The explainable network trained with ECLoss can easily learn the facial part-based representation on the target convolutional layer, where an individual channel can detect a certain face part. Our experiments on dozens of datasets show that ECLoss achieves superior explainability metrics, and at the same time improves the performance of face verification without face alignment. In addition, our visualization results also illustrate the effectiveness of the proposed ECLoss.

I. INTRODUCTION

Face recognition has made significant achievement in the past ten years [18], due to the development of deep learning, large-scale labeled faces datasets and so forth, and has been extensively used in many applications [9]. Despite of the great success, most of these face recognition algorithms are still black-box.

What face recognition networks have learned? It is commonly believed that the shallow layers are more likely to capture low-level information like colors, edges, and textures, while deep layers tend to describe high-level features [29]. However, such observation is too general to understand a face recognition network.

A typical approach of previous studies for this problem is to use salience map to understand a network when it makes a decision [23], [15]. However, most salience map based methods are not suitable for face recognition, as the areas of maximum activation are focused on the whole face as shown in Fig. 1, which does not provide effective evidence and semantic information for explainability. Prior works have demonstrated that the representation extracted by deep networks may be mixed with high-dimensional abstract features that humans cannot understand.

One potential solution for this issue is to guide the networks to learn a structural representation [27]. Yin et al. [27] suggested that in explainable face recognition network, each channel shall be able to represent a face part, and proposed a spatial activation diversity loss to learn an explainable representation. However, in order to learn explainable representation patterns, occluded faces are required for training. This additional step undoubtedly increases the difficulty of training. How can we find out a model-agnostic method to construct an explainable face recognition network without any additional occlusion face dataset?

Inspired by previous studies on the explainability and interpretability of deep learning [30], [27], [2], [14], we proposed a generic Explainable Channel Loss (ECLoss) to construct an explainable face recognition network. Fig. 1 shows the difference in feature maps among different methods. By adopting ECLoss, the recognition network can stand a more readable representation for its decision behavior, which draws a remarkable progress over the existing explainable solutions. As shown in Fig. 2, a channel of the explainable face recognition network trained with ECLoss can learn semantic parts like eyes, mouth, nose, etc. To the best of our knowledge, this is the first method to construct a feature level explainable face recognition network that does not require any additional dataset or manual annotation. In this way, the explainable network is able to effectively demonstrate what has been learned.

In conclusion, the proposed Explainable Channel Loss (ECLoss) has the following properties: (1) Networks trained with ECLoss can explain what is learned inside the network. As shown in Fig. 5, the visualization results show that different channels learned different semantic parts. (2) As shown in Fig. 2 and Fig. 5, a single channel within the explainable face recognition network trained with ECLoss can learn different semantic parts like eyes, mouth, nose, etc. (3) Compared to [30], the training of ECLoss in the network is simple and effective, and does not require any manual annotation, i.e., labeling of nose, mouth, or other parts are not necessary. (4) As a model-agnostic method, ECLoss can be applied.
Fig. 2. The training pipeline of explainable network with the proposed ECLoss. (a) The activation templates are used for ECLoss to force the representations of the target convolutional layer to contain maximum two activation peaks. (b) Compared with the network trained without ECLoss, the explainable network trained with ECLoss can obtain more explainable representation on the target convolution layer. (c) The pipeline of training: The proposed ECLoss used mutual information, to force the target convolutional layer to represent the patterns of the activation templates.

II. RELATED WORK

A. Face Recognition

Face recognition has been studied for several decades. Before deep learning, many previous studies are based on feature engineering [4]. Due to the rapid development of deep learning, face recognition has made a huge achievement [18]. Although these previous methods have reached significant performance in face recognition, the deep networks still performs in a back-box manner, as we don’t know what information the networks have learned.

B. Explainability & Interpretability

The explainability and the discrimination power are two crucial properties of a deep network [2], [30]. At present, there is a trade-off between high explainability and strong discrimination power. Trough deep neural networks, high discrimination ability can be achieved but lack of the explainability [30].

Post-hoc Explanation via Visualization. Post-hoc explanation aims to provide an understanding of what knowledge has been learned by networks, in an intuitive manner of humans. Visualization is the most direct and intuitive way to explore the pattern hidden inside the deep network. Specifically, in face recognition, [33] shows a method to find what difference between two similar-looking faces by a pair-wise patch-by-patch occlusion method, but it does not indicate what knowledge has been acquired by networks. [28] applied a CAMs method to generate heatmaps to indicate the active regions of different faces from different people; the results showed that the active regions tend to cover a whole face, so it does not provide useful information.

These post-hoc visualization methods have been successfully used in various fields of computer vision, like object classification [15], attacking [31], biometrics [21] and medical image processing [20], but they are difficult to generalize to face recognition. Therefore, model-intrinsic explanation methods are usually considered in face recognition.

Model-intrinsic Explanation via Visualization. Model-intrinsic explanations are often derived from intrinsic, transparent, or white-box models [13], which has simple structures, such as decision trees or linear models. In this paper, we mainly focus on the methods to improve the explainability of network. These models have better partial explainability, to some extent, than traditional models. [27] proposed a spatial diversity loss and a feature diversity loss to enhance explainable representations, which reach an outstanding performance on three face recognition benchmarks, but it requires occluded faces before training. [12] proposed a novel similarity metric, named explainable cosine (xCos), that can be plugged into various traditional networks to provide meaningful explanations on face recognition. Although it obtained comparable results, it did not reveal the knowledge
learned inside the network.

In this study, we aim to constrain the deep network to learn an explainable facial part-based structural representation, without requirement for any manual annotation or occluded faces.

III. METHOD

In this study, our goal is to guide the network to learn a facial part-based structural representation. Formally, we need to match the activation peaks of feature maps with the face part of the input face. In object detection and classification, the object always occupies adjacent pixels in the picture, but in face recognition, the face part, such as two eyes may be separated. To go a step further, we proposed a loss function to guide each channel of the feature map to contain maximum two activation peaks, and make an individual channel learn the same semantic parts.

As shown in Fig. 2-(a), our activation templates contain all possible variations of activation patterns in the feature maps, which have zero, one, or two activation peaks. As we know, the representation of initial deep learning networks without training is random. Even for a well trained network, its representation is usually mixed with high-dimensional abstract features that humans cannot understand.

In this study, the task of the proposed ECLoss is to match the feature maps of the target layer within the network with a subset of activation templates $\mathcal{T}$. Therefore, we extend the definitions of mutual information in information theory from variable to sets of variables, which is a measure of the mutual dependence between the two sets. Our proposed ECLoss enables the network to learn the representation of activation templates by increasing the mutual information between activation templates and feature maps of the network. As shown in Fig. 2-(c), our proposed ECLoss is applied to the last convolutional layer of the network to increase the mutual information $\text{MI}(\mathcal{T};\mathcal{X})$ between activation templates $\mathcal{T}$ and feature maps $\mathcal{X}$. As a result, after training, the patterns of the feature maps $\mathcal{X}$ will match the activation templates $\mathcal{T}$.

The proposed ECLoss is a generic and model-agnostic method that can be used for various face recognition networks. For example, Fig. 2-(c) shows that our generic channel loss function works on the last layer of the encoder, and can be used with other loss functions. Formally, in this work, our total loss function is defined as:

$$\text{Total Loss} = \alpha \text{Loss}(y, \hat{y}) + \beta \text{ECLoss}$$  \hspace{1cm} (1)

where $\text{Loss}(y, \hat{y})$ indicates the loss between true subject ID label and the prediction, the $\alpha$ and $\beta$ are weighting coefficients.

A. Activation Templates

Given the face images $I$ with the size $s \times s$ as the input, the face recognition network with parameter $\theta$ can represent $I$ as the feature maps $\mathcal{X}$ from the target convolutional layer after ReLU operation. In particular, $\mathcal{X} = \{x_1, x_2, \ldots, x_t, \ldots, x_r\}$, where $x_i$ refers to the $i$-th channel of $\mathcal{X}$.

Activation templates with one peak. For an unknown face, the face parts may appear in any pixel of the input face image, i.e., any pixel of the feature map may be activated if there is a face part of interest in the receptive field, as shown in Fig. 4.

Let $T^{(i,j)}$ denotes the activation template containing an activation peak in coordinate $(i, j)$, and $T_{(u,v)}^{(i,j)}$ indicates the value in coordinate $(u, v)$ of the activation template $T^{(i,j)}$. For each activation template, the value of the activation peak is $\tau$ and the value of other pixels extend outward the peak is defined as:

$$T_{(u,v)}^{(i,j)} = \tau \cdot \max(1 - \frac{\|v(u,v) - (i,j)\|_2}{r}, -1)$$  \hspace{1cm} (2)

where the $\tau$ is a constant greater than 0, $\|\cdot\|_2$ denote L2-Norm distance, $r$ is a constant that denotes the radius of the area greater than 0. To go a step further, the one-peak activation template set $\mathcal{T}_1$ is defined as $\mathcal{T}_1 = \left\{ T^{(i,j)} \mid 1 \leq i, j \leq s \right\}$.

In addition, a negative template $T^-$ is defined to represent a face map without any activation peaks, $T^-_{(i,j)} = -\tau$ for every coordinate $(i,j)$ within the template $T^-$. 

Activation templates with two peaks. In order to match activation peaks with face parts, we allow a feature map to contain at most two activation peaks. In this situation, we define the two-peak templates to be the combination of any two one-peak templates, as shown in Fig. 2. The combination $T^{(i,j)} \cup T^{(i,2)}$ of two one-peak templates $T^{(i,j)}$ and $T^{(i,2)}$ is defined as $T_{(u,v)} = \max(T_{(u,v)}^{(i,j)}, T_{(u,v)}^{(i,2)})$, where $T_{(u,v)}$ is a
two-peak template combining two one-peak template $T^{(i_1,j_1)}_{(u,v)}$ and $T^{(i_2,j_2)}_{(u,v)}$. Then, the two-peak template set $T_2$ is defined as:

$$T_2 = \left\{ T^{(i_1,j_1)} \cup T^{(i_2,j_2)} \mid 1 \leq i_1,i_2,j_1,j_2 \leq s \right\} (3)$$

As shown in Fig. 2, a set of activation templates $T$ used for ECLoss contain all of the one-peak templates, two-peak templates and a negative template $T^-$, $T = T_1 \cup T_2 \cup \{ T^- \}$.

### B. Explainable Channel Loss

In this study, our major goal is to make each dimension of the representation tend to represent a face part. Generally, before training, the feature maps is a random output when the input face images go through the network, so the feature maps $X$ and the locations of all face parts are unrelated. In this case, after network training, each dimension of the feature map belonging to $X$ will be able to match activation templates $T$.

**Problem Definition:** Given a template set of activation pattern $T$, and a feature map $X$ of the target convolutional layer from the face recognition network, the task of ECLoss is to find parameters $\theta$ such that feature map $X$ maximally represent a subset of the face part set $X$.

By extending the definitions of mutual information in information theory from variable to sets of variables [5], [10], the mutual information between $X$ and $T$ is defined as:

$$MI(T;X) = \sum_{T \in T} \sum_{x \in X} P(T,x) \log \left( \frac{P(T,x)}{P(T)P(x)} \right)$$

$$= \sum_{T \in T} \sum_{x \in X} P(x)P(T|x) \log \left( \frac{P(T|x)}{P(T)} \right) (4)$$

where $T(\mu)$ represents the prior probability of a activation template from $T$. $P(x)$ denotes the probability of appearance of feature maps of different patterns, $P(x) = \sum_{T \in T} P(T)P(x|T)$.

To simplify the problem, we assume that all templates are uniformly distributed, so the $P(T)$ is given as a constant of $\frac{1}{|T|}$. $P(T|x)$ is the conditional likelihood denoting the fitness between $T$ and $x$:

$$P(T|x) = \frac{e^{\text{tr}(x^*T)}}{\sum_{T \in T} e^{\text{tr}(x^*T)}} (5)$$

where $\cdot$ indicates hadamard product (element-wise product), $tr(\cdot)$ indicates the trace of a matrix, so $tr(x^*T)$ can be rewritten as $tr(x^*T) = \sum_{i,j} x_i T_{ij}$.

To increase the mutual information between $T$ and $X$, ECLoss is defined as $ECLoss = -MI(T;X)$. Then, the total loss function is defined as $\text{Total Loss} = \alpha \text{Loss}(y, \hat{y}) - \beta \text{MI}(T;X)$.

### C. Visualization of Structural Representation

The feature map $x$ is a $n \times n$ matrix encoded by the face recognition network. In order to ensure the same resolution of $x$ as the input face, feature map $x$ will be resized to $s \times s$ by bi-linear upsampling.

To focus on the area of maximum activation, the area of top 10% intensity will be highlighted, while other area will be discarded. A threshold $t$ is found to highlight the area of top 10% intensity in the feature map $x$:

$$\hat{x}_{(i,j)} = \begin{cases} x_{(i,j)}, & x_{(i,j)} \geq t \\ 0, & \text{otherwise} \end{cases} (6)$$

Then the threshold $t$ is defined as:

$$\hat{t} \equiv \arg \max \{ P(\hat{x}_{(i,j)} > t) \geq 0.1 \} (7)$$

where $P$ is the probability when the condition is true. Some examples of visualization for top 10% activation are shown on Fig. 5, which are the result of transforming $\hat{x}$ into the heatmap style and superimposing it on the original face image.
Fig. 6. Comparison of the feature map between networks with and without ECLoss. For the same input face, the feature map of the explainable network trained with ECLoss has at most two activation peaks, and usually activates a certain face part. Instead, the feature map of the original network has more noises, and the activation peaks are located randomly on different parts.

### TABLE I

| Stage | Dataset | Landmark | Pose | Image / Pairs |
|-------|---------|----------|------|---------------|
| Training | Glint360K [1] | - | Near-frontal | 17,091,657 / - |
| | LFW [8] | - | Near-frontal | 17,091,657 / - |
| | CFP-FF [24] | - | Near-frontal | 17,091,657 / - |
| | CALFW [32] | - | Near-frontal | 17,091,657 / - |
| | VGGFace [18] | - | Near-frontal | 17,091,657 / - |
| Testing | AgeDB [17] | 68 | Near-frontal | 17,091,657 / - |
| | 300-W [22] | 68 | Near-frontal | 17,091,657 / - |
| | AFW [35] | 68 | Near-frontal | 17,091,657 / - |
| | Helen [11] | 68 | Near-frontal | 17,091,657 / - |
| | LaPa [16] | 106 | Near-frontal | 17,091,657 / - |
| | LFPW [3] | 68 | Near-frontal | 17,091,657 / - |

### IV. EXPERIMENT

**A. Dataset & Pre-processing**

**Training Dataset** The Glint360K [1] is a large scale dataset containing 17 million face images from 360K individuals. All face images are downloaded from Internet and then pre-processed to size of 112 × 112 [6].

**Testing Dataset** For testing robustness of face recognition, we explore various face verification datasets: LFW [8], CFP-FF [24] and CALFW [32]. Besides, in this study, many explainability metrics are assessed using several datasets, such as VGGFace[18], AgeDB [17], Helen [11], 300-W[22], LaPa [16], AFW [35] and LFPW [3]. All datasets used in training and testing are listed in Table I.

**Pre-processing** For face recognition, face alignment is an essential prerequisite, which can align facial components, e.g., eye, nose, mouth, and contour to a standard facial space. In this study, in order to reduce the influence of the structural representation by the standard facial space, the data used for training and testing are not aligned. All face images used for training and testing are detected and cropped by MTCNN [26], and resized to 224 × 224.

**B. Implementation Details**

We trained three explainable networks based on VGG13 [25], VGG16 [25], and VGG19 [25] for face recognition in the Glint-360K dataset. And original VGG13, VGG16, and VGG19 were trained on the same dataset and settings for comparison. As the skip connections of the residual networks mix different pattern of multiple layers in a single feature map [7], we do not use residual networks and skip connection in this study. In these experiments, the input face images are detected by MTCNN and resized to 224 × 224, without any face alignment.

For activation templates, we found in our experiments that the results are not sensitive to the value of \( \tau \). As ECloss might overflow when the value of \( \tau \) is big, we set \( \tau \) to a small number, i.e. \( t = 0.001 \), and \( r = 4 \). For total loss function, we set the \( \alpha \) as a constant of 1, and \( \beta \) is set as \( 1e^{-5} \) initially and then automatically changed by the program to reduce the total loss.

In the VGG architecture, the shape of last convolutional layer is \( 14 \times 14 \), i.e., for each channel, tens of thousands of activation templates can be generated (nearly \( 14^2 \)-peak templates, \( 14^2 \) one-peak templates and 1 negative template). To reduce the complexity of calculation, we take 400 templates evenly from original template set \( \mathbb{T}^{tw} \) as a new set \( \mathbb{T} \) used for all experiments.

For mutual information MI(\( T; X \)) between \( T \) and \( X \), we assume that all templates of \( \mathbb{T} \) are uniformly distributed, so we set \( |\mathbb{T}| = 400 \), \( P(T) = \frac{1}{|\mathbb{T}|} = 0.0025 \). Because \( P(x) = \sum_{T} P(T)P(x|T) \), \( T \in \mathbb{T} \) is computed using numerous prior feature map requiring large amounts of memory. We approximate \( P(x) \) as \( \sum_{T} P(T)E_{t} \left[ \frac{e^{\beta(x|T)}}{\sum_{x}e^{\beta(x|T)}T} \right] \), where \( E_{t}[\cdot] \) indicates the mathematic expectation, \( tr(\cdot) \) indicates the trace of a matrix, and \( P(T) = 0.0025 \).

All experiments are conducted on a workstation with eight NVIDIA Tesla V100 GPUs. Moreover, the learning rate is set to 0.03 and the batch size is set to 64.

### V. RESULT & ANALYSIS

**A. Visualization of Activation**

As mentioned in the section III-C, we visualized the receptive fields of activation of a channel. Previous studies
in [2], [34] and [30] have introduced this method to compute the receptive fields on a given feature map.

For both networks trained with and without ECLoss, we visualize the top 10% activation values of the last convolutional layer. As shown in Fig. 5, the maximal activation area in the network trained with ECLoss can focus on a certain facial part, like face, cheek, nose, etc., but the original network randomly focuses on some unexplainable semantic information. Moreover, for explainable networks trained with ECLoss, these clear disentanglements of structural part-based representation can easily help people to know what knowledge the network has learned, which is helpful for quantifying the contribution of different face parts to the network.

We found that, within the explainable face recognition networks trained with ECLoss, an individual layer usually detected a certain face part even if the input face images are from different people. For example, as shown in Fig. 5, the 2nd channel detected the mouth, the 122nd channel detected the left eye, the 462nd channel detected the eyes, etc. In addition, it not only detected face components, but also learned some patterns that make sense to the network. For example, the 124th channel learned the area to the right of the right eye. Further, we also compare the feature map of the explainable network and the original network in Fig. 6. Given the same face as input, our explainable networks usually have at most two activation peaks in a feature map, and the activation peaks tend to activate a certain face part. Instead, the activation peaks of the feature map of the original network usually contain a lot of noises, and are randomly located on different areas.

B. Quantitative Analysis

1) Face Verification: Face verification mainly computes the similarity between a reference and a probe face image to decide if these face images are of the same identity. In the evaluation of face verification, LFW, CFP-FF and CALFW datasets are used to assess performance of the explainable face recognition networks and the original networks.

In all experiments, we did not perform face alignment. As shown in Tab. 2, compared to the original networks, the explainable network with the proposed ECLoss attained a significant improvement.

2) Part Explainability: This evaluation metric of part explainability was proposed by Bau et al. [2] and Zhang et al. [30], which measures the object-part explainability of channels from a network.

For assessing part explainability of various network, five face parsing datasets of LaPa, AFW, Helen, LFPW and 300-W datasets are used. Before evaluation of part explainability, as shown in Fig. 7, we generate facial component masks for these face datasets by their landmarks, which contain hair, forehead, cheek, eyes, nose, mouth, and jaw. Formally, a facial components set is defined as $K = \{\text{hair, forehead, cheek, etc.}\}$.

For assessing part explainability, the feature map $x$ is first resized to $s \times s$ by bi-linear upsampling, where $s$ is the size of the input image. A threshold $t$ is then applied to find out the top 10% activation area using Equation (7). Then, the fitness of feature map $x$ regarding to the facial component $k \in K$ is computed by

$$\text{IoU}_{\text{Mask}_k, x} = \frac{P[\text{Mask}_k \cap x]}{P[\text{Mask}_k \cup x]} \quad (8)$$

where $P$ is the probability when the condition is true, and $\text{Mask}_k$ is the binary mask that represents the ground-truth of the facial component $k$, which is derived from our re-annotation of various datasets shown on Fig. 7.

For evaluation of part explainability, we compute $\text{PE}_I = \mathbb{E}_{x \sim X} \left[ \max_k (\text{IoU}_{\text{Mask}_k, x}) \right]$ for each input face image. As shown on Tab. III, for each network, we conduct several independent experiments, where a random selection of 100 face images from each face parsing dataset is used to calculate the part explainability.

3) Location Consistency: The metrics of location consistency assess the consistency of the location of the activation peaks [30], [2], which indicate whether the channel of feature maps is able to detect the same face part regardless of the input image, as shown in Fig. 5.
TABLE III
THE EVALUATION METRIC OF PART EXPLAINABILITY ON VARIOUS DATASET: LaPa, AFW, Helen, LFPW and 300-W.

| Network           | LaPa  | AFW   | Helen | LFPW  | 300-W |
|-------------------|-------|-------|-------|-------|-------|
| VGG13 + Softmax   | 0.1642| 0.1612| 0.1646| 0.1625| 0.1554|
| VGG13 + ECLoss    | 0.1710| 0.1687| 0.1711| 0.1702| 0.1623|
| VGG16 + Softmax   | 0.1479| 0.1447| 0.1530| 0.1443| 0.1396|
| VGG16 + ECLoss    | 0.1632| 0.1490| 0.1566| 0.1431| 0.1443|
| VGG19 + Softmax   | 0.0424| 0.0396| 0.0396| 0.0376| 0.0381|
| VGG19 + ECLoss    | 0.0432| 0.0426| 0.0466| 0.0430| 0.0376|

TABLE IV
THE EVALUATION METRIC OF LOCATION CONSISTENCY ON VARIOUS DATASET: LaPa, AFW, Helen, LFPW and 300-W.

| Network           | LaPa  | AFW   | Helen | LFPW  | 300-W |
|-------------------|-------|-------|-------|-------|-------|
| VGG13 + Softmax   | 0.02537| 0.01971| 0.02082| 0.02187| 0.02122|
| VGG13 + ECLoss    | 0.04029| 0.03654| 0.03661| 0.03796| 0.03770|
| VGG16 + Softmax   | 0.02743| 0.03516| 0.03032| 0.04230| 0.02867|
| VGG16 + ECLoss    | 0.02898| 0.03589| 0.03327| 0.04175| 0.03417|
| VGG19 + Softmax   | 0.08075| 0.08307| 0.08767| 0.08217| 0.08450|
| VGG19 + ECLoss    | 0.09382| 0.09503| 0.09592| 0.09453| 0.09418|

TABLE V
THE EVALUATION METRIC OF ACTIVATION ROBUSTNESS ON VARIOUS DATASET: VGGFace, LFW, CFP, AgeDB, Helen and LaPa.

| Network           | VGGFace | LFW   | CFP   | AgeDB | Helen | LaPa |
|-------------------|---------|-------|-------|-------|-------|------|
| VGG13 + Softmax   | 3.733 ± 0.094 | 3.945 ± 0.022 | 3.580 ± 0.205 | 3.984 ± 0.147 | 4.070 ± 0.164 | 3.918 ± 0.236 |
| VGG13 + ECLoss    | 2.596 ± 0.050 | 2.784 ± 0.077 | 2.495 ± 0.126 | 2.718 ± 0.221 | 2.816 ± 0.105 | 2.889 ± 0.166 |
| VGG16 + Softmax   | 3.714 ± 0.081 | 3.994 ± 0.041 | 3.994 ± 0.041 | 3.889 ± 0.167 | 3.946 ± 0.144 | 3.847 ± 0.195 |
| VGG16 + ECLoss    | 3.144 ± 0.086 | 3.272 ± 0.013 | 3.210 ± 0.078 | 3.429 ± 0.165 | 3.242 ± 0.151 | 3.395 ± 0.186 |
| VGG19 + Softmax   | 2.987 ± 0.111 | 2.965 ± 0.247 | 2.646 ± 0.069 | 3.053 ± 0.123 | 3.083 ± 0.148 | 3.068 ± 0.042 |
| VGG19 + ECLoss    | 1.672 ± 0.108 | 1.743 ± 0.164 | 1.493 ± 0.037 | 1.775 ± 0.183 | 1.829 ± 0.053 | 1.849 ± 0.065 |

For evaluation of location consistency, we first find the face component $\hat{k}$ with maximum correspondence by $\hat{k} = \arg\max_k IoU_{\text{Mask}_k, x}$. Therefore, for the $i$-th channel of feature maps $x_i$, we compute its corresponding face part $\hat{k}$ for different input images $I$.

Given a channel’s index $i$, $A_{i,k}$ indicates the probability of the feature map $x_i$ corresponding to the $k$-th face part, i.e., for an arbitrary channel’s index $i$, $\sum_{k \in K} A_{i,k} = 1$. For the $i$-th channel, we use $S_i = \text{var}[A_{i,k}] = \sqrt{\sum_k (A_{i,k} - \bar{A}_i)^2 / |K|}$, $k \in K$ to assess its consistency. For a network architecture, we use $LS = E[S_i] = \sum_i S_i / c$ to evaluate location consistency of the network.

The results are shown in Tab. IV. For each network, we conduct several independent experiments, using a random selection of 100 face images from each face parsing dataset (LaPa, AFW, Helen, LFPW and 300-W).

4) Activation Robustness: The main goal of the proposed ECLoss is to force a feature map to contain at most two activation peaks. As shown in Fig. 5 and Fig. 6, compared to the feature maps of explainable networks trained with ECLoss, the feature maps of original networks usually involve more than two activation peaks. Therefore, when the number of activation peaks is smaller, the activation pattern is more robust.

In simple terms, the purpose of Activation Robustness is to evaluate the average number of peaks in the feature map. Given a feature map $x$ after ReLU operation from the last convolutional layer, then, a contour method [19] will be used to compute peak prominence on the feature map $x$, which uses enclosing contours to define the local maxima.

In the experiments of activation robustness, for all original
networks and explainable networks, we conduct 5 independent experiments to calculate the mean and standard deviation of the number of activation peaks. The results of these experiments have been shown in Tab. V, which demonstrate that our proposed ECLoss achieves better robustness on the six datasets of VGGFace, LFW, CFP, AgeDB, Helen and LaPa. 

VI. CONCLUSION

In this paper, we proposed an Explainable Channel Loss (ECLoss) to construct an explainable face recognition network by forcing the target convolution layer to learn the facial part-based representation. Each channel of the feature maps represents a certain face part, without any manual annotation or additional occlusion datasets. We assess the proposed ECLoss by three traditional deep learning architectures VGG13, VGG16, and VGG19 on dozens of datasets like Glin360K, LFW, CFP, etc. All of our experiments have shown that our proposed ECLoss can achieve significant explainability metrics while improving the performance of face recognition.

VII. ACKNOWLEDGEMENTS

This research was supported by National Natural Science Foundation of China under grant no. 91959108 and 62006156, and Technology Project of Guangdong Province under grant no. 2022A1515012125.

REFERENCES

[1] X. An, X. Zhu, Y. Gao, Y. Xiao, Y. Zhao, Z. Feng, L. Wu, B. Qin, M. Zhang, D. Zhang, et al. Partial Ic: Training 10 million identities on a single machine. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 1445–1449, 2021.

[2] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba. Network dissection: Quantifying interpretability of deep visual representations. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 6541–6549, 2017.

[3] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. IEEE transactions on pattern analysis and machine intelligence, 35(12):2930–2940, 2013.

[4] T. Berg and P. N. Belhumeur. Pool: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 955–962, 2013.

[5] L. Brillouin. Science and information theory. Courier Corporation, 2013.

[6] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5203–5212, 2018.

[7] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

[8] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In Workshop on faces in Real-Life Images: detection, alignment, and recognition, 2008.

[9] X. Huang, J. Xia, and L. Shen. One-class face anti-spoofing based on attention auto-encoder. In Chinese Conference on Biometric Recognition, pages 365–373. Springer, 2021.

[10] S. Kublack. Information theory and statistics. Courier Corporation, 1997.

[11] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang. Interactive facial feature localization. In European conference on computer vision, pages 679–692. Springer, 2012.

[12] Y.-S. Lin, Z.-Y. Liu, Y.-A. Chen, Y.-S. Wang, Y.-L. Chang, and W. H. Hsu. xcos: An explainable cosine metric for face verification task. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 17(3):1–16, 2021.

[13] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis. Explainable ai: A review of machine learning interpretability methods. Entropy, 23(1):18, 2021.

[14] H. Liu, H. Ji, Y. Li, N. He, H. Wu, F. Liu, L. Shen, and Y. Zheng. Robust representation via dynamic feature aggregation. arXiv preprint arXiv:2205.07466, 2022.

[15] H. Liu, H. Wu, W. Xie, F. Liu, and L. Shen. Group-wise inhibition based feature regularization for robust classification. arXiv preprint arXiv:2103.02152, 2021.

[16] Y. Liu, H. Shi, H. Shen, Y. Si, X. Wang, and T. Mei. A new dataset and boundary-attention semantic segmentation for face parsing. In AAM, pages 11637–11644, 2020.

[17] H. Moschoglou, A. Paparoanoos, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou. Agdb: the first manually collected, in-the-wild age database. In proceedings of the IEEE conference on computer vision and pattern recognition workshops, pages 51–59, 2017.

[18] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. 2015.

[19] M. B. Prime and A. T. DeWald. The contour method. Practical residual stress measurement methods, pages 109–138, 2013.

[20] Y. Qi, H. Lin, Y. Li, and L. Chen. Parameter-free attention in fnri decoding. IEEE Access, 9:48704–48712, 2021.

[21] Y. Qi, M. Qiu, H. Lin, J. Chen, Y. Li, and H. Lei. Research on gender-related fingerprint features, extracting fingerprint features using autocoder networks for gender classification. 2022.

[22] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: Database and results. Image and vision computing, 47:3–18, 2016.

[23] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision, pages 618–626, 2017.

[24] S. Sengupta, J.-C. Chen, C. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs. Frontal to profile face verification in the wild. In 2016 IEEE winter conference on applications of computer vision (WACV), pages 1–9. IEEE, 2016.

[25] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[26] J. Xiang and G. Zhu. Joint face detection and facial expression recognition with mtcnn. In 2017 4th international conference on information science and control engineering (ICISCE), pages 424–427. IEEE, 2017.

[27] B. Yin, L. Tran, H. Li, X. Shen, and X. Liu. Towards interpretable face recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 9348–9357, 2019.

[28] T. Zee, G. Gali, and I. Nwogu. Enhancing human face recognition with an interpretable neural network. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, pages 0–0, 2019.

[29] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In European conference on computer vision, pages 818–833. Springer, 2014.

[30] Q. Zhang, Y. N. Wu, and S.-C. Zhu. Interpretable convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 8827–8836, 2018.

[31] W. Zhang, H. Liu, R. Ramachandra, F. Liu, L. Shen, and C. Busch. Face presentation attack detection using taskonomy feature. arXiv preprint arXiv:2111.11046, 2021.

[32] T. Zheng, W. Deng, and J. Hu. Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments. arXiv preprint arXiv:1708.08197, 2017.

[33] Y. Zhong and W. Liu. Deep channel attention in similar-looking face recognition. In 2018 24th International Conference on Pattern Recognition (ICPR), pages 3353–3358. IEEE, 2018.

[34] B. Zhou, D. Bau, A. Oliva, and A. Torralba. Interpreting deep visual representations via network dissection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9348–9357, 2019.

[35] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2879–2886. IEEE, 2012.