Visual Explanation for Deep Metric Learning

Sijie Zhu, Taojiannan Yang, *Graduate Student Member, IEEE*, and Chen Chen, *Member, IEEE*

***Abstract*— This work explores the visual explanation for deep metric learning and its applications. As an important problem for learning representation, metric learning has attracted much attention recently, while the interpretation of the metric learning model is not as well-studied as classification. To this end, we propose an intuitive idea to show where contributes the most to the overall similarity of two input images by decomposing the final activation. Instead of only providing the overall activation map of each image, we propose to generate point-to-point activation intensity between two images so that the relationship between different regions is uncovered. We show that the proposed framework can be directly applied to a wide range of metric learning applications and provides valuable information for model understanding. Both theoretical and empirical analyses are provided to demonstrate the superiority of the proposed overall activation map over existing methods. Furthermore, our experiments validate the effectiveness of the proposed point-specific activation map on two applications, i.e. cross-view pattern discovery and interactive retrieval. Code is available at https://github.com/Jeff-Zilence/Explain_Metric_Learning**

***Index Terms*— Deep metric learning, visual explanation, convolutional neural networks, activation decomposition.**

I. Introduction

LEARNING the similarity metrics between arbitrary images is a fundamental problem for a variety of tasks, such as image retrieval [1], verification [2], [3], localization [4], video tracking [5], etc. Recently the deep Siamese network [6] based framework has become a standard architecture for metric learning and achieves exciting results on a wide range of applications [7]. However, there are surprisingly few works conducting visual analysis to explain why the learned similarity of a given image pair is high or low. *Specifically, which part contributes the most to the similarity is a straightforward question and the answer can reveal important hidden information about the model as well as the data.*

Previous visual explanation works mainly focus on the interpretation of deep neural network for classification [8]–[11]. Guided back propagation (guided BP) [8] has been used for explanation by generating the gradient from prediction to input, which shows how much the output will change with a little change in each dimension of the input. Another representative visual explanation approach, class activation map (CAM) [9], generates the heatmap of discriminative regions corresponding to a specific class based on the linearity of global average pooling (GAP) and fully connected (FC) layer. However, the original method only works on this specific architecture configuration (i.e. GAP+FC) and needs re-training for visualizing other applications. Based on the gradient of the last convolutional layer instead of the input, Grad-CAM [12] is proposed to generate activation maps for all convolutional neural network (CNN) architectures. Besides, other existing methods explore network ablation [10], the winner-take-all strategy [13], inversion [14], and perturbation [11] for visual explanation.

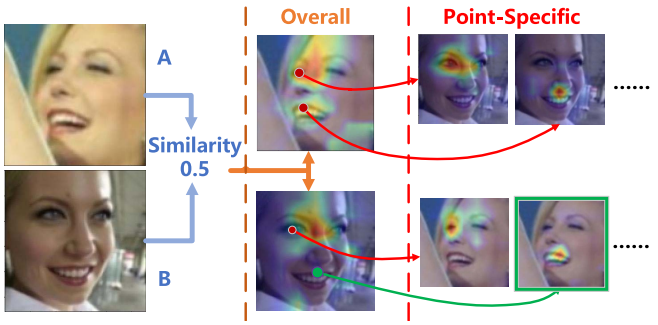


Fig. 1. An overview of activation decomposition. The overall activation map (the second column) on each image highlights the regions contributing the most to the similarity. The point-specific activation map (the third column) highlights the regions in one image that have large activation responses on a specific position, e.g. mouth or eye, in the other image.

Since verification applications like person re-identification (re-id) [3] usually train metric learning models along with classification, recent work [15] leverages the classification activation map to improve the performance, but the activation map of metric learning is still not well explored. For two given images, a variant of Grad-CAM has been used for visualization of image retrieval [16] by computing the gradient from the cosine similarity of the embedding features to the last convolutional layers of both images. However, Grad-CAM only provides the overall highlighted regions of two input images, the relationship between each activated region of two images is yet to be uncovered. Since the similarity is calculated from two images and possibly based on several similar patterns between them, the relationship between these patterns is critical for understanding the model.

In this paper, we propose an **activation decomposition** framework for visual explanation of deep metric learning and explore the relationship between each activated region

Manuscript received July 30, 2020; revised March 28, 2021 and June 29, 2021; accepted August 7, 2021. Date of publication September 1, 2021; date of current version September 8, 2021. This work was supported in part by the National Science Foundation under Grant 1910844. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jianbing Shen. *(Corresponding author: Chen Chen.)*

The authors are with the Center for Research in Computer Vision, University of Central Florida, Orlando, FL 32816 USA (e-mail: sizhu@knights.ucf.edu; taoyang1122@knights.ucf.edu; chen.chen@crcv.ucf.edu).

Digital Object Identifier 10.1109/TIP.2021.3107214

1941-0042 © 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: New Jersey Institute of Technology. Downloaded on September 30,2021 at 20:38:34 UTC from IEEE Xplore. Restrictions apply.

by *point-to-point activation* response between two images. As shown in Fig. 1, the overall activation maps (the second column) of two input images are generated by decomposing their similarity (score) along each image. In this example, the query image (A) has a high/strong activation on **both the eyes and mouth areas**, but the overall map of the retrieved image (B) **only highlights the eyes**. Therefore, it is actually hard to understand how the model works only based on the overall maps. For image B, the mouth region (green point) has a low activation which means the activation between the mouth and **the whole image A** is low compared to the overall activation (similarity). However, by further decomposing this activation (green point) along image A, that is to compute the activation between the mouth region of image B and **each position in image A**, the resulting *point-specific* activation map (green box) reveals that the mouth region of image B still has a high response on the mouth region of image A. This point-specific activation map can be generated for each pixel, which encodes the point-to-point activation intensity for representing the relationship between regions in both images, e.g. eye-to-nose or mouth-to-mouth. Compared with the overall activation map, the point-specific activation map provides much more refined information about the model which is crucial for explanation.

The main contributions of this paper are summarized as follows.

- We propose a novel explanation framework for deep metric learning architectures based on activation decomposition. It can be served as a white-box explanation tool to better understand the metric learning models without modifying the model architectures, and is applicable to a host of metric learning applications, e.g. face recognition, person re-identification, image retrieval, etc.
- The proposed point-specific activation map uncovers the point-to-point activation intensity which, to the best of our knowledge, has not been explored by existing methods. Our experiments further show the importance of the point-specific activation map on two practical applications, i.e. cross-view pattern discovery and interactive retrieval.
- We provide both theoretical and empirical analysis to show the superiority of the proposed overall activation map for deep metric learning over the popular Grad-CAM algorithm.

The remainder of this paper is organized as follows. Section II provides a brief review on existing visual interpretation methods for classification and metric learning. Section III introduces the idea of activation decomposition for visual interpretation on a simple network architecture. Section IV further extends the framework to more complex architectures for metric learning as a unified solution. Section V presents a detailed comparison between the proposed method and Grad-CAM. Section VI validates the effectiveness of the overall activation map of our method, and Section VII demonstrates the unique advantages of the point-specific activation map on two practical applications. Finally, Section VIII concludes the paper.

II. Related Work

A. Interpretation for Classification

As a default setting, most existing interpretation methods [8], [9], [11], [13] are designed in the classification context where the output score is generated from one input image. Among the approaches that do not require architecture change, one intuitive idea [9], [13] is to look into the model and see where contributes the most to the final prediction. CAM [9] and its variant Grad-CAM [12] are among the most widely used approaches, but recent works [11], [12], [17] simply consider CAM as a heuristic linear combination of convolutional feature maps which is limited to the original architecture, global average pooling (GAP) and one FC layer [9], [12], [17]. Specifically, to generate CAM for standard VGG [18] with global max pooling (GMP) and three FC layers, [9], [17] have to replace this architecture with GAP+FC and re-train the model. In this paper, we show that CAM is actually a *special case* of activation decomposition on the specific architecture, and the activation decomposition idea applies to more complex architectures without re-training, even beyond the classification problem.

Grad-CAM [12] is considered as a generalization of CAM for arbitrary CNN architectures, but the original paper only provides the proof on CAM's architecture. We show that Grad-CAM is not equivalent to activation decomposition for some architectures (Section V).

Another direction of explanation is to check the black box model by modifying the input and observing the response in the prediction. A representative method [11] aims to optimize the blurred region and see which region has the strongest response on the output signal when it is blurred. Perturbation optimization is applicable for any black box model, but the optimization is computationally expensive.

B. Interpretation for Metric Learning

As ignored by most existing methods, the explanation for metric learning is important for a wide range of applications [19], [20], e.g. weakly supervised localization. In this work, we further shows its potential significance on diagnosis of metric learning losses and new applications including cross-view pattern discovery and interactive retrieval for person re-identification [21]–[25] and face recognition. Gradient-based methods, e.g. guided BP [8] can be adopted to metric learning, but recent works [11], [26] claim that gradient can be irrelevant to the model and guided-BP fails on the sanity check [26]. Among methods that pass the sanity check, Grad-CAM has been used for visualization of image retrieval in [16], but no quantitative result is provided in the paper. Instead of adding more gradient heuristics like Grad-CAM++ [27], we propose to explain the result of Grad-CAM under the decomposition framework. As shown in Section V, by removing the heuristic step in Grad-CAM, the meaning of the generated activation map can be clearly explained using the proposed framework.

Recent work [19] also explores visualization techniques on metric learning, but it only provides qualitative results on several specific architectures. Another variant [20] replaces the

Authorized licensed use limited to: New Jersey Institute of Technology. Downloaded on September 30,2021 at 20:38:34 UTC from IEEE Xplore. Restrictions apply.

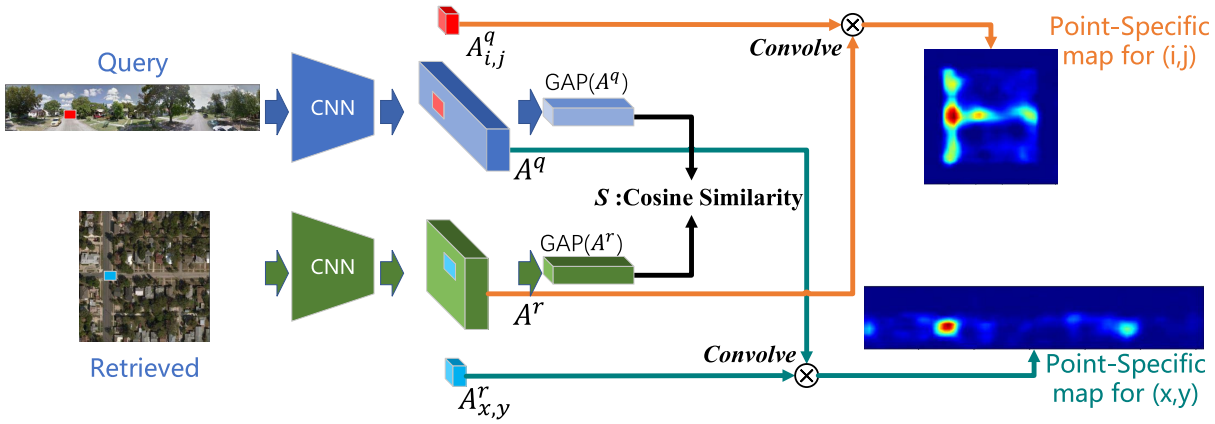


Fig. 2. Activation decomposition on a simple architecture for deep metric learning.

classification score with triplet loss in Grad-CAM framework and deploys weights transfer from the training data in the inference phase. The method requires three samples to form the triplet and its usage in the inference phase is dependent upon the specifically designed transfer procedure, thus is not a general method given only the pretrained model. Moreover, all the aforementioned methods only focus on the overall activation map which is not satisfactory for understanding the model as discussed in Section I. Apart from white-box explanation, black-box optimization method, such as perturbation [11], needs reformulation for metric learning and can be more computationally expensive if the point-to-point activation map is desired.

III. Activation Decomposition on a Simple Architecture

To better introduce our method, we first review the formulation of CAM and illustrate our idea based on a simple architecture (CNN+GAP in Fig. 2) for metric learning. *Note that the proposed method is not limited to this architecture, the general formulation will be presented in Section IV*. As shown in Eq. 1, CAM is actually a spatial decomposition of the prediction score of each class, and the original method only applies for CNN with GAP and one FC layer without bias. The decomposition clearly shows how much each part of the input image contributes to the overall prediction of one class and provides valuable information about how the decision is made inside the classification model. The idea is based on the linearity of GAP:

$$S_c = \sum_k \omega_{k,c} \left(\frac{1}{Z} \sum_{i,j} A_{i,j,k} \right) = \frac{1}{Z} \sum_{i,j} \left(\sum_k \omega_{k,c} A_{i,j,k} \right). \quad (1)$$

S_c denotes the overall score (before softmax) of class c and $\omega_{k,c}$ is the FC layer parameter for the k-th channel of class c. $A_{i,j,k}$ denotes the feature map of the last convolutional layer at position (i, j), and Z is the normalization term of GAP. Here, Z equals to $m \times n$, which is the spatial size of feature map $A \in \mathbb{R}^{m \times n \times p}$. In fact, the result $\sum_k \omega_{k,c} A_{i,j,k}$ can be considered as a decomposition of S_c along (i, j). For a two-stream Siamese-style architecture (see Fig. 2) with GAP and the cosine similarity metric (S) for metric learning applications (e.g. image retrieval), we propose to do decomposition along (i, j, x, y) so that the relationship between different parts of two images are uncovered:

$$\begin{aligned} S &= (E^q \cdot E^r)/|E^q||E^r| \\ &= \sum_k GAP(A^q_k) GAP(A^r_k)/|E^q||E^r| \\ &= \frac{1}{Z} \sum_k \left(\sum_{i,j} A^q_{i,j,k} \sum_{x,y} A^r_{x,y,k} \right) \\ &= \frac{1}{Z} \sum_{i,j,x,y} \left(\sum_k A^q_{i,j,k} A^r_{x,y,k} \right). \end{aligned} \quad (2)$$

Here, E^q,E^r denote the embedding feature of query and reference image. $|E|$ denotes L2 norm of E and $\cdot$ is the inner product. The normalization term Z equals to $m^q n^q m^r n^r |E^q||E^r|$, if the spatial size of A^q, A^r are $m^q \times n^q$, $m^r \times n^r$. We use (i, j) and (x, y) for different streams because some applications such as cross-view image retrieval [4] may have different image sizes for two streams. A denotes the feature map of the last convolutional layer. The superscripts q and r respectively denote the query and retrieved images in this paper. For each query point (i, j) in the query image, the corresponding **point-specific activation map** in the retrieved image is given by $\sum_k A^q_{i,j,k} A^r_{x,y,k}$ which is the contribution of features at (i, j, x, y) to the overall cosine similarity. Like CAM and Grad-CAM, bilinear interpolation is implemented to generate the contribution of each pixel pair and the full resolution map. The **overall activation maps** of two images are generated by a simple summation along (i, j) or (x, y), e.g. $\sum_{i,j} (\sum_k A^q_{i,j,k} A^r_{x,y,k})$. Although we only show the positive activation in the map, the negative value is also available for counterfactual explanation like [12].

Recent works [28], [29] have highlighted the superiority of L2 normalization, we therefore adopt the cosine similarity S as the default metric. With L2 normalization, the squared

Authorized licensed use limited to: New Jersey Institute of Technology. Downloaded on September 30,2021 at 20:38:34 UTC from IEEE Xplore. Restrictions apply.

Euclidean distance D equals to $2-2S$ so that S and D are equivalent as metrics. Although there are still a number of methods [3], [30] utilizing the Euclidean distance without L2 normalization, [3] shows that the cosine similarity performs well as the evaluation metric. We further empirically show that the cosine similarity works well for explanation in this case (Section VII-B).

IV. Generalization for Complex Architectures

Recent metric learning approaches usually leverage more complex architectures to improve performance, e.g. adding several FC layers after the flattened feature or global pooling. Although different metric learning applications have different head architectures (e.g. GAP+FC) on CNN, the basic components are highly similar. Given the introduction of activation decomposition on a simple architecture in the previous section, we present an unified extension to make our method applicable to most existing state-of-the-art CNN architectures for various applications, including image retrieval [7], face recognition [31], person re-identification [3], and image geo-localization [4], [32]. Specifically, in Section IV-A, we address linear components by considering them together as one linear transformation. Then Section IV-B focuses on transforming nonlinear components to linear ones in the *inference phase*.

A. Linear Component

For the last convolutional layer feature $A \in \mathbb{R}^{m\times n\times p}$ (m, n and p denote the width, height, and number of channels), its GAP (global average pooling) is equivalent to the flattened feature $\hat{A} \in \mathbb{R}^{mnp}$ multiplied by a transformation matrix $T_{GAP} \in \mathbb{R}^{p\times mnp}$:

$$GAP(A) = \frac{1}{mn}\sum_{i,j} A_{i,j} = T_{GAP}\hat{A}. \tag{3}$$

Here (i, j) denotes the spatial coordinates. By reshaping T_{GAP} to $p \times m \times n \times p$ as T^*_{GAP}, the matrix is given by:

$$T^*_{GAP}(k', i, j, k) = \begin{cases} \frac{1}{mn} & k' = k \\ 0 & k' \neq k \end{cases} \tag{4}$$

The matrix T_{GAP} is simply calculated by reshaping T^*_{GAP} to $p \times mnp$. Therefore, the GAP can be considered as a special case of flattened layer multiplied by a transformation matrix. Without loss of generality, we consider a two-stream framework with flattened layer followed by one FC layer with weights $\hat{W} \in \mathbb{R}^{l\times mnp}$ and biases $B \in \mathbb{R}^l$ (l denotes the length of the feature embedding vector). Since the visual explanation is generated in the inference phase when typical components such as FC layer and batch normalization (BN) are linear, all the linear components together are formulated as one linear transformation $g(\hat{A}) = \hat{W}\hat{A} + B = \sum_{i,j} W_{i,j}A_{i,j} + B$ in the FC layer. Here $W_{i,j} \in \mathbb{R}^{l\times p}$ and $A_{i,j} \in \mathbb{R}^p$ denote the weights matrix and feature vector corresponding to position (i, j). Although B is ignored in CAM, we keep it as a residual term in the decomposition. Then Eq. 2 is re-formulated as:

$$\begin{aligned} SZ &= g^q(A^q)\cdot g^r(A^r) \\ &= \left(\sum_{i,j} W^q_{i,j}A^q_{i,j} + B^q\right)\cdot\left(\sum_{x,y} W^r_{x,y}A^r_{x,y} + B^r\right) \\ &= \sum_{i,j,x,y} \underbrace{(W^q_{i,j}A^q_{i,j})\cdot(W^r_{x,y}A^r_{x,y})}_{\textbf{point-to-point activation}} + \sum_{i,j}(W^q_{i,j}A^q_{i,j}\cdot B^r) \\ &\quad + \sum_{x,y}(W^r_{x,y}A^r_{x,y}\cdot B^q) + B^q\cdot B^r. \end{aligned} \tag{5}$$

Here Z denotes the normalization term for cosine similarity (L2 norm of embedding features), and $\cdot$ is the inner product. As can be seen from Eq. 5, the decomposition of S has four terms, and the last three terms contain the bias B. The first term clearly shows the activation response for location pair (i, j, x, y) and the **point-specific** map for query position (i, j) is given by $I(x, y) = (W^q_{i,j}A^q_{i,j})\cdot(W^r_{x,y}A^r_{x,y})$. The second and third terms correspond to the activation between one image and the bias of the other image. Although they can be considered as negligible bias terms, they actually contribute to the overall activation map. For the overall activation map of the query image, the second term can be included, because $W^q_{i,j}A^q_{i,j}\cdot B^r$ varies at different (i, j) positions while the third and last terms stay unchanged. Similarly, the first and third terms are considered when calculating the overall map for the retrieved image. We investigate both settings about the bias term (i.e. w or w/o) on the overall activation map and they are referred to as "Decomposition": $(W^q_{i,j}A^q_{i,j})\cdot(\sum_{x,y} W^r_{x,y}A^r_{x,y})$, and "Decomposition+Bias": $(W^q_{i,j}A^q_{i,j})\cdot(\sum_{x,y} W^r_{x,y}A^r_{x,y}+B^r)$ in Section VI. The last term is pure bias which is the same for every input image. Therefore, we ignore this term if not mentioned.

B. Non-Linear Component

Non-linear transformation increases the difficulty of activation decomposition, our idea is to find its approximate linear transformation which can be directly integrated into the formulation of Section IV-A. For a non-linear transformation $f(x)$ with its current input data $x \in [x_0 - \delta, x_0 + \delta]$ (δ is a small number), the first two terms of the Taylor expansion of $f(x)$ can be considered as its linear transformation *in the inference phase*:

$$\begin{aligned} f(x) &= f(x_0) + \frac{f'(x_0)}{1!}(x - x_0) + \dots \\ &\approx f(x_0) + f'(x_0)(x - x_0). \end{aligned} \tag{6}$$

Therefore, our method can generalize to more complex architectures as long as the network is differentiable, since the formulation of gradient $f'(x_0)$ can be computed by chain rule.

In this paper, we shed light on the practical guidelines for deploying our method on image retrieval, face recognition, person re-identification, and cross-view geo-localization frameworks. For these popular applications, the most widely used non-linear components – global maximum pooling (GMP) and rectified linear unit (ReLU) – can be transformed as linear

Authorized licensed use limited to: New Jersey Institute of Technology. Downloaded on September 30,2021 at 20:38:34 UTC from IEEE Xplore. Restrictions apply.

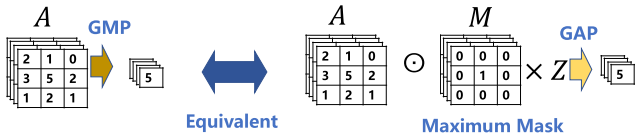


Fig. 3. Maximum mask for GMP. Z is the constant normalization term for GAP.

operations in the inference phase by multiplying a mask.[1] In the inference phase, GMP can be considered as a combination of a maximum mask M and GAP as shown in Fig. 3. The transformation matrix of GMP is given by

$$T^*_{GMP} = mn(T^*_{GAP} \odot M). \tag{7}$$

$M \in R^{m\times n\times p}$ is the maximum matrix of A where only the maximum position of each channel has nonzero value as 1. T_{GMP} is computed by reshaping T^*_{GMP} to $p \times mnp$. The result of Hadamard product $(M \odot A)$ is considered as the new feature map which can be directly applied to Eq. 5. Similarly, by adding a mask for ReLU, the FC layer with ReLU can be included in W and B in the inference phase.

V. Comparison With Grad-CAM

In this section, we provide a comprehensive comparison between the proposed method and Grad-CAM with theoretical derivation. Apparently, the proposed method can generate point-specific activation map which uncovers more fine-grained information than Grad-CAM, as Grad-CAM only generates overall activation map. We will further show the relationship between our overall activation map and Grad-CAM. As another way for generating overall activation map on more complex architectures, Grad-CAM follows a heuristic design without a clear meaning for the generated map. When we calculate the activation map based on Grad-CAM, what do we get? In [12], it has shown that Grad-CAM is equivalent to CAM on GAP based architecture, while it is not true on other architectures. For a classification architecture with flattened feature and one FC layer, the prediction score for class c is formulated as $S_c = \sum_{i,j,k} W_{i,j,k,c} A_{i,j,k}$, where W is the reshaped weights of the FC layer following the shape of the last convolutional feature (A). From our activation decomposition perspective, the contribution of each position (i, j) is clearly given by $\sum_k W_{i,j,k,c} A_{i,j,k}$, while the Grad-CAM map for class c at (i, j) position is given by:

$$\begin{aligned} GradCAM_{i,j,c} &= \sum_k A_{i,j,k} GAP\left(\frac{\partial S_c}{\partial A_k}\right) \\ &= \sum_k A_{i,j,k} GAP\left(W_k\right). \end{aligned} \tag{8}$$

In this case, the result of Grad-CAM is different from activation decomposition because of the GAP operation in Eq. 8. For architectures using GMP, they are also different, because only the maximal value of each channel contributes to the overall prediction score with the activation decomposition framework, while Grad-CAM puts the same weight for features at all positions. Concretely, for a classification architecture with GMP and one FC layer ($W \in \mathbb{R}^{p\times l}$ and no bias), the prediction score for class c is given by $S_c = \sum_k max_{i,j}(A_{:,:,k}) W_{k,c} = \sum_{i,j,k} A_{i,j,k} M_{i,j,k} W_{k,c}$, where M is the maximum matrix in Eq. 7. The decomposition of position (i, j) is $\sum_k A_{i,j,k} M_{i,j,k} W_{k,c}$ and only the maximum position of each channel has a nonzero value. However, for Grad-CAM, the activation map is given by:

$$\begin{aligned} GradCAM_{i,j,c} &= GAP\left(\frac{\partial S_c}{\partial A}\right) A_{i,j} \\ &= \frac{1}{mn} \sum_k A_{i,j,k} W_{k,c}, \end{aligned} \tag{9}$$

where all (i, j) positions have a nonzero weight resulting in a more scattered activation map.

Although [12] empirically shows that the heuristic GAP step can help improve the performance of object localization, this step makes the meaning of the generated map unclear. By removing this step, which means combining the gradient and feature directly as $A_{i,j,k} \frac{\partial S_c}{\partial A_{i,j,k}}$, the modified version would generate the same result as activation decomposition for flattened and GMP based architectures. As for the metric learning architecture discussed in Section IV-A, the Grad-CAM map of query image is written as (*the detailed derivation is presented in Appendix A*):

$$\begin{aligned} GradCAM_{i,j} = \frac{1}{Z} \Bigg(&\frac{\partial (E^q/|E^q|)}{\partial E^q} GAP(W^q) A^q_{i,j} \Bigg) \\ &\cdot \underbrace{\left(\sum_{x,y} W^r_{x,y} A^r_{x,y} + B^r \right)}_{E^r}. \end{aligned} \tag{10}$$

E^q and E^r denote the embedding vectors of the query and retrieved images before L2 normalization and Z is the normalization term. The gradient term $\frac{\partial (E^q/|E^q|)}{\partial E^q}$, which comes from the L2 normalization, would put less weights on the dominant channels so that the generated activation map becomes more scattered as shown in Fig. 4 (*please also refer to Appendix A for proof*). It can be removed by calculating gradient from $E^q \cdot E^r$ without L2 normalization, denoted as "Grad-CAM (no norm)". If we remove the gradient term as well as the GAP term, the result is actually equivalent to the overall map of our "Decomposition+Bias" given by the first two terms of Eq. 5 as $(W^q_{i,j} A^q_{i,j}) \cdot (\sum_{x,y} W^r_{x,y} A^r_{x,y} + B^r)$.

VI. Evaluating Overall Activation Map

To validate the advantage of the proposed method, we provide results of human evaluation as well as several practical applications. We first verify the effectiveness of the proposed overall activation map with weakly supervised localization and human evaluation (Sections VI-A and VI-B). We also show how the activation map brings potential insights on the generalization ability of different metric learning models (Section VI-C).

A. Weakly Supervised Localization

Following previous works [9], [12] on visual explanation, we conduct weakly supervised localization experiment in the

[1]The mask is computed once for each input image.

Authorized licensed use limited to: New Jersey Institute of Technology. Downloaded on September 30,2021 at 20:38:34 UTC from IEEE Xplore. Restrictions apply.

TABLE I

WEAKLY SUPERVISED LOCALIZATION ACCURACY (IOU = 0.5) WITH DIFFERENT THRESHOLDS ON CUB VALIDATION SET

Threshold	Grad-CAM	Grad-CAM (no norm)	Decomposition+Bias (ours)	Decomposition (ours)
0.15	**16.71%**	16.83%	23.49%	17.38%
0.2	16.26%	17.06%	32.01%	19.77%
0.3	14.48%	21.59%	**44.94%**	34.92%
0.4	9.43%	35.01%	37.85%	**50.64%**
0.5	4.43%	47.00%	21.99%	45.78%
0.6	1.54%	**48.27%**	9.35%	27.98%
0.7	0.27%	20.90%	2.25%	9.11%

TABLE II

WEAKLY SUPERVISED LOCALIZATION ACCURACY (IOU = 0.5) OF VARIOUS METHODS ON CUB VALIDATION SET

	Method	Accuracy
Classification	CAM [9]	41.0%
Metric Learning	Grad-CAM	16.7%
	Grad-CAM (no norm)	48.3%
	Decomposition+Bias (ours)	44.9%
	Decomposition (ours)	**50.6%**

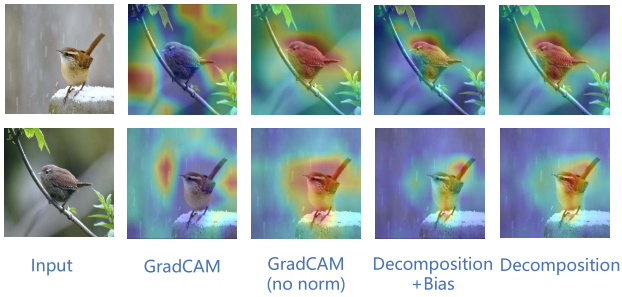


Fig. 4. The activation maps of different approaches on an example image pair from the CUB dataset.

context of metric learning. Among the popular datasets for image retrieval, CUB [33] is a challenging dataset with over 10,000 images from 200 bird species, and the bounding box annotation is available for localization evaluation. We first train a metric learning model on CUB using the state-of-the-art image retrieval method [7]. Then the weakly supervised localization is conducted to show the differences between the variants of the proposed visual explanation method.

We first generate the activation maps of different methods, then the mask is generated by segmenting the heatmap with a threshold. Following previous works [9], [12], the bounding box that covers the largest connected component in the mask is generated as the prediction. A predicted bounding box is considered correct if the IoU (Intersection over Union) between the bounding box and ground-truth is greater than 0.5. Apparently, different thresholds would lead to different predictions, we thus apply a grid search on the threshold to find the best performance of each competing method for a fair comparison. Table I lists the localization accuracy on CUB validation set under different thresholds. For convenience, we report the best result of each method in Table II, including the result of CAM (classification based method) adopted from the original paper [9].

In these two tables, "Grad-CAM (no norm)" means Grad-CAM with gradient computed from the product of two embedding vectors without normalization and "Decomposition+Bias" denotes the first two terms of Eq. 5. Since the architecture of [7] is based on GMP, "Grad-CAM (no norm)" is not equivalent to "Decomposition+Bias" as shown in Section V. The proposed framework outperforms Grad-CAM for metric learning as well as CAM based on classification. As justified in Section V, computing the gradients before normalization does improve the performance of Grad-CAM by a large margin. The qualitative result in Fig. 4 also illustrates the scattering effect of Grad-CAM with normalization as demonstrated in our theoretical analysis (Appendix A). The information in the bias term does not help the weakly supervised localization, as witnessed by the performance of "Decomposition+Bias".

B. Human Evaluation

Following the work of Grad-CAM [12], we also conduct human evaluation on the widely used platform (AMT: Amazon Mechanical Turk) to further evaluate the quality of overall activation maps generated by different methods. Based on Table II, two comparison pairs are formed as "Decomposition" vs. "Grad-CAM" and "Decomposition" vs. "Grad-CAM (no norm)". For each comparison pair, the AMT raters are asked to carefully check the visual explanation generated from two different methods as shown in Fig. 5. Each visual explanation contains an image pair with birds from the same species and the top activated regions (2% pixels) are highlighted based on the overall activation map of one competing method. We denote these two methods as robot A and robot B in the questionnaire. The raters will decide which robot they think is more reasonable in a five-point Likert scale as shown in Fig. 5. The five scales denote method A is "clearly more", "slightly more", "equally", "slightly less", "clearly less" reasonable than method B.

720 evaluations of comparison pairs are collected from at least 40 different AMT workers for two comparison pairs (half for each) and the result is consistent with the outcome of Section VI-A as shown in Fig. 6. Over 60% evaluations support that the proposed method is clearly more reasonable than the vanilla Grad-CAM. Removing the normalization for the gradient computing does improve the quality of Grad-CAM's activation map significantly, but the proposed method still holds more supportive evaluations (41.1% vs. 25.2%) in the AMT test.

C. Model Diagnosis

A large number of loss functions have been proposed for metric learning [1], [2], [7], while only the performance and embedding distribution are evaluated. We show that the overall

Authorized licensed use limited to: New Jersey Institute of Technology. Downloaded on September 30,2021 at 20:38:34 UTC from IEEE Xplore. Restrictions apply.

Fig. 5. Human evaluation interface for evaluating the overall activation maps on CUB.

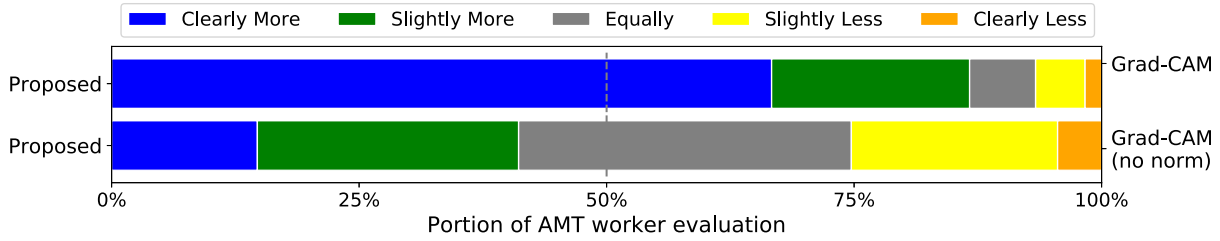


Fig. 6. Human evaluation results on the proposed method vs. Grad-CAM variants. The "clearly more" corresponds to the portion of workers who select the option 1 in Fig. 5, which means they think Robot A is clearly more reasonable than Robot B.

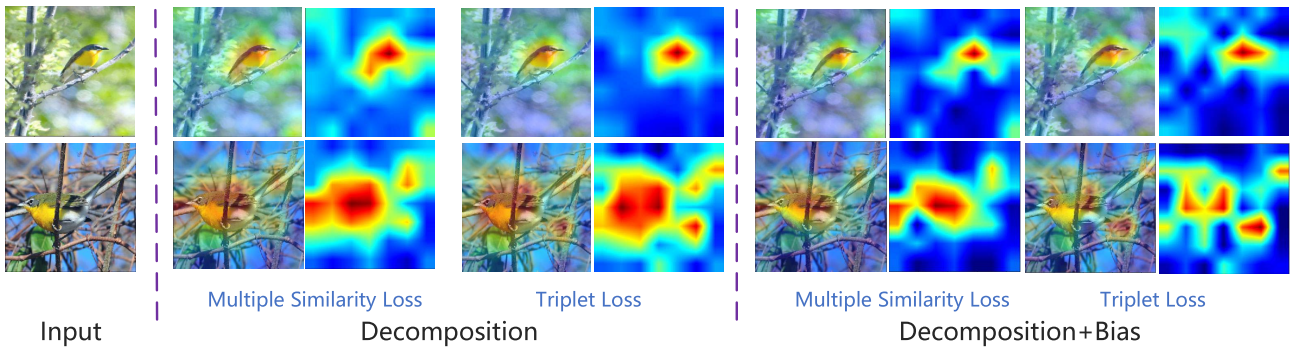


Fig. 7. Qualitative results (overall activation maps corresponding to different metric learning losses) for model diagnosis. Although the metric learning models with two different losses yield similar training accuracies (82.69% and 82.98% in Table III), their overall activation maps generated by the activation decomposition framework reveal great differences in the regions to which the models pay attention. These activation maps also provide valuable insights on the model generalization ability to the validation/test set.

TABLE III
TOP-1 RECALL ACCURACY ON CUB

Loss	Train	Validation
MS	82.69%	65.45%
Triplet	82.98%	60.55%

TABLE IV
LOCALIZATION ACCURACY (IOU = 0.5) ON CUB TRAINING SET

	Decomposition+Bias	Decomposition
MS	52.10%	58.99%
Triplet	39.93%	55.60%
Accuracy Drop	12.17%	3.39%

activation map can also help evaluate the generalization ability of different metric learning methods.

We follow the setting of [7] and train two metric learning models with Multiple Similarity (MS) loss [7] and Triplet loss [2], respectively. As shown in Table III, although they have almost the same accuracy on the training set, there is a big gap between their generalization abilities on the validation set. *Before checking the result on the validation set, is there any clue in the training set for such gap?* Despite the similar training accuracies, the predictions of

Authorized licensed use limited to: New Jersey Institute of Technology. Downloaded on September 30,2021 at 20:38:34 UTC from IEEE Xplore. Restrictions apply.

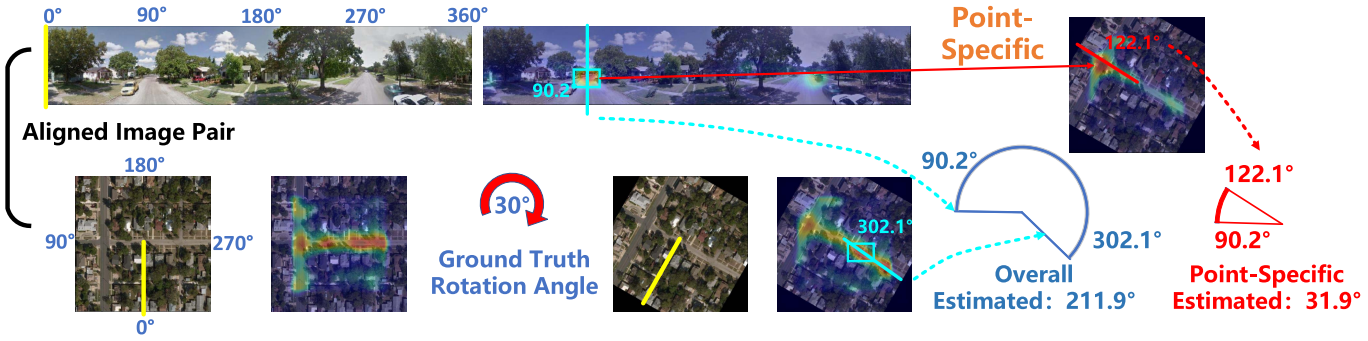


Fig. 8. An example of illustrating cross-view pattern discovery – image orientation estimation – by using the activation maps from the metric learning (i.e. image matching/retrieval) model. Based on the overall activation maps from two views, the estimated orientation angle between two views is 211.9°. By generating the point-specific activation map of a maximum activated region in the other view, the estimated orientation is much more accurate, i.e. 31.9°, with only 1.9° of angle difference as compared with the ground truth in this example.

two models are actually based on different regions from the activation maps generated by our activation decomposition framework (see Fig. 7). The activation maps also provide valuable insights on the model generalization ability to the validation/test set. It is evident from Fig. 7, the model with MS loss is more focused on the salient object area than the one with Triplet loss, which leads to better validation performance as confirmed in Table III. The quantitative results of localization in Table IV also support the fact that "Triplet" model is more likely to focus on the background region rather than the object as we witness the localization accuracy drop for Triplet loss. Although "Decomposition+Bias" can be more sensitive to different loss functions, implying that the bias term does provide valuable information on the overall map, both settings of the proposed method have the same trend on accuracy. Therefore, our activation decomposition framework can shed light on the generalization ability of loss functions for metric learning, thereby providing an useful tool for model diagnosis.

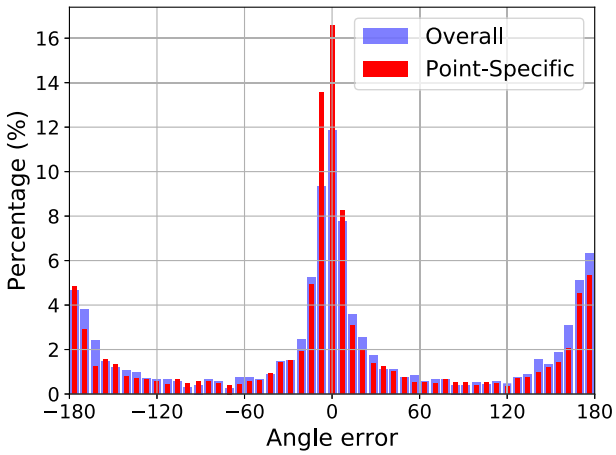


Fig. 9. The results of cross-view image orientation estimation (i.e. angle error distributions) obtained by the overall and point-specific maps based on the image matching experiment on the CVUSA dataset.

VII. Evaluating Point-Specific Activation Map

The proposed activation decomposition framework not only generates the overall activation map, but also can produce the point-specific activation map for fine-grained visual explanation of metric learning. In this section, we introduce two applications to demonstrate the importance of the proposed point-specific activation map.

A. Application I: Cross-View Pattern Discovery

When metric learning is applied to cross-view applications, e.g. image retrieval [4], [34], [35], the model is capable of learning similar patterns in different views which may provide geometric information of the two views, e.g. the camera pose or orientation information. We conduct orientation estimation experiment to show the advantage of point-specific activation map compared with the overall activation map on providing geometric information based on cross-view patterns. In our experiment, we take street-to-aerial image geo-localization [4], [34] as an example.

The objective of cross-view image geo-localization is to find the best matched *aerial-view image* (with GPS information) in a reference dataset for a query *street-view* image. We conduct the experiment on CVUSA [34], which is the most popular benchmark for this problem containing 35,532 training pairs and 8,884 test pairs. Each pair consists of a query street image and the corresponding **orientation-aligned** aerial image at the same GPS location. For example, as shown in Fig. 8, the left two images (street-view panorama and aerial-view images) are the original aligned image pair and the yellow line denotes the 0° which corresponds to the South direction (180° corresponds to the North direction). The location of panorama always lies at the center of the corresponding aerial image. We use [0, 360]° to denote different angles as marked on those two images.

We train a simple Siamese-VGG [18], [36] network following the modified triplet loss in [4]. During the training, if the image pairs are not aligned (e.g. randomly rotate the aerial view images), the activation map can be used for orientation estimation. Specifically, we train the Siamese-VGG with randomly rotated aerial images so that the model is approximately rotation-invariant and so is the overall activation map for aerial image. In the example of Fig. 8, the overall

Authorized licensed use limited to: New Jersey Institute of Technology. Downloaded on September 30,2021 at 20:38:34 UTC from IEEE Xplore. Restrictions apply.

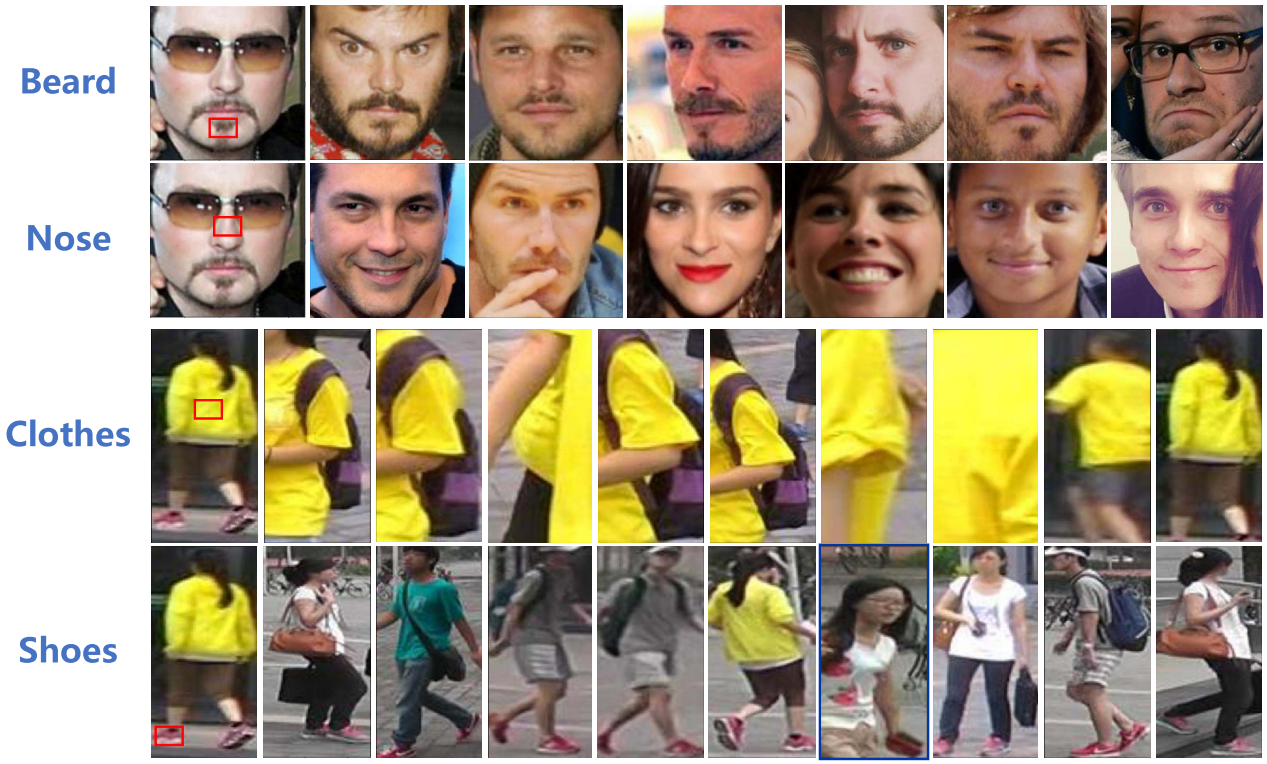


Fig. 10. Top retrieved images by interactive retrieval on face and re-id datasets. The first column shows the query images and the red box on each of them indicates the region of interest (RoI). In the last row, the blue bounding box highlights a failure retrieval case, which is further analyzed in Fig. 13.

activation maps for the original aligned pair are first generated to show the highlighted regions contributing to the similarity score, and both views focus on the road areas in this example. The most activated regions are highly relevant in both views and most of them are similar patterns. When we randomly rotate the aerial image (30° in this example and is denoted as the ground truth rotation angle in Fig. 8), the aerial view activation map still focuses on the road area which is relevant to the street view.

In this paper, we simply use the pixel with the maximum activation value for orientation estimation, because the most activated areas (highlighted in cyan boxes in Fig. 8) from two views are likely to be relevant. For the street-view image, the selected pixel lies in the angle of 90.2° (the cyan line) based on the angle marks on the left aligned image. For the rotated aerial view, the selected pixel lies in the angle of 302.1° (the cyan line). Since the overall map contains multiple activated regions, the selected (maximum activation) pixels in both views actually do not correspond to the same object, which reveals one disadvantage of using only the overall activation map. In this example, the estimated angle is $302.1° - 90.2° = 211.9°$, which is not correct. However, for the selected pixel (the one with the maximum activation value) in the street-view image, if we generate its corresponding point-specific activation map on the aerial-view image (as shown on the very right of Fig. 8), the new selected pixel (maximum activation) on this point-specific activation map lies in the angle of 122.1° (the red line). Then, the estimated angle is calculated as $122.1° - 90.2° = 31.9°$, which is very close to the ground truth (30°). This demonstrates the advantage of the point-specific activation decomposition for finding more fine-grained information in this application.

We also present the quantitative results in Fig. 9. The angle error is computed by `error = ground truth - estimated angle`. We then add or subtract 360° to the error to fit it in the range of $[-180, 180]°$ if the absolute value of the error is greater than 180°. For example, when the error is 359°, we subtract it by 360° and get −1° as the error. This is a reasonable setting adopted from the previous work [34]. As evident from Fig. 9, the point-specific activation map significantly outperforms the overall map for cross-view orientation estimation. Point-specific activation based orientation estimation has over 16% samples (a total of 8,884 test sample pairs in the CVUSA dataset) with angle error less than ±3.5° (the red bar at 0° in Fig. 9), while overall activation map based method has less than 12%. The result further validates the superiority of the point-specific map compared with the overall map. Moreover, we also provide a demo[2] to show how the point-specific map changes according to the query pixel. Failure samples of both methods tend to have an angle error of 180 degree, because the activation maps of both views are likely to focus on roads which are symmetric in aerial images.

B. Application II: Interactive Retrieval

Verification applications like face recognition and person re-id usually retrieve the most similar images to the query image. *However, we may be more interested in a specific*

[2]Please check the demo (the "Geo-localization" section) at https://github.com/Jeff-Zilence/Explain_Metric_Learning

Authorized licensed use limited to: New Jersey Institute of Technology. Downloaded on September 30,2021 at 20:38:34 UTC from IEEE Xplore. Restrictions apply.

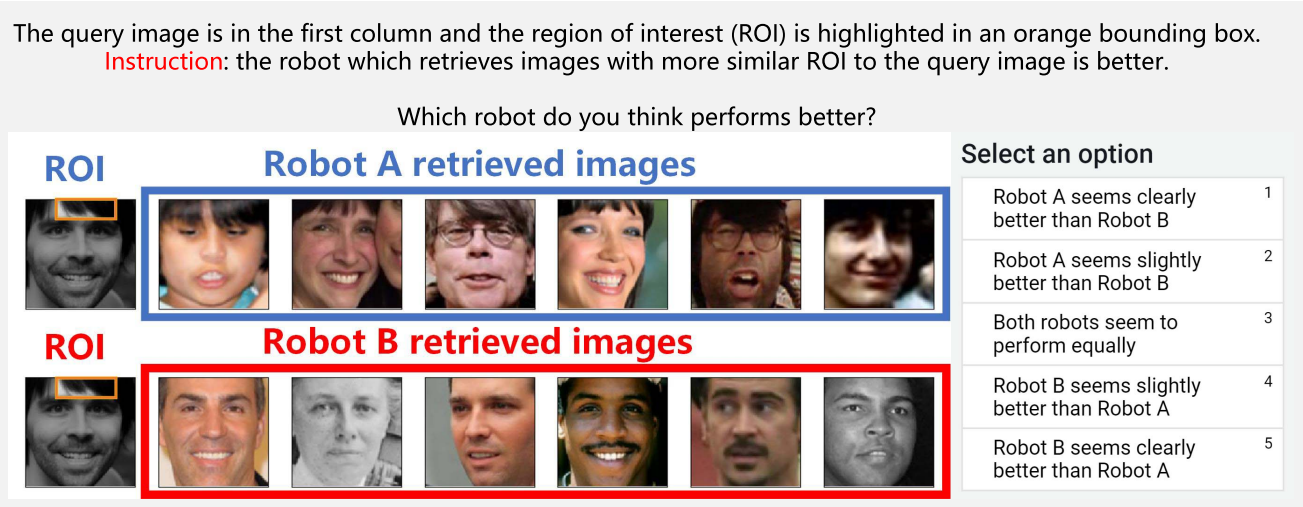


Fig. 11. Human evaluation interface for interactive retrieval.

RoI (region of interest) than the overall image for certain circumstances, for example finding people with a similar bag or pair of shoes (i.e. RoI) in surveillance applications. Since our framework provides the point-to-point activation map, it can be a reasonable tool for measuring partial similarity to serve this purpose. With the partial similarity, we are able to *interactively* retrieve images with only parts/regions similar to the query image instead of the *overall* most similar images. As shown in Fig. 10, the point-specific activation generated by the proposed method works well on retrieving people with similar clothes and faces with similar beard. **No explicit supervision** is adopted except the pre-trained retrieval model.

Specifically, we follow the pipeline of recent approaches on face recognition [31] and person re-id [3]. For re-id, the model is trained and evaluated on Market- 1501 [37] where some validation images only contain a small part of a person. Although the model is trained with Euclidean distance without L2 normalization, cosine similarity still works well as the evaluation metric. For face recognition, we take the trained model on CASIA-WebFace [38] and evaluate it on FIW [39] where the face identification and kinship relationship are available.

The interactive search is conducted by simply matching the equivalent partial feature in Eq. 5 ($W^q_{i,j}A^q_{i,j}$) with the reference embedding features. We first compute $W^q_{i,j}A^q_{i,j}$ as the feature of each position in the last convolutional layer, and a bilinear interpolation is adopted to generate the feature for every pixel of the original image. For a point of interest (i, j), we compute the cosine similarity between the calculated feature on (i, j) and the embedding feature of the reference images as the point-specific similarity (note that the equivalent feature is normalized with the l_2 norm of the original embedding as $|E^q|$). In this case, the embedding features of the reference dataset do not have to be recomputed. This similarity can be also considered as the summation of the values in the point-specific map corresponding to (i, j). In the case of Eq. 2 (CNN+GAP), the similarity can be simplified as $\sum_{x,y}(\sum_k A^q_{i,j,k}A^r_{x,y,k})/|E^q||E^r|$ (E^q and E^r denote the original embedding features of the query and retrieved images). If the ROI is very large, we can merge multiple equivalent features in the ROI to generate better result. Since all the objects have similar size in the face and re-id datasets, we simply adopt the equivalent feature of the center pixel in the ROI.

Apparently, the images retrieved by the original feature are likely to be from the same identity as the query image, because the models are trained for identification task. However, the proposed interactive retrieval is able to retrieve different images based on the same query image with different RoIs as shown in Fig. 10. To further quantitatively evaluate the advantage of the proposed point-specific map, we conduct human evaluation on the overall retrieval (based on the original feature) vs. the interactive retrieval (based on the point-specific map). Similar to Section VI-B, the AMT workers are asked to decide which retrieval result is better given a specific RoI (see Fig. 11). We conduct experiments on both face verification and person re-id tasks. 360 evaluations of comparison pairs in total are collected from at least 40 different AMT workers. As shown in Fig. 12, the interactive retrieval holds considerably more supportive evaluations than the overall retrieval (57.5% vs. 28.0% on face, and 47.5% vs. 31.2% on person re-id), which demonstrates the effectiveness of the proposed point-specific activation map.

However, we do find some failure cases as shown in the last row of Fig. 10 with the blue bounding box. There is no shoes in the failure image, but why is this image retrieved with a top rank (i.e. 6th rank)? We provide an in-depth analysis from the perspectives of both overall and point-specific activation maps in Fig. 13. In the overall activation map, three regions of the query image have a high activation on the retrieved image. From the point-specific activation maps, there is a high activation on the purse in the retrieved image corresponding to the shoes in the query image. This might be because of their similar red color and the arm of the retrieved person may appear like a leg from a specific viewpoint. This example also validates the importance of the point-specific activation map generated by our framework for explanation.

Authorized licensed use limited to: New Jersey Institute of Technology. Downloaded on September 30,2021 at 20:38:34 UTC from IEEE Xplore. Restrictions apply.

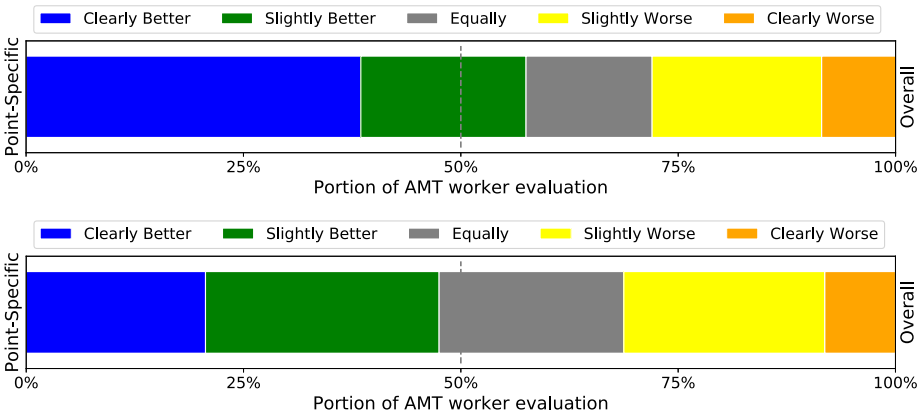


Fig. 12. Human evaluation results on the point-specific (interactive) vs. overall retrieval for face verification (first row) and person re-identification (second row). The "clearly better" corresponds to the portion of workers who select the option 1 in Fig. 11, which means they think Robot A performs clearly better than Robot B.

Fig. 13. Explanation of the failure case by the point-specific activation map.

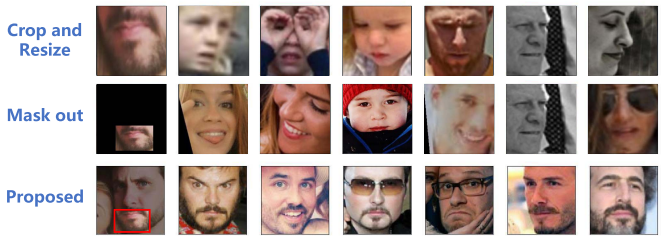


Fig. 14. Comparison between our method and cropping/masking for interactive retrieval on face verification. Query image is in the first column.

Fig. 15. Comparison between our method and cropping/masking for interactive retrieval on person re-identification. Query image is in the first column.

1) Comparison with Cropping/Masking: One may argue that another possible way to achieve interactive retrieval is to crop or mask out a specific RoI. Then we can generate the feature with the specific part of image and use this feature as query to retrieve images. However, generally speaking, this operation will cause the feature to be very different from features of the original images, thus would not generate satisfactory retrieval results. And we validate this by providing the qualitative comparison of retrieval results in Figs. 14 and 15.

For face verification, flattened features followed by FC layers are usually used, because the input image size is relatively small and flattened features reserve all the information in the last convolutional layer. However, this architecture does not allow change in the input size, so one may need to resize the image to the original size after cropping a specific RoI. Another way is to assign 0 for all pixels outside the RoI. The comparison in Fig. 14 shows that cropping with resizing or masking are not able to generate satisfactory results.

For person re-identification, GAP is usually used so that we do not need to resize the cropped image. The results in Fig. 15 also show the superiority of the proposed method over cropping/masking.

VIII. Conclusion

We propose a simple yet effective framework for visual explanation of deep metric learning based on the idea of activation decomposition. The framework is applicable to a host of applications, e.g. image retrieval, face recognition, person re-id, geo-localization, etc. Experiments show the importance of visual explanation for metric learning as well as the superiority of both the overall and point-specific activation maps generated by the proposed method. Furthermore, we introduce two applications, i.e. cross-view pattern discovery and interactive retrieval, which reveal the importance of the

Authorized licensed use limited to: New Jersey Institute of Technology. Downloaded on September 30,2021 at 20:38:34 UTC from IEEE Xplore. Restrictions apply.

Fig. 16. Human evaluation interface for interactive retrieval on person re-identification.

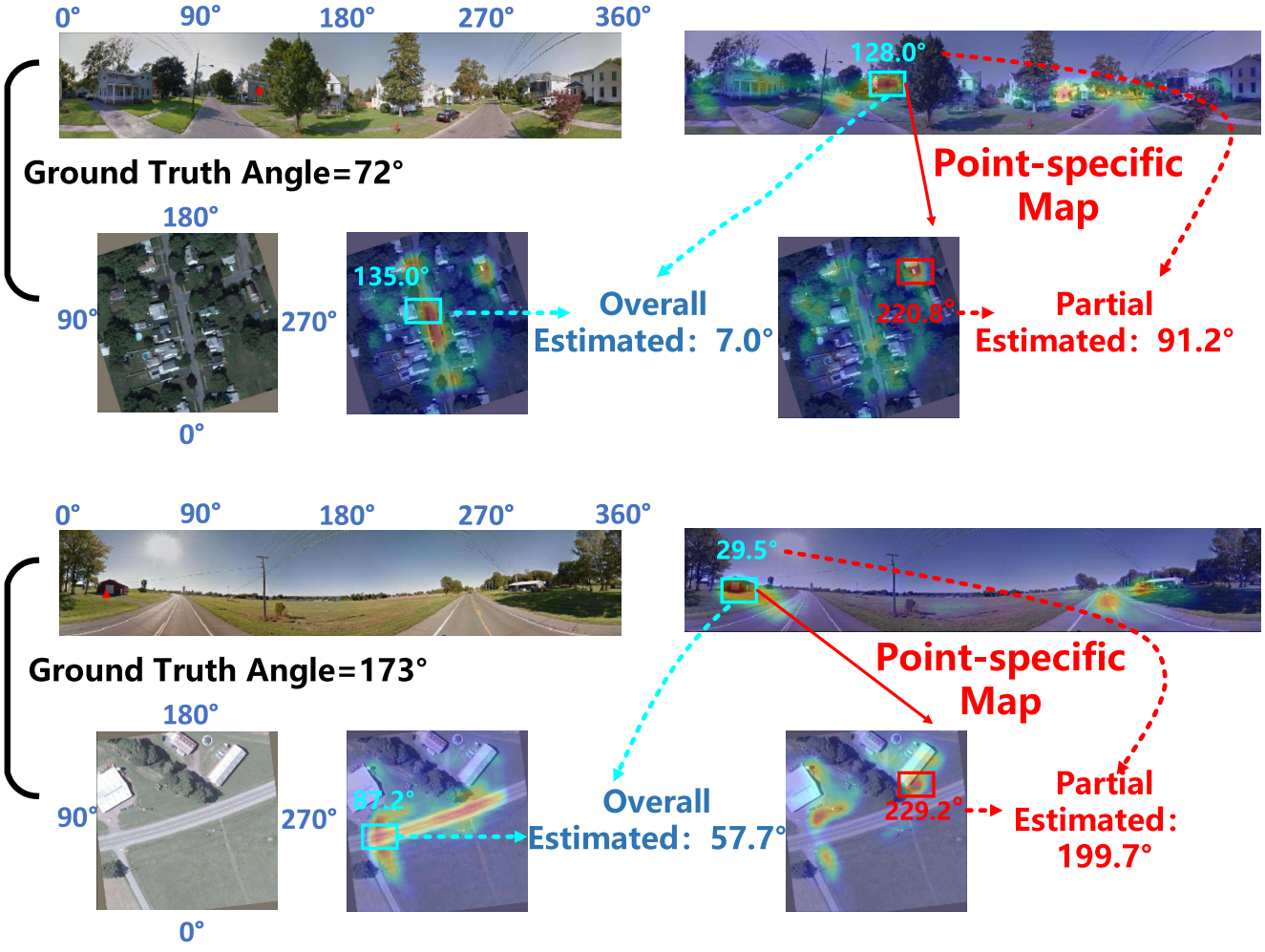


Fig. 17. Examples of cross-view pattern discovery, i.e., image orientation estimation.

point-specific activation map for explanation. Our work also points to interesting directions in exploring the point-specific activation maps for fine-grained information discovery and analysis.

Authorized licensed use limited to: New Jersey Institute of Technology. Downloaded on September 30,2021 at 20:38:34 UTC from IEEE Xplore. Restrictions apply.

Fig. 18. Top retrieved images by interactive retrieval on face recognition. Red boxes on the query images (first column) highlight the Region of Interest (RoI).

Fig. 19. Top retrieved images by interactive retrieval on person re-identification. Red boxes on the query images (first column) highlight the Region of Interest (RoI).

APPENDIX A
GRAD-CAM FOR METRIC LEARNING

For metric learning architecture discussed in Section IV-A of the main paper, the similarity is formulated as $S = \frac{E^q \cdot E^r}{|E^q||E^r|}$, where $E^q \in \mathbb{R}^l$ and $E^r \in \mathbb{R}^l$ are the embedding vectors of the query and retrieved images. $|x|$ denotes the L2 norm and $a \cdot b$ is the inner product of a and b. The Grad-CAM map of the query image is given by:

$$GradCAM_{i,j} = GAP\left(\frac{\partial S}{\partial A^q}\right) A^q_{i,j} \tag{11}$$

With the gradient chain rule, the gradient is written as:

$$\frac{\partial S}{\partial A^q} = \frac{\partial \frac{(E^r)^T E^q}{|E^q||E^r|}}{\partial A^q} = \left(\frac{E^r}{|E^r|}\right)^T \frac{\partial (E^q/|E^q|)}{\partial E^q} \frac{\partial E^q}{\partial A^q}$$

$$= \left(\frac{E^r}{|E^r|}\right)^T \frac{\partial (E^q/|E^q|)}{\partial E^q} W^q \tag{12}$$

By expanding E^r as $\sum_{x,y} W^r_{x,y} A^r_{x,y} + B^r$ (Section IV-A) and merging Eq. 11 with Eq. 12, the Grad-CAM map is reformulated as:

$$GradCAM_{i,j}$$
$$= GAP\left(\left(\frac{E^r}{|E^r|}\right)^T \frac{\partial (E^q/|E^q|)}{\partial E^q} W^q\right) A^q_{i,j}$$
$$= \frac{1}{Z}\left(\sum_{i^*,j^*} (E^r)^T \frac{\partial (E^q/|E^q|)}{\partial E^q} W^q_{i^*,j^*}\right) A^q_{i,j}$$
$$= \frac{1}{Z}(E^r)^T \frac{\partial (E^q/|E^q|)}{\partial E^q}\left(\sum_{i^*,j^*} W^q_{i^*,j^*}\right) A^q_{i,j}$$

Authorized licensed use limited to: New Jersey Institute of Technology. Downloaded on September 30,2021 at 20:38:34 UTC from IEEE Xplore. Restrictions apply.

$$= \frac{1}{Z}\left(\sum_{x,y} W^r_{x,y} A^r_{x,y} + B^r\right) \cdot \left(\frac{\partial(E^q/|E^q|)}{\partial E^q} GAP(W^q) A^q_{i,j}\right)$$
$$= \frac{1}{Z}\left(\frac{\partial(E^q/|E^q|)}{\partial E^q} GAP(W^q) A^q_{i,j}\right) \cdot \left(\sum_{x,y} W^r_{x,y} A^r_{x,y} + B^r\right) \quad (13)$$

Here the Z is the normalization term for simplicity. $\frac{\partial E/|E|}{\partial E}$ is the $l \times l$ Jacobian matrix given by:

$$\frac{\partial(E/|E|)}{\partial E} = \left(\frac{\partial(E_i/|E|)}{\partial E_j}\right)_{i,j} = \begin{cases} \frac{1}{|E|}\left(1 - \frac{E_i^2}{|E|^2}\right) & i = j \\ -\frac{E_i E_j}{|E|^3} & i \neq j \end{cases} \quad (14)$$

$\frac{1}{|E|}$ is the normalization term. The $\frac{\partial E/|E|}{\partial E}$ term can be removed, if we compute the gradient from $E^q \cdot E^r$ instead of $\frac{E^q \cdot E^r}{|E^q||E^r|}$. For the dominant channel i, the weight $(1 - \frac{E_i^2}{|E|^2})$ is small resulting in a more scattered activation map.

Appendix B
Qualitative Results

In Fig. 16, we show an example of human evaluation interface for person re-identification. We also provide extra qualitative results for cross-view pattern discovery, interactive retrieval (both face verification and person re-identification) in Figs. 17, 18, 19.

References

[1] H. O. Song, Y. Xiang, S. Jegelka, and S. Savarese, "Deep metric learning via lifted structured feature embedding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4004–4012.

[2] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.

[3] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1–9.

[4] S. Hu, M. Feng, R. M. H. Nguyen, and G. H. Lee, "CVM-Net: Cross-view matching network for image-based ground-to-aerial geo-localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7258–7267.

[5] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr, "Fully-convolutional Siamese networks for object tracking," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 850–865.

[6] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 539–546.

[7] X. Wang, X. Han, W. Huang, D. Dong, and M. R. Scott, "Multi-similarity loss with general pair weighting for deep metric learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5022–5030.

[8] J. Tobias Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," 2014, *arXiv:1412.6806*. [Online]. Available: http://arxiv.org/abs/1412.6806

[9] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.

[10] B. Zhou, D. Bau, A. Oliva, and A. Torralba, "Interpreting deep visual representations via network dissection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 9, pp. 2131–2145, Sep. 2019.

[11] R. C. Fong and A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3429–3437.

[12] 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 *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[13] J. Zhang, S. A. Bargal, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff, "Top-down neural attention by excitation backprop," *Int. J. Comput. Vis.*, vol. 126, no. 10, pp. 1084–1102, Oct. 2018.

[14] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5188–5196.

[15] W. Yang, H. Huang, Z. Zhang, X. Chen, K. Huang, and S. Zhang, "Towards rich feature discovery with class activation maps augmentation for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1389–1398.

[16] A. Gordo and D. Larlus, "Beyond instance-level image retrieval: Leveraging captions to learn a global visual representation for semantic retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6589–6598.

[17] B. Zhou, Y. Sun, D. Bau, and A. Torralba, "Interpretable basis decomposition for visual explanation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 119–134.

[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[19] A. Stylianou, R. Souvenir, and R. Pless, "Visualizing deep similarity networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 2029–2037.

[20] L. Chen, J. Chen, H. Hajimirsadeghi, and G. Mori, "Adapting grad-CAM for embedding networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2794–2803.

[21] M. Ye, J. Shen, X. Zhang, P. C. Yuen, and S.-F. Chang, "Augmentation invariant and instance spreading feature for softmax embedding," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Aug. 3, 2020, doi: 10.1109/TPAMI.2020.3013379.

[22] M. Ye and J. Shen, "Probabilistic structural latent representation for unsupervised embedding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5457–5466.

[23] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jan. 26, 2021, doi: 10.1109/TPAMI.2021.3054775.

[24] M. Ye, X. Lan, Q. Leng, and J. Shen, "Cross-modality person re-identification via modality-aware collaborative ensemble learning," *IEEE Trans. Image Process.*, vol. 29, pp. 9387–9399, 2020.

[25] M. Ye, J. Shen, and L. Shao, "Visible-infrared person re-identification via homogeneous augmented tri-modal learning," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 728–739, 2020.

[26] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 9505–9515.

[27] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 839–847.

[28] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "SphereFace: Deep hypersphere embedding for face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 212–220.

[29] H. Wang *et al.*, "CosFace: Large margin cosine loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5265–5274.

[30] E. Ristani and C. Tomasi, "Features for multi-target multi-camera tracking and re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6036–6046.

[31] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4690–4699.

[32] S. Zhu, T. Yang, and C. Chen, "VIGOR: Cross-view image geo-localization beyond one-to-one retrieval," 2020, *arXiv:2011.12172*. [Online]. Available: http://arxiv.org/abs/2011.12172

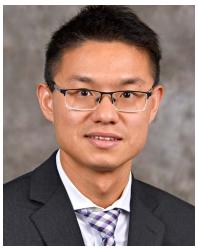
[33] P. Welinder *et al.*, "Caltech-UCSD birds 200," California Inst. Technol., Pasadena, CA, USA, Tech. Rep. CNS-TR-2010-001, 2010.

Authorized licensed use limited to: New Jersey Institute of Technology. Downloaded on September 30,2021 at 20:38:34 UTC from IEEE Xplore. Restrictions apply.

[34] M. Zhai, Z. Bessinger, S. Workman, and N. Jacobs, "Predicting ground-level scene layout from aerial imagery," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 867–875.
[35] Y. Tian, C. Chen, and M. Shah, "Cross-view image matching for geo-localization in urban environments," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3608–3616.
[36] S. Zhu, T. Yang, and C. Chen, "Revisiting street-to-aerial view image geo-localization and orientation estimation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 756–765.
[37] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1116–1124.
[38] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," 2014, *arXiv:1411.7923*. [Online]. Available: http://arxiv.org/abs/1411.7923
[39] J. P. Robinson, M. Shao, Y. Wu, H. Liu, T. Gillis, and Y. Fu, "Visual kinship recognition of families in the wild," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 11, pp. 2624–2637, Nov. 2018.

Sijie Zhu received the B.E. degree from the University of Science and Technology of China in 2015 and the M.S. degree from the University of Chinese Academy of Sciences in 2018. He is currently pursuing the Ph.D. degree with the Center for Research in Computer Vision, University of Central Florida. His research interests include visual representation learning and computer vision applications.

Taojiannan Yang (Graduate Student Member, IEEE) received the bachelor's degree from the University of Science and Technology of China in 2018. He is currently pursuing the Ph.D. degree with the Center for Research in Computer Vision, University of Central Florida. His research interests include computer vision and deep learning.



Chen Chen (Member, IEEE) received the Ph.D. degree from the Department of Electrical Engineering, University of Texas at Dallas, Richardson, TX, USA, in 2016. He is currently an Assistant Professor with the Center for Research in Computer Vision, University of Central Florida. His research interests include computer vision, deep learning, and image and video processing. He published over 80 papers in refereed journals and conferences in these areas. He is an Associate Editor of *Journal of Real-Time Image Processing* and IEEE JOURNAL ON MINIATURIZATION FOR AIR AND SPACE SYSTEMS.

Authorized licensed use limited to: New Jersey Institute of Technology. Downloaded on September 30,2021 at 20:38:34 UTC from IEEE Xplore. Restrictions apply.