Aligning Characteristic Descriptors with Images for Human-Expert-like Explainability

Bharat Chandra Yalavarthi
University at Buffalo-SUNY
Buffalo, NY
byalavar@buffalo.edu

Nalini Ratha
University at Buffalo-SUNY
Buffalo, NY
nratha@buffalo.edu

Abstract

In mission-critical domains such as law enforcement and medical diagnosis, the ability to explain and interpret the outputs of deep learning models is crucial for ensuring user trust and supporting informed decision-making. Despite advancements in explainability, existing methods often fall short in providing explanations that mirror the depth and clarity of those given by human experts. Such expert-level explanations are essential for the dependable application of deep learning models in law enforcement and medical contexts. Additionally, we recognize that most explanations in real-world scenarios are communicated primarily through natural language. Addressing these needs, we propose a novel approach that utilizes characteristic descriptors to explain model decisions by identifying their presence in images, thereby generating expert-like explanations. Our method incorporates a concept bottleneck layer within the model architecture, which calculates the similarity between image and descriptor encodings to deliver inherent and faithful explanations. Through experiments in face recognition and chest X-ray diagnosis, we show that our approach offers a significant contrast over existing techniques, which are often limited to the use of saliency maps. Our approach represents a significant step toward making deep learning systems more accountable, transparent, and trustworthy in the critical domains of face recognition and medical diagnosis.

1 Introduction

Deep learning models have revolutionized various applications, enhancing accuracy, efficiency, and scalability. However, their lack of transparency poses significant risks in security applications like face recognition, where accountability is crucial to identify biases or failures [38]. In legal contexts, identity search decisions must be justified, similar to how facial forensic examiners testify in court [32]. Investigations have exposed biases in commercial face recognition systems, exemplified by a wrongful arrest due to a flawed match [2] [3]. In computer-aided diagnosis, the black-box nature of these models makes it difficult for physicians to trust and explain decisions, raising concerns about biases and spurious correlations [24]. Incorporating explainability or interpretability mechanisms can address these challenges by clarifying decision-making processes, aiding in debugging, and revealing model biases [5]

Although there are several existing methods relating to explainable face recognition [38, 25, 18, 22] they focus on indistinct visual explanations which are less interpretable than textual explanations [32]. Visual explanations can lack the nuanced detail and context often provided in textual explanations and may be subject to misinterpretation [32]. While prior works dealt with explaining the chest X-ray diagnosis made by deep learning models, none have explored using fine-grained, and atomic characteristic descriptors for providing radiologist-like explanations. The prior work can be mainly categorized into the explainable X-ray report generation [34] [19], or saliency maps for the X-ray
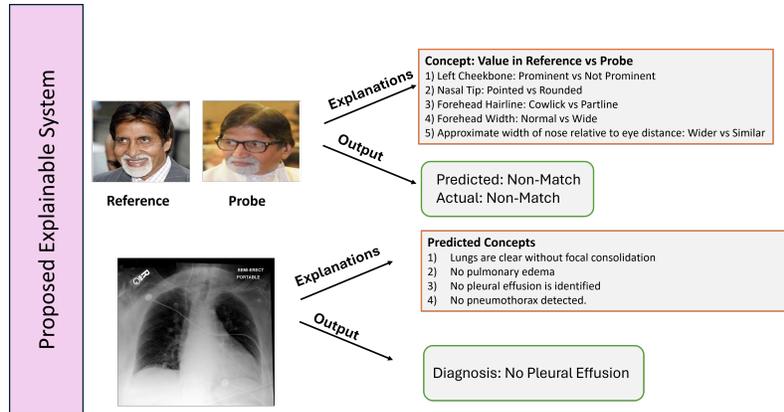
Figure 1: Illustration of our proposed explainable system for face recognition and chest X-ray diagnosis.

images to depict the important regions for the diagnosis [31]. The ability of deep learning systems to provide human-expert-like explanations can help in regulatory compliance especially in legal proceedings, improved communication, and increased user trust and acceptance [32]. Contrary to the prior work, our approach employs precisely defined characteristic descriptors which are textual concepts to emulate the explanatory abilities of human experts in justifying facial recognition and chest X-ray diagnosis decisions, while also ensuring the faithfulness of the explanations through self-explainable architecture. Examples of the explanations given by our method is given in figure 1.

Our framework is designed to function in both supervised and unsupervised environments and we demonstrate the efficacy of our framework in both contexts. For the unsupervised case, we utilize face recognition, and for the supervised case, we apply it to chest X-ray diagnosis. The characteristic descriptors or concepts (used interchangeably) for face recognition are derived from the facial features standard published by FISWG [4] for morphological analysis in face comparison for face recognition. For X-ray diagnosis, we extracted the descriptors from the radiology reports of the MIMIC-CXR [20] dataset. Our key contribution is proposing a generic explainable framework capable of functioning in both supervised and unsupervised contexts, which can be used to provide expert-like explanations to any classification decisions made by deep learning models. We show the results of our method which provides a coherent, faithful, user-friendly, and expert-like textual explanation for the tasks of face recognition, and X-ray diagnosis.

2 Related Work

The related prior work can be broadly divided into three categories i) Vision Language Models (VLMs) for explainability ii) explainable face recognition and iii) explainable X-ray diagnosis.

VLMs for Explainability: Pre-trained VLMs like CLIP [33], and ALIGN [17] have shown good performance in various tasks across several domains in zero-shot and fine-tuning settings [39, 12]. Recently there have been several works where VLM's like CLIP were used to design explainable image classification models. [41, 30, 26] uses the concept bottleneck layer formed by aligning textual concepts and images for explainable image classification. Candidate concepts are usually generated by prompting the LLMs and undergo a selection process. We extend the use of VLMs for explainability to provide expert-level explanations using characteristic descriptors.

Explainable Face Recognition: Most prior work (examples in supplementary material) focuses on generating visual explanations in the form of saliency maps, highlighting the regions of the face that the model considers when making a decision. [27, 18, 28, 38, 7, 22, 15, 25] identify these important regions through different techniques including occlusion, perturbation, similarity measurement, and attention. Visual explanations have several disadvantages, they are not precise, fine-grained, and may be subject to interpretation [32]. In human communications, an explanation response is usually in a textual medium either in spoken or written form as this can provide clear and concise explanations in most cases [32]. Unlike the above works, [8] offers both textual and visual explanations by training

separate networks to identify face attributes and using counterfactual examples to determine key attributes. However, it relies on a limited set of constrained attributes, reducing explanation precision, depth, and coverage compared to our method. [13] [10] evaluate the usage of chatGPT for performing and explaining face recognition. But this method can have reliability issues as these models can hallucinate and faithfulness of the explanations can't be ensured. Moreover, unlike ours, [8] [13] [10] do not provide explanations similar to that of a forensic expert justifying a face match decision.

Chest X-ray Diagnosis: Prior works like [24] and [40], employ saliency maps and report features to enhance disease classification, while others, such as [31], model radiologist gaze for an explanation. Methods like [34] and [9] link image regions to report sections and utilize longitudinal data for interpretable report generation, respectively. Additionally, [37] leverages model-produced explanations to improve classification, and [36] addresses explaining image-text alignment in vision-language models. In contrast, our approach provides radiologist-like explanations using radiologist-defined descriptors for diagnostic decisions.
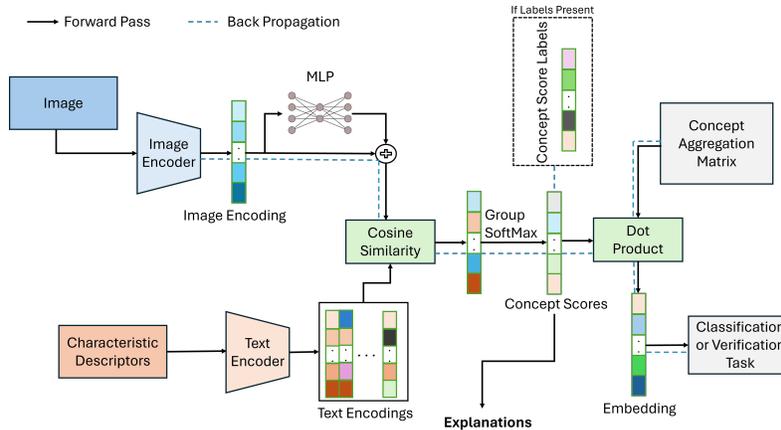
3   Methodology

Figure 2: Proposed architecture produces explanations and embeddings for tasks like classification or face recognition. It can work with or without concept supervision. If concept labels are available, they are used for back-propagating the loss; otherwise, only the task loss is back-propagated.

Our proposed novel explainable methodology is based on using characteristic descriptors for providing human-expert-like explanations. Based on such characteristic descriptors described by human experts we create a set of textual concepts denoted by $C = \{c_1, c_2, ..c_n\}$. These concepts are divided into groups based on the component they are describing. Inspired by the concept bottleneck models (CBM) [23] to design inherently explainable models we use a bottleneck layer in our proposed methodology and employ CLIP [33] [16] to identify these textual concepts in the images. The similarity score between the textual and image embeddings produced by CLIP can give us the concept scores that form the bottleneck layer in our architecture. The concept scores of each image reveal the extent of a concept's presence and are used for providing explanations. Let $X \in R^{H*W*D}$ denote an image where H, W, D are the height, width, and channels of the image, and y its label. We denote the image encoder module of the CLIP as E_i and the text encoder module as E_t. The dot product between the encodings of E_i and E_t shows the match between the image and the text modalities. To fine-tune CLIP we use a skip connection to extract image embeddings as first detailed in [12]. Image encoding $I \in R^d$ is extracted using the following equation:

$$I = \alpha * E_i(X) + (1 - \alpha) * F_i(E_i(X)) \tag{1}$$

where F_i denotes a two-layer network with ReLU activation, which down-samples and then up-samples the embedding back to its original size. On the other hand, each of the tokenized concepts is fed into E_t to get the text encodings $T \in R^{N*d}$. We compute the cosine similarity between I and and each encoding in T to get concept scores $S \in R^N$ which represents the presence of concepts in an image. To better represent these concept scores and the dependencies within a concept group

3

we apply SoftMax independently within each group (denoted as Group SoftMax) of the concept set to obtain S_{sm}. Group SoftMax enhances the representation of embeddings by emphasizing the most activated concept within each group, thereby improving both accuracy and explainability, as demonstrated in an ablation study in the supplementary material. We then transform the S_{sm} using a learned concept aggregation matrix $W \in R^{N*m}$ to get the final embedding $X_{emb} \in R^m$ used for downstream tasks like face recognition or disease classification. The concept aggregation matrix can be further used to understand the feature weights of the concepts. Figure 2 shows the architecture of the proposed explainable framework.
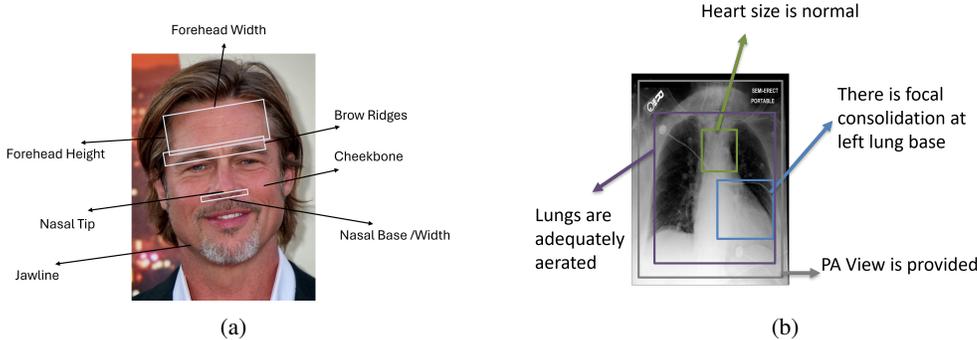

(a)   (b)

Figure 3: Illustrative examples of some of the a) facial characteristic descriptors as recommended in the FISWG Facial Image Comparison Guide. b) characteristic descriptors for chest X-ray diagnosis as defined by radiologists.

3.1 Face Recognition

The FISWG guide listing the facial features to be used for face comparison and morphological analysis is commonly followed by forensic experts. It has detailed characteristic descriptors for each of the nineteen components of the human face. Figure 3 (a) shows some examples of the characteristic descriptors presented in the FISWG guide [4]. We use 120 such descriptors from the guide in our explainable face recognition system. Given a reference and probe face images, we get the face embeddings of these by passing it through the proposed architecture and compute if it is a match or not based on their cosine similarity (shown in figure 4 (a)). Since no face dataset includes labeled characteristic descriptors for faces, and most face descriptors are general features fairly detectable by a generic model, we rely on pre-trained image and text encoders of CLIP[16] trained on LAION 2B internet data to identify these descriptors in an unsupervised manner. We further fine-tuned our model for face recognition using the quality adaptive margin loss function proposed in AdaFace [21]. We tune only the fully connected layer towards the end of the image encoder. Additionally, we adopted the idea from [12] to add a fully connected layer on top of the image encoder and form a residual connection between their respective outputs (denoted as Adaptive FT). The ablation studies involved in determining the optimal architecture are described in the supplementary material.

3.2 Chest X-Ray Diagnosis

Features in chest X-ray images are more niche and subtle, making it challenging for a generic CLIP model to capture them. We use a CLIP model pre-trained on chest X-ray images and radiologist reports which is proposed by [6] as our image and text encoders. The CLIP image encoder and the text encoder are frozen while only the fully connected layers are trained. As we can extract the concept labels from radiologist reports, we can supervise the concepts unlike in face recognition where there was no supervision. Given a corpus of radiology reports, we extract the atomic, fine-grained characteristic descriptors from them using the Mistral 7B language model. We prompt it to disentangle the descriptors to separate sentences from the report. These concepts can describe the presence/absence of a disease or any anatomical finding. Figure 3 (b) shows the illustration of some of the characteristic descriptors for chest X-rays used by radiologists. Given a chest X-ray, its corresponding concept labels, and its diagnosis label we calculate the cosine similarity between the X-ray image and the characteristic descriptor embeddings to obtain concept scores. For obtaining concept score labels we set the concepts present in the report to max value in the calculated concept
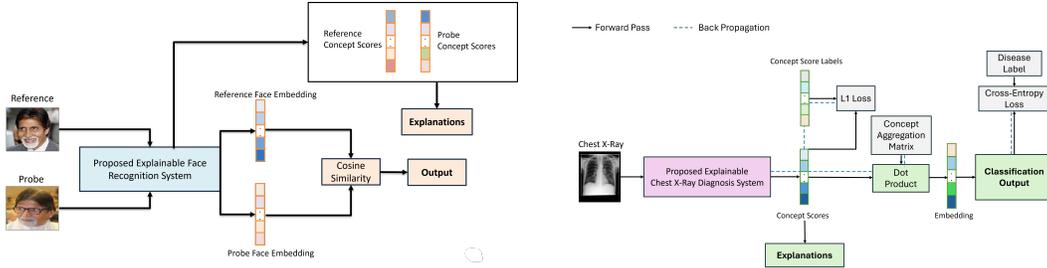
(a) Embeddings are generated from concept scores for each face and used for verification. These scores are also analyzed to produce explanations.

(b) Concept scores are compared with concept score labels to compute the loss. These scores provide explanations and are transformed into the final embedding for classification.

Figure 4: Proposed Explainable Pipelines for (a) Face Verification and (b) Chest X-Ray Diagnosis

scores. L1 loss calculated between the concept scores and concept scores label is used for supervising the model to give appropriate explanations. The concept scores are further passed through a concept aggregation layer to get the logits used for making the diagnosis prediction. As the standard, we use the cross-entropy loss for classification (the entire pipeline is shown in figure 4 (b)).

4 Experiments and Results

4.1 Face Recognition

Ablation Study: Unlike for chest X-ray diagnosis, there is no vision language model fine-tuned for face recognition, so we performed a set of ablation experiments as shown below to determine the best combination of model, architecture, and training methods.

CLIP Model: We evaluate various CLIP variants based on their zero-shot performance in 1:1 face verification tasks. We report the average accuracy on the five validation datasets of each variant in table 1. Based on these results we chose ViT-L/14 trained on the LAION-2B dataset variant for all further experiments.

Table 1: CLIP Variants and Zero-Shot Accuracy on Face Verification.

Variant	Training Dataset	Accuracy
ViT-B/16	DataComp-1B	68.72%
ViT-L/14	OpenAI's WIT	71.21%
ViT-H-14-378-quickgelu	dfn5b	72.29%
ViT-H-14-quickgelu	dfn5b	72.87%
ViT-L/14	**LAION-2B**	**74.11%**

Architecture: In our experiments with various architectural configurations, we find that implementing group SoftMax improved verification accuracy. Furthermore, the combination of CLIP MLP unfreezing, group SoftMax, and a linear layer yielded the highest performance for face recognition tasks among the configurations tested as shown in table 2.

Fine-Tuning: It is crucial to fine-tune the CLIP backbone without compromising its text-image alignment capability, as the faithfulness and validity of the explanations are directly dependent on it. Previous work has demonstrated the strong performance of CLIP in zero-shot image classification settings due to its robust capability to align text and images. We use the zero-shot performance of our fine-tuned CLIP on CIFAR-10 as a proxy for its alignment capability to determine the appropriate level of fine-tuning. Our goal in this experiment is to fine-tune the model to enhance face recognition accuracy without affecting its alignment capability. Table 3 shows that fine-tuning the entire image encoder of CLIP achieved the best face recognition accuracy but compromised text-image alignment, making it unsuitable for explanations. Weight ensemble fine-tuning [39], which interpolates zero-

5

Table 2: Ablation Study of Architecture Choices for Fine-Tuning.

Architecture	Face Recognition Accuracy
CLIP Zero-Shot (No FT)	74.11%
CLIP Zero-Shot (No FT) + Group SoftMax	76.91%
CLIP Zero-Shot (No FT) + Group SoftMax + Linear Layer	76.91%
CLIP Image-Encoder MLP FT + Group SoftMax + Linear Layer (Base)	83.72%

shot and fine-tuned model weights, improved alignment but failed to match zero-shot accuracy. Experiments revealed that Adaptive-FT [12] and fine-tuning the CLIP image encoder's last MLP layer, combined with group SoftMax and a linear layer (Base), enhanced face recognition with minimal loss in alignment. Combining these two techniques achieved the optimal balance between recognition performance and text-image alignment.

Table 3: Ablation Study of CLIP Fine-Tuning methods.

Fine-Tuned Model	Face Recognition Accuracy	Alignment Proxy Accuracy
CLIP Zero-Shot (No FT)	74.11%	96.91%
Entire Image Encoder FT	94.80%	12.59%
Weight Ensemble FT [39]	76.72%	36.16%
Adaptive FT [12]	83.50%	96.82%
Base	83.72%	93.59%
Base + Adaptive FT	**89.50%**	**94.27%**

Face Recognition Method: We have experimented with existing state-of-the-art methods of AdaFace [21], ArcFace [11], and CosFace [35], for training face recognition models to find the best suitable one. AdaFace was shown to be the best performing one with an average accuracy of 89.50% followed by ArcFace with 83.24% and CosFace with 80.14%.

Datasets and Results: A random subset of 500K images from MS1MV2 [11] dataset containing 5.8M images with 85K identities was used for fine-tuning our model. Standard face recognition datasets including LFW [14], CFP-FP [1], AgeDB [29], CPLFW [42], CALFW [43] are used for validation. We resize the cropped and aligned MS1MV2 images to 224 x 224 for compatibility with CLIP's image encoder. The model is fine-tuned for 5 epochs using the AdamW optimizer with a learning rate of 0.0003, following the approach in [39]. Table 4 shows the 1:1 verification performance of our proposed approach on these five datasets compared to the black box SOTA model AdaFace [21]. Although we lose some accuracy compared to the black box model, our approach provides explanations leading to insights and debugging capabilities, which are as important as performance in certain critical applications.

Explanations: We show the ability of our approach to produce faithful explanations justifying its face-matching decisions. Figure 5 shows the examples of the model justifying its decision based on important concepts. The top closest concepts of reference and probe which are commonly activated in each concept group are shown for matching cases, while the top concept groups where the activated concepts differ the most between the reference and probe are shown for non-match cases.

4.2 Chest X-ray Diagnosis

Dataset and Training: We use the MIMIC-CXR dataset which has chest X-rays, corresponding radiologist reports, and disease labels. For this work, we classify the presence of Pleural Effusion conditions from X-ray images using a subset of around 20K samples. We make an 85% train and 15% test split from the chosen subset of the MIMIC-CXR data. We used the pre-trained CLIP model [6] as a baseline and compared its classification performance in detecting pleural effusion against our

Face Recognition

Reference   Probe

Predicted: Non-Match
Actual: Non-Match

Differentiating Concepts in Reference vs Probe (Top-5)
1) Left, Right eyebrow shape is Straight vs Arch
2) Face shape: Square vs Oblong
3) Nasal Body: Straight vs Convex
4) Approximate width of nose relative to eye distance: Similar vs Wider
5) Proportions of features on face: Wider vs Equal

Chest X-Ray Diagnosis

Diagnosis: No Pleural Effusion

Predicted Concepts
1) Cardiac contours are unchanged.
2) Cardiac silhouette are stable.
3) Lungs appear clear.
4) Patient has undergone median sternotomy.

Concepts from Radiologist Report (GT)
1) Cardiac silhouette is normal.
2) Bronchovascular structures are crowded.
3) Cardiac contours are unchanged.
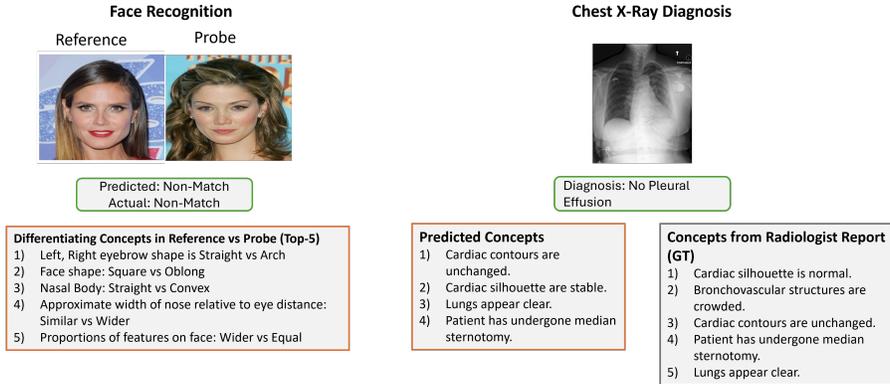4) Patient has undergone median sternotomy.
5) Lungs appear clear.

Figure 5: Examples of explanations provided by our model for its decisions in face recognition and chest X-ray diagnosis.

proposed explainable architecture. We observe from table 5 that the classification performance of our proposed architecture is similar to that of the black box while our approach has an added advantage of providing explanations. The model was fine-tuned for 10 epochs using the Adam optimizer with a learning rate of 0.005 with 1:2 proportions of loss signals from classification and concept prediction.

Explanations: We do a quantitative evaluation of the explanations provided by our model with the labels extracted from radiologist reports. We use two metrics for our evaluation, i) METEOR ii) Rouge-L score for measuring the quality of explanations when compared against the ground truth. The evaluation results are 0.31 for ROUGE-L and 0.27 for METEOR. This exhibits the fidelity of our explanations. One possible reason for the METEOR score not being higher could be the difference in the ordering of concepts between the predictions and the ground truth. We do not optimize for ordering as we treat all ground truth concepts as equally important. Figure 5 shows example results of our explainable chest X-ray diagnosis system.

Table 4: Performance (1:1 verification accuracy) of our proposed explainable face recognition approach on 5 benchmark datasets compared with black box SOTA [21].

Dataset	Proposed (Explainable)	Baseline (Black box - SOTA)
LFW [14]	98.38%	99.82%
CFP-FP [1]	90.88%	98.49%
CALFW [43]	89.36%	96.08%
AgeDB [29]	81.18%	98.05%
CPLFW [42]	87.71%	93.13%
Average	89.50%	97.11%

Table 5: Performance of our proposed explainable chest X-ray diagnosis approach on MIMIC-CXR dataset compared with black box model.

Metric	Proposed (Explainable)	Baseline (Black Box)
Accuracy	83.78%	86.64%
Precision	84.77%	86.4%
Recall	83.78%	87.3%
F1-Score	84.19%	86.8%

5 Conclusion

In this work, we propose a methodology to explain the decisions made by deep learning systems similar to that of a human expert thereby providing improved explainability and expert-level insight. We show that we can design models that give faithful and concrete explanations like a human expert using characteristic descriptors. Through our experiments, we show the performance on benchmark datasets proving the efficacy of our model in providing explanations without significantly affecting the performance in comparison with black-box models. We hope our proposed method with the ability to produce consumable and verifiable descriptions can address transparency and trustworthiness in face recognition and X-ray diagnosis systems.

References

[1] Frontal to profile face verification in the wild. In: 2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016. 2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016, Institute of Electrical and Electronics Engineers Inc. (May 2016). https://doi.org/10.1109/WACV.2016.7477558

[2] https://www.nytimes.com/2023/08/06/business/facial-recognition-false-arrest.html (2023)

[3] https://www.nytimes.com/2019/12/19/technology/facial-recognition-bias.html (2023)

[4] https://fiswg.org/FISWG_Morph_Analysis_Feature_List_v2.0_20180911.pdf (2023)

[5] Abid, A., Yuksekgonul, M., Zou, J.: Meaningfully debugging model mistakes using conceptual counterfactual explanations. In: Proceedings of the 39th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 162. PMLR (2022), https://proceedings.mlr.press/v162/abid22a.html

[6] Boecking, B., Usuyama, N., Bannur, S., Castro, D.C., Schwaighofer, A., Hyland, S., Wetscherek, M., Naumann, T., Nori, A., Alvarez-Valle, J., Poon, H., Oktay, O.: Making the most of text semantics to improve biomedical vision–language processing. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022. pp. 1–21. Springer Nature Switzerland, Cham (2022)

[7] Brito, J., Proença, H.: A deep adversarial framework for visually explainable periocular recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1453–1461 (2021)

[8] Chen, R., Li, J., Zhang, H., Sheng, C., Liu, L., Cao, X.: Sim2word: Explaining similarity with representative attribute words via counterfactual explanations. ACM Transactions on Multimedia Computing, Communications and Applications **19**(6), 1–22 (2023)

[9] Dalla Serra, F., Wang, C., Deligianni, F., Dalton, J., O'Neil, A.: Controllable chest X-ray report generation from longitudinal representations. In: Bouamor, H., Pino, J., Bali, K. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 4891–4904. Association for Computational Linguistics, Singapore (Dec 2023). https://doi.org/10.18653/v1/2023.findings-emnlp.325, https://aclanthology.org/2023.findings-emnlp.325

[10] DeAndres-Tame, I., Tolosana, R., Vera-Rodriguez, R., Morales, A., Fierrez, J., Ortega-Garcia, J.: How good is chatgpt at face biometrics? a first look into recognition, soft biometrics, and explainability. IEEE Access (2024)

[11] Deng, J., Guo, J., Yang, J., Xue, N., Kotsia, I., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(10), 5962–5979 (Oct 2022). https://doi.org/10.1109/tpami.2021.3087709, http://dx.doi.org/10.1109/TPAMI.2021.3087709

[12] Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y.: Clip-adapter: Better vision-language models with feature adapters. International Journal of Computer Vision **132**, 1–15 (09 2023). https://doi.org/10.1007/s11263-023-01891-x

[13] Hassanpour, A., Kowsari, Y., Shahreza, H.O., Yang, B., Marcel, S.: Chatgpt and biometrics: an assessment of face recognition, gender detection, and age estimation capabilities. arXiv preprint arXiv:2403.02965 (2024)

[14] Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Tech. Rep. 07-49, University of Massachusetts, Amherst (October 2007)
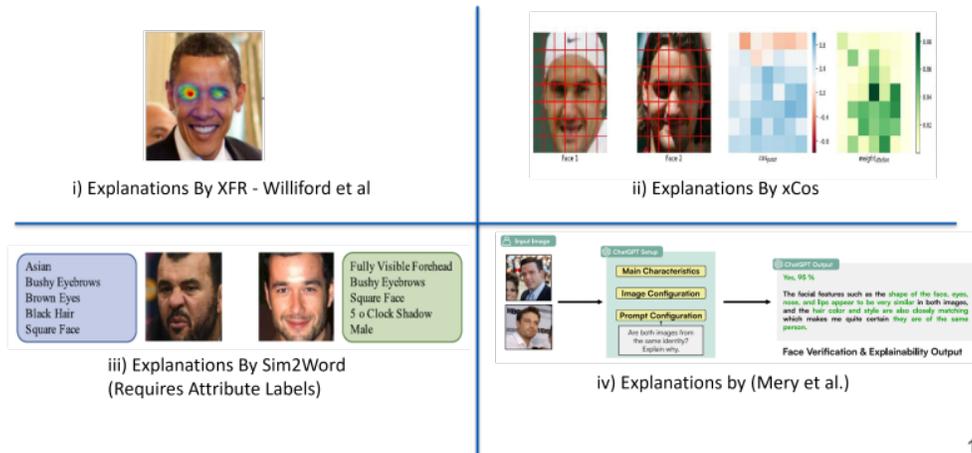
[15] Huber, M., Luu, A.T., Terhörst, P., Damer, N.: Efficient explainable face verification based on similarity score argument backpropagation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 4736–4745 (2024)

[16] Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., Schmidt, L.: Openclip (Jul 2021). https://doi.org/10.5281/zenodo.5143773, `https://doi.org/10.5281/zenodo.5143773`, if you use this software, please cite it as below.

[17] Jia, C., Yang, Y., Xia, Y., Chen, Y., Parekh, Z., Pham, H., Le, Q.V., Sung, Y., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. CoRR **abs/2102.05918** (2021)

[18] Jiang, H., Zeng, D.: Explainable face recognition based on accurate facial compositions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1503–1512 (2021)

[19] Jin, H., Che, H., Lin, Y., Chen, H.: Promptmrg: Diagnosis-driven prompts for medical report generation (2024)

[20] Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data **6**(1), 317 (2019)

[21] Kim, M., Jain, A.K., Liu, X.: Adaface: Quality adaptive margin for face recognition (2023)

[22] Knoche, M., Teepe, T., Hörmann, S., Rigoll, G.: Explainable model-agnostic similarity and confidence in face verification. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 711–718 (2023)

[23] Koh, P.W., Nguyen, T., Tang, Y.S., Mussmann, S., Pierson, E., Kim, B., Liang, P.: Concept bottleneck models. In: III, H.D., Singh, A. (eds.) Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 5338–5348. PMLR (13–18 Jul 2020), `https://proceedings.mlr.press/v119/koh20a.html`

[24] Lei, Y., Li, Z., Shen, Y., Zhang, J., Shan, H.: Clip-lung: Textual knowledge-guided lung nodule malignancy prediction. In: Greenspan, H., Madabhushi, A., Mousavi, P., Salcudean, S., Duncan, J., Syeda-Mahmood, T., Taylor, R. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. pp. 403–412. Springer Nature Switzerland, Cham (2023)

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

[26] Menon, S., Vondrick, C.: Visual classification via description from large language models (2022)

[27] Mery, D.: True black-box explanation in facial analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1596–1605 (2022)

[28] Mery, D., Morris, B.: On black-box explanation for face verification. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3418–3427 (2022)

[29] Moschoglou, S., Papaioannou, A., Sagonas, C., Deng, J., Kotsia, I., Zafeiriou, S.: Agedb: The first manually collected, in-the-wild age database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (July 2017)

[30] Oikarinen, T., Das, S., Nguyen, L.M., Weng, T.W.: Label-free concept bottleneck models (2023)

[31] Pham, T., Brecheisen, J., Nguyen, A., Nguyen, H., Le, N.: I-ai: A controllable and interpretable ai system for decoding radiologists' intense focus for accurate cxr diagnoses. In: 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 7835–7844. IEEE Computer Society, Los Alamitos, CA, USA (jan 2024). https://doi.org/10.1109/WACV57701.2024.00767, `https://doi.ieeecomputersociety.org/10.1109/WACV57701.2024.00767`

[32] Phillips, P.J., Przybocki, M.A.: Four principles of explainable AI as applied to biometrics and facial forensic algorithms. CoRR **abs/2002.01014** (2020)

[33] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML (2021)

[34] Tanida, T., Muller, P., Kaissis, G., Rueckert, D.: Interactive and explainable region-guided radiology report generation. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7433–7442. IEEE Computer Society, Los Alamitos, CA, USA (jun 2023). https://doi.org/10.1109/CVPR52729.2023.00718, `https://doi.ieeecomputersociety.org/10.1109/CVPR52729.2023.00718`

[35] Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W.: Cosface: Large margin cosine loss for deep face recognition. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5265–5274 (2018). https://doi.org/10.1109/CVPR.2018.00552

[36] Wang, Y., Rudner, T.G.J., Wilson, A.G.: Visual explanations of image-text representations via multi-modal information bottleneck attribution (2023)

[37] Watson, M., Hasan, B.A.S., Al Moubayed, N.: Learning how to mimic: Using model explanations to guide deep learning training. In: 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 1461–1470 (2023). https://doi.org/10.1109/WACV56688.2023.00151

[38] Williford, J.R., May, B.B., Byrne, J.: Explainable face recognition. In: European conference on computer vision. pp. 248–263. Springer (2020)

[39] Wortsman, M., Ilharco, G., Kim, J.W., Li, M., Kornblith, S., Roelofs, R., Lopes, R.G., Hajishirzi, H., Farhadi, A., Namkoong, H., Schmidt, L.: Robust fine-tuning of zero-shot models. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7949–7961 (2022). https://doi.org/10.1109/CVPR52688.2022.00780

[40] Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 21315–21326. IEEE Computer Society, Los Alamitos, CA, USA (oct 2023). https://doi.org/10.1109/ICCV51070.2023.01954, `https://doi.ieeecomputersociety.org/10.1109/ICCV51070.2023.01954`

[41] Yang, Y., Panagopoulou, A., Zhou, S., Jin, D., Callison-Burch, C., Yatskar, M.: Language in a bottle: Language model guided concept bottlenecks for interpretable image classification (2023)

[42] Zheng, T., Deng, W., Hu, J.: Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments (2017)

[43] Zheng, T., Deng, W., Hu, J.: Cross-age LFW: A database for studying cross-age face recognition in unconstrained environments. CoRR **abs/1708.08197** (2017), `http://arxiv.org/abs/1708.08197`

A   Appendix / supplemental material

A.1   Explanations provided by prior work

Figures 6 and 7 show the explanations provided by current explainability methods in face recognition and chest X-ray diagnosis respectively.

i) Explanations By XFR - Williford et al

ii) Explanations By xCos

iii) Explanations By Sim2Word
(Requires Attribute Labels)

iv) Explanations by (Mery et al.)

Figure 6: Some examples of existing explainable face recognition methods. Most existing methods focus on saliency maps, while some use textual features.

i) Explanations by Tanida et. al

ii) Explanations by Pham et. al
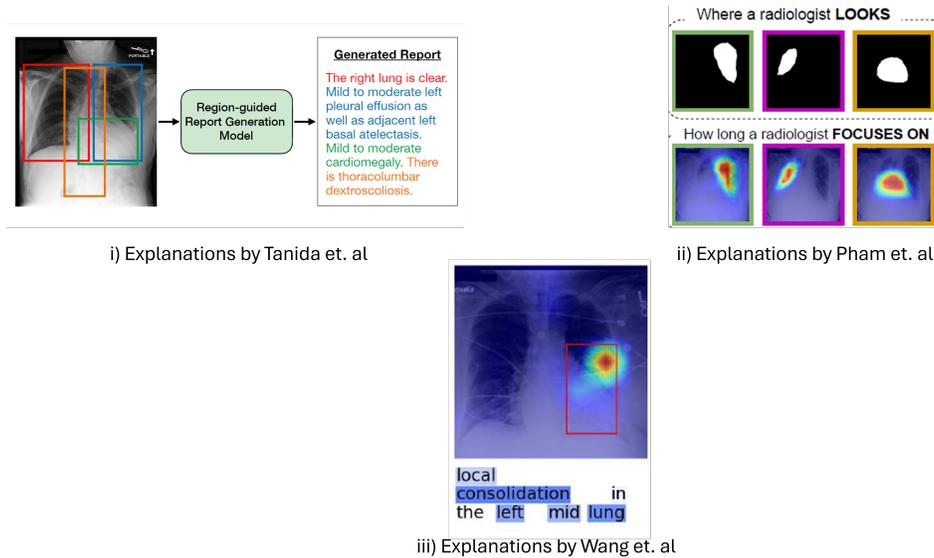
iii) Explanations by Wang et. al

Figure 7: Some examples of existing explainable chest X-ray diagnosis methods. A major focus of work in this area has been finding the salient features in the image, text, or both.