IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 31, 2019, accepted October 2, 2019, date of publication October 31, 2019, date of current version November 15, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2950701

Trigraph Regularized Collective Matrix Tri-Factorization Framework on Multiview Features for Multilabel Image Annotation

JUNYI ZHANG[1,2], YUAN RAO[1], JULI ZHANG[3], AND YONGQIANG ZHAO[1]

[1]The Lab of Social Intelligence and Complex Data Processing, School of Software Engineering, Xi'an Jiaotong University, Xi'an 710049, China
[2]China Minsheng Bank, Beijing 110000, China
[3]Department of Research and Development, Xi'an Microelectronics Technology Institute, Xi'an 710068, China

Corresponding author: Yuan Rao (raoyuan@mail.xjtu.edu.cn)

This work was supported in part by the World-Class Universities (Disciplines) and the Characteristic Development Guidance Funds for the Central Universities under Grant PY3A022, in part by the Shenzhen Science and Technology Project under Grant JCYJ20180306170836595, in part by the National Natural Science Fund of China under Grant F020807, in part by the Ministry of Education Fund Project\Cloud Number Integration Science and Education Innovation under Grant 2017B00030, in part by the Basic Scientific Research Operating Expenses of Central Universities under Grant ZDYF2017006, and in part by the Shaanxi Provincial Science and Technology Department Collaborative Innovation Project under Grant 2015XT-21.

ABSTRACT Due to the explosive growth of image data, image annotation has been one of the most popular research directions in computer vision. It has been widely used in image retrieval, image analysis and understanding. Because traditional manual image annotation is time consuming, more advanced automatic annotation methods are needed. A major challenge in developing an efficient image annotation method is how to effectively use all available information contained in the data. To this end, this paper proposes a novel image annotation framework that uses multiple information from data. It employs nonnegative matrix tri-factorization (NMTF) to simultaneously factorize image-to-label, image-to-feature, and feature-to-label relation matrices using their intertype relationships and incorporates the intratype information through manifold regularizations. This method can be referred to as the trigraph regularized collective matrix tri-factorization framework (TG-CMTF). TG-CMTF captures the correlations among different labels, different images and different features. By taking advantage of these relations from images, features and labels, TG-CMTF can achieve better annotation performance than most state-of-the-art methods. The promising experimental results on three standard benchmarks have shown the effectiveness of this information. Furthermore, we show the annotation process as a precise optimization problem and solve it by an iterative algorithm, which proves the correctness of the proposed method from the mathematical theory.

INDEX TERMS Image annotation, nonnegative matrix tri-factorization, manifold regularization, multilabel.

I. INTRODUCTION

The rapid growth of the Internet and modern technologies has exponentially increased accessible digital images. Among these images, most of them are unlabeled. Accessing and retrieving this considerable number of unlabeled images is rather difficult. Therefore, annotating unlabeled images is beneficial for vision-based tasks, such as image retrieval and image classification. However, traditional manual image annotation methods are time consuming and tedious. Some researchers have devoted efforts to associating images with several labels through their similarities, which is called automatic image annotation, which formulates the annotation task as a text retrieval problem and has both high accuracy and efficient computation. Due to these advantages, automatic image annotation has become a very popular subject of research [1]–[3]. However, there are several challenges in this topic. Thus, effective image annotation is needed.

Existing content-based image annotation methods [4]–[6] usually extract visual features from images and then predict the related labels. These methods rely heavily on low-level visual content. Different visual features capture different aspects or views of the image, thereby providing different information. However, if each feature represents the same

The associate editor coordinating the review of this manuscript and approving it for publication was Eduardo Rosa-Molinar.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see http://creativecommons.org/licenses/by/4.0/

image, they all capture the same underlying latent structure. There are some inherent relations in these features. Nevertheless, most of the visual-based annotation methods disregard the consistencies among different views. Some of them combine several types of features [7], [8] into one feature vector, and then compare it with other images. This causes the dimensionality problem. Moreover, there is another well-known problem, semantic gap [9] between low-level features and high-level semantics. This occurs because the low-level visual features cannot abstract the visual content of the image very well. Thus, multiview features have been considered by many researchers to improve the annotation performance. Recently, deep convolutional neural networks [10]–[12] have shown significant improvements in vision tasks. Considering each view describes one aspect of images, we not only extract low-level visual features but also introduce convolutional neural networks (CNN) features as one view of the visual features in our method.

To effectively annotate images, some researchers have been devoted to exploring available label information, such as label-label correlation (Coherent Language Model [13] and WordNet-based method [14]) and image-label relation. These works prove that adding image semantics or visual content can also improve image annotation performance. Thus, semantic and visual contents are both used in the proposed method to boost the annotation performance. However, how to use such information is vital for the task. Most of the previous annotation approaches [15], [16] only rely on intratype relationships, i.e., image-image and label-label, and rarely use intertype relationships, such as image-to-label. However, we utilize intertype relationships also. The aim is to explore both intertype relationships and intratype information to minimize the semantic gap, which also maximizes annotation performance. The rich structures of multitype relational data provide a potential opportunity to improve the annotation performance. To go one step further, we annotate images using visual features, label semantics, and relations between images and labels. To achieve this, this method utilizes these relations in the joint relational model by a collective matrix factorization framework, which is a novel framework for image annotation.

Nonnegative matrix factorization (NMF) [17] is used to discover the latent structures embedded in the data, in which the negative data matrix is approximated by two lower dimension nonnegative matrices. Although these NMF-based methods have achieved considerable success in different fields such as image processing [18], image classification [19] and image retrieval [20], the performances of these approaches are not desirable since they do not consider the geometrical structure of samples. Fortunately, Laplacian graphs [21] can encode the geometrical information. In the graph, images are considered nodes and weights are their edges. In [22], the Laplacian regularization term was first introduced in NMF, which encodes the intrinsic geometric information contained in the data. To exploit multiple relations described above in the terms of Laplacian graphs, we use a trigraph regularization NMF framework to solve the image annotation problem. First, we extract the multiple visual features for each image by a variety of methods, including handcraft features and deep CNN features. Then, we construct the relationships of intraimages, intrafeatures and intralabels. To fully exploit all available information, we build the intertype relations among features, images and labels, respectively. We utilize these relations to discover the hidden global structures of data and underlying data distribution. Then, the labels for testing images can be predicted from these relations. Generally, we use a collective nonnegative matrix tri-factorization (NMTF) [23] to decompose these relations by graph regularizations. This method explores the intertype and intrarelationships simultaneously. Moreover, the view heterogeneity, label heterogeneity, and the images' feature heterogeneity are utilized. These relations of data enhance the information from the data and affect the matrix factorization. Furthermore, the similarity constraint between labels captures the correlations among different labels, the similarity constraint between images captures the correlations among different images, and the features similarity captures the correlations among different views of features. This generates a semantic space in which visual patterns and text terms are represented together. The main idea is described in Figure 1.

As Figure 1 describes, the image-to-feature relation, image-to-label relation and feature-to-label relation are factorized into three low-rank matrices. Then, they are combined with a collective nonnegative matrix factorization (CNMF). To utilize the intertype relationships of images, features and labels, we construct the image graph, feature graph and label graph. These three graphs are used as three regularization terms in the CNMF, which is referred to as trigraph regularized collective matrix tri-factorization (TG-CMTF).

One of the main contributions of the proposed method is finding the meaningful latent factors from each view feature and each relation in such a way that one image has a ''close representation'' in each space, and similar images have similar relations in image-to-feature, image-to-label and feature-to-label. These relations impact each other and impact the final prediction of the image annotation, which aims to narrow the semantic gap by joining the visual and test features of images. Furthermore, it maps different aspects of the image to a common latent semantic space.

It is worthwhile to highlight the main contributions of the proposed method. We summarize them as follows,

- The proposed method utilizes a collective matrix factorization to factorize the multiple relations.
- This method maps different aspects of the image to a common latent semantic space, which can exploit multimodal interactions among images, text annotations and multiview visual features.
- The proposed method introduces an iterative optimization scheme to solve the multigraph regularized collective matrix factorization problem and calculates the time complexity. The experiments are conducted, and

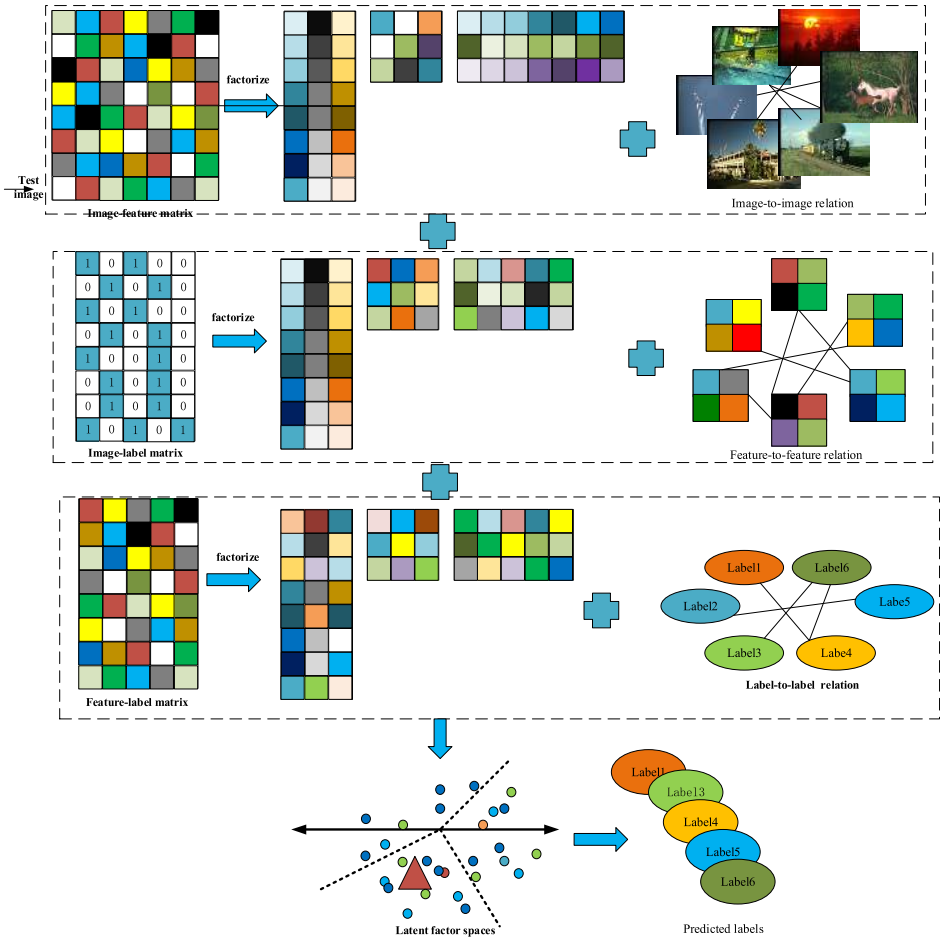


FIGURE 1. The overview of the proposed method.

experimental results on three benchmarks are discussed. Tuning all the parameters is discussed.

The rest of this paper is organized as follows: Section II introduces the related work about automatic image annotation methods and the multiview feature learning methods. We describe the theoretical formula and the optimization scheme of the proposed method in Section III. In Section IV, we conduct extensive experiments and discuss the performance. Furthermore, we analyze the parameters used in this paper. Finally, we conclude this paper in Section V.

II. RELATED WORK

In this subsection, we briefly review the related works on automatic image annotation and multiview feature learning methods.

A. AUTOMATIC IMAGE ANNOTATION

With the growth of multimedia content, autoannotation methods have attracted increasing attention, such as image retrieval [1], [24] and image classification [19], [25]. Given an unlabeled image, automatic image annotation methods assign this image with several textual labels to describe this image. As a very important research topic in recent years, many approaches [2], [11], [15], [16] have been proposed to solve the image annotation problem. Existing literature usually utilizes machine learning strategies to bridge the semantic gap between visual features and semantic concepts, which combine multiple sources of information such as visual features with semantic concepts. The semantic gap is one of the challenges for image annotation. According to different solutions, existing image annotation methods can be roughly categorized into model-based and retrieval-based methods. Model-based methods contain generative models and discriminative models. These types of methods usually utilize machine learning methods or knowledge models to find a mapping function. They narrow the semantic gap by exploring the relationships between high-level semantic concepts and low-level visual content features. Retrieval-based methods directly provide candidate labels from labeled images by some retrieval strategies, such as nearest neighbor-based methods.

Generative models assume images are sampled from a statistical distribution and then assign labels to images by computing a joint distribution over image features or labels [26]–[28]. This leads to a conditional probability over labels. It contains mixture models [28] and topic

models [27], [29]. In [30], an image was divided into several real-valued feature vectors, then a joint probability between image features and labels was calculated, and the probability of each label was predicted. Topic models [29] assume samples consist of a mixture of topics. In addition, constrained nonnegative matrix factorization [31] and probabilistic latent semantic analysis (pLSA) [32] was proposed to address a multivariate binary response variable in the annotation data.

Discriminative models [33]–[35] usually learn a separate classifier for each class and predict the labels for each image. A support vector machine (SVM)-based multilabel annotation method was proposed in [34], which modifies the SVM hinge loss function. Another study [35] used discriminative feature mapping and feature selection to address the image annotation issue. More recently, the label-specific features (LIFT) [36] algorithm was developed to promote discrimination of different class labels by utilizing visual feature information. However, these approaches fail to explore the correlation among different labels. In addition, the relationships between labels and the visual features have not yet been considered.

Nearest neighbor-based methods assume that similar images share common labels [37]. To this end, many of these methods implement a label transfer from training labels to the test image by retrieving a set of visually similar images. Various kinds of features are usually combined to compute similarities among images. Despite their simplicity and effectiveness, nearest neighbor-based approaches [16], [38] have attracted more attention and achieved promising results in image annotation. More recently, weighted nearest neighbor annotation models [39]–[41] have been proposed to weight rare and common labels, which can capture the semantic correlation between labels. In TagProp [39], a weighted nearest neighbor technique combined with metric learning addressed the image annotation problem well. In 2PKNN [40], a two-step variant of the k-nearest neighbor(KNN) based method was proposed to learn weights for multiple features. In the first step, it uses image-to-tag relations, and in the second step, it exploits image-to-image similarities. Finally, it predicts labels for the testing images. However, these methods computed image similarity only according to visual features. Their performance depends heavily on the number of training examples. Moreover, exact neighbor search is very time consuming in large datasets. Thus, an efficient image annotation method is urgent.

To overcome these issues, some methods [4], [11] improve annotation performance by using more elegant visual features. Some [36] consider the correlation of different labels as additional information. Some of them [42] consider the relationships between the labels and features. Some [15] introduce the image-to-label relation to the annotation issue. For better annotation, the proposed method considers rich information together. Multiple information will help to improve image annotation performance.

B. MULTIVIEW FEATURE LEARNING

Many multiview feature learning algorithms have been proposed to solve the vision tasks, such as image retrieval [43], clustering [44] and annotation [45], [46]. Generally, multiview learning approaches can be roughly grouped into two types [47]: feature-level fusion-based [48], [49] and classifier-level fusion [50]. Among these methods, canonical correlation analysis (CCA) explores a three-view embedding space. Some experimental studies have shown that classifier-level fusion is better than simple feature fusion. However, the existing works [50], [51] also proved that sophisticated feature-level fusion usually performs better than classifier-level fusion. As is known, multiview features improve the accuracy, but the increasing features decrease efficiency due to the feature dimension. Moreover, most multiview learning methods are unsupervised, which cannot make use of the label information. To deal with the dimensionality problem, we utilize the NMTF to reduce the dimensions. To explore the label information, we the employ collective matrix tri-factorization framework to decompose the image-to-label matrix and feature-to-label matrix. There is an assumption in this paper: the relation between each view and label is consistent with the relation between each image and label. Different views should have the same annotation result. We use this assumption to constrict the NMTF to capture some latent features during the matrix factorization. Then, we impose the similarity constraints between related labels to capture the correlations among different labels, and the similarity constraints between related features to capture the relationships among different views. Since all the views capture the same latent structure, the constraints can enforce the method to discover a consistent latent structure for all views [44]. In NMF-KNN [45], a weighted extension of the multiview NMF method was proposed to address the dataset imbalance and feature-fusion problems. In this method, tag was treated as another feature view of the image. Then, it learns a set of bases across all the views which respond to the same underlying features.

More recently, some visual-based image annotation methods improve the performance by using deep learning features, which have shown significant improvement in computer vision tasks [11], [12]. To benefit from the efficiency of CNN methods, we use the CNN features as another view of image features. This means the CNN feature is just one of the multiview features in this paper. We not only extract the low-level feature but also learn the CNN features, which is helpful for studying the consistencies among all the views.

III. THE PROPOSED METHOD

In the following, we investigate the annotation problem with a trigraph regularized collective matrix tri-factorization framework.

A. PROBLEM FORMULATION

To clearly describe the proposed method, we describe the image-feature-label tritype graph model for a set of given

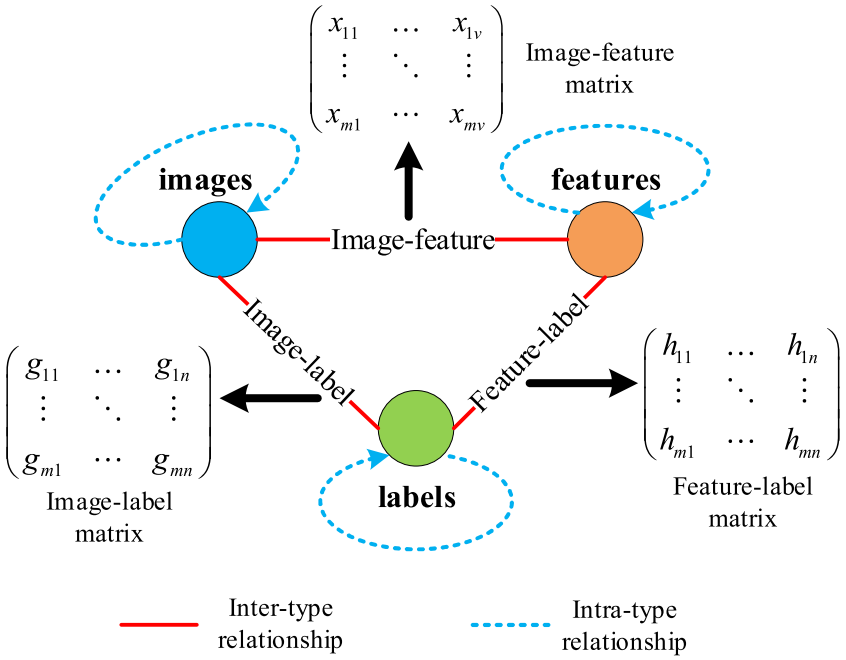


FIGURE 2. Overview of the proposed method for image annotation.

images, based on which the image annotation framework is developed. Let $Q = \langle V, E, R, S \rangle$, where V denotes the set of vertices that consists of images set $P = \{p_1, p_2, \ldots, p_m\}$, the features set $F = \{f_1, f_2, \ldots, f_v\}$ and the label set $W = \{w_1, w_2, \ldots, w_n\}$, i.e., $V = P \cup F \cup W$. m is the number of images, v is the number of features, and n is the total number of labels. E is the set of edges that connect the vertices in V. R is a set of intertype relationship matrices that consists of image-to-feature relation $R_{PF} \in \mathbb{R}^{m \times v}$, image-to-label relation $R_{PW} \in \mathbb{R}^{m \times n}$ and feature-to-label relation $R_{FW} \in \mathbb{R}^{m \times n}$. S is a set of intratype relationship matrices that consists of $S_{PP} \in \mathbb{R}^{m \times m}$, $S_{FF} \in \mathbb{R}^{m \times m}$ and $S_{WW} \in \mathbb{R}^{n \times n}$. We use Figure 2 to illustrate the trigraph of the method.

Figure 2 shows the overview of the proposed method based on the relationships. This framework captures informative semantic relations between the (a) image feature, (b) image-label, (c) feature-label, (d) image-image, (e) feature-feature, and (f) label-label. With the help of these relationships, this framework can deal with the complex semantics of image annotation tasks well.

B. INCORPORATING THE INTRATYPE RELATIONSHIPS

In this paper, we explore the following three types of intrarelationships: intraimage, intrafeature, and intralabel. To utilize such information, we construct three graphs. Then, we incorporate the three graphs through Laplacian regularizations, which are in the form of pairwise affinity matrices.

1) MODELING THE INTRAIMAGE RELATIONSHIPS

There is no doubt that if two images are visually similar to each other, they have similar labels. To discover the intrinsic discriminative structure of images, we model the manifold structure among images by using Laplacian [21]. We construct an image-image graph $G_P = (v_P, \varepsilon_P)$, where vertex set v_P denotes the images set $\{p_1, p_2, \ldots, p_m\}$. Each node in the graph represents an image in the dataset, and an edge represents the affinity angle between two images. There is an assumption that the labels of one image are usually interdependent on their similar images or neighbors in the image graph. To weight the edge, we use the heat kernel [22] as follows

$$W_{ij}^P = e^{-\frac{\|p_i - p_j\|^2}{\sigma}}$$

where $\sigma \in \mathbb{R}$ weights the distance between image p_i and image p_j. According to this equation, we can construct the image similarity matrix for pairwise images.

If two images are similar, then their new representations in the new space are also similar, which we can formulate as the following equation

$$O_1 = \frac{1}{2} \sum_{i,j}^{m} \left\| (Q_P)_i - (Q_P)_j \right\|^2 W_{ij}^P = Tr((Q_P)^T L_P (Q_P))$$

where $Tr(.)$ represents the trace of the matrix. $L_P = D^P - W^P$ is the Laplacian matrix of the image graph. $W^P = [W_{ij}^P]$ is a symmetric nonnegative similarity matrix. $D^P \in \mathbb{R}^{m \times n}$ is a diagonal matrix and $D_{ii}^P = \sum_{j-1}^{m} W_{ij}^P$.

2) MODELING THE INTRAFEATURE RELATIONSHIPS

Low-level visual features play a very important role in many vision tasks, as image retrieval, image annotation and image classification. Since the semantic gap between low-level features and high-level semantic concepts, using only one type of feature cannot achieve a promising result. Multiview

of features capture multiple aspects of the image and can improve the performance [45]. As is known, designing and choosing a suitable set of features is rather challenging. Thus, we used 15 publicly available handcrafted features, which have been proven useful in image annotation [39]. The visual features can be categorized into local features and global features. Global features contain gist features [52] and color histograms of red/green/blue(RGB), LAB and hue saturation value (HSV). These features are computed over three equal horizontal divisions except for gist features. The local features consist of scale-invariant feature transform (SIFT) [53] and hue descriptor. These features are extracted from Harris-Laplacian interest points and multiscale grids, respectively. The 15 distinct features form 15 visual data matrices for each image.

Additionally, due to the significant performance in computer vision [10]–[12], [54], we use the CNN feature as a view of the visual features. However, designing and tuning a new network is rather challenging. Thus, we use a publicly available network named VGG-net, first introduced in [54]. This is a 16-layer network. We use it to extract a 4096-dimensional visual feature vector for each image. According to the original paper and to compatible with the VGG-net architecture, we resize all the images to 224*224 as the network input. The visual features are the activations of the last fully connected layer. This has proven to be good in image recognition and image classification [12], [55].

Since it is theoretically clear and its efficient result in prior work, we use cosine similarity as the evaluation measure of feature similarity, which can be defined as follows:

$$sim_{VS}(f_i^A, f_j^A) = \frac{\langle f_i^A, f_j^A \rangle}{\|f_i^A\| \, \|f_j^A\|}$$

where $\langle . \rangle$ represents the inner product of the two image-feature vectors. f_i^A denotes the large view feature of the i^{th} image.

We construct a weighted graph according to this similarity. In this graph, the weight of the edges signifies the level of similarity between the features. An edge weight of zero denotes that the two features are not associated. The weight of this graph can be defined as follows:

$$W_{ij}^F = \begin{cases} sim_{VS}(f_i^A, f_j^A) & \text{if two vectors are associated.} \\ 0 & otherwise. \end{cases}$$

We minimize the fitting errors of all the features to capture the interfeature relationships as

$$O_2 = \frac{1}{2} \sum_{i,j}^{m} \left\| (Q_F)_i - (Q_F)_j \right\|^2 W_{ij}^F = Tr((Q_F)^T L_F (Q_F))$$

where $D_{ii}^F = \sum_j W_{ij}^F$ is a diagonal matrix, and $L_F = D^F - W^F$ is the Laplacian matrix of the feature graph. Moreover, the graph Laplacian preserves the local geometric information.

3) MODELING THE INTRALABEL RELATIONSHIPS

If two labels are colabeled by the same image, they are more likely to be colabeled by other images. Thus, we calculate the cooccurrence percentage of the two labels by counting the label pairs tagged by the same images in the whole dataset. To compute the cooccurrence percentage of the two labels, we construct a weighted graph $G^W = (v_W, \varepsilon_W)$, where each node represents one label, and each edge shows the affinity angle between two labels. In this paper, we use $\{w_1, w_2, \ldots, w_n\}$ to denote the label set. We can define the affinity matrix W_{ij}^w on the label graph as follows:

$$W_{ij}^w = \begin{cases} \text{corr}\,(w_i, w_j), & if\ w_i, w_j \in L(p_i) or w_i, w_j \in L(p_j) \\ 0, & \text{otherwise.} \end{cases}$$

where $w_i, w_j \in L(p_i)$ indicates that image p_i is labeled by labels w_i and w_j together. corr(w_i, w_j) denotes the correlation between the two labels. To calculate the cooccurrence relations among labels, we need to build a binary image-to-label matrix T for the labeled images in the dataset. We use the rows in the matrix to indicate the different images, and the columns denote the relations with each label. If image x_i is labeled by label l_j, then $t_ij = 1$ and otherwise 0. We can compute the label cooccurrence between two labels as follows:

$$corr(w_i, w_j) = \frac{\langle t_{:i}, t_{:j} \rangle}{\|t_{:i}\| \, \|t_{:j}\|}$$

where $\langle t_{:i}, t_{:j} \rangle$ counts the number of images annotated by labels w_i and w_j together. According to this, we can construct the graph regularization of the label graph as follows:

$$O_3 = \frac{1}{2} \sum_{i,j}^{c} \left\| (Q_w)_i - (Q_w)_j \right\| W_{ij}^w = Tr((Q_W)^T L_W (Q_W))$$

where W_{ij}^w encodes the label information and $L_W = D^w - W_w$ is the Laplacian matrix of the weighted matrix W, $D_{ii}^w = \sum_{j=1}^{n} W_{ij}^w$.

C. INCORPORATING THE INTERTYPE RELATIONSHIPS

In this section, we incorporate the intertype relationships as image-to-label, feature-to-label and image-to-feature relations. Intertype relationships are deemed very important in image annotations that have not been used.

1) MODELING THE IMAGE-FEATURE RELATIONSHIPS

Assume we have m images and n labels. We extract v views visual features by using the v type of feature extraction algorithms. Let d_j denote the dimension of the j^{th} visual features. We use x_j^s to indicate the s^{th} image in the j^{th} view. Since there are m images in the dataset, we can construct an $m \times d_j$ nonnegative matrix $X_j = [x_j^1, x_j^2, \ldots, x_j^m]^T$, where the rows are images, and the columns are the different views of features. To interpret it easily, we can rewrite the matrix as

follows:

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1v} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mv} \end{pmatrix} = \begin{pmatrix} x_1^1 & \cdots & x_j^1 \\ \vdots & \ddots & \vdots \\ x_1^m & \cdots & x_j^m \end{pmatrix}$$

2) MODELING THE IMAGE-LABEL AND FEATURE-LABEL MATRICES

To utilize the relationship between images and labels, we first build the image-label matrix $G \in \mathbb{R}_+^{m\times n}$ for the labeled images whose labels are known in the dataset. In this matrix, rows are images, and columns are labels. Specifically, if image p_i is annotated with the j^{th} label, we set $g_{ij} = 1$; otherwise, $g_{ij} = 0$. Similarly, we can construct the feature-label matrix $H \in \mathbb{R}^{m\times n}$. For the j^{th} label and i^{th} view feature for the s^{th} image, we define $h_{ij}(\mathrm{s}) = \mathrm{f}_s^v$ to indicate that the s^{th} image of the i^{th} view feature is related to the j^{th} label. We compute $\mathrm{f}_s^v \geq 0$ by combining all the features of one image into a single large ''view''. Then, we use it to denote the relationships between features and labels. The construction of the two matrices can leverage the label information from all labels and features. This helps to annotate the remaining unlabeled images with each label and can provide the complement information for the unlabeled images. Furthermore, it ensures the consistencies among different views of features and the correlations across related labels. For each dataset, we need to build the image-label matrix and feature-label matrix.

D. OBJECTIVE FUNCTION OF TG-CMTF

In this paper, we aim to propagate the label information from labeled images to unlabeled images, by using multiview features, image-to-label, image-to-image, image-to-feature, feature-to-feature and label-to-label relationships. We formulate the image annotation problem as an optimization problem, which utilizes a trigraph regularized collective matrix tri-factorization framework. This framework solves the optimization problem by sharing some low-rank matrices, such as Q_P, Q_F and Q_W. The shared matrices help the factorizations to find more interpretable low-rank representations, which explores the multiple relationships among images, labels and features. The mathematical pattern of this optimization problem can be described as follows

$$\begin{aligned} \min\ & L(Q_P, V_{PF}, Q_F, V_{PW}, Q_W, V_{FW}) \\ &= \frac{1}{2}\left\|X - Q_P V_{PF} Q_F^T\right\|^2 + \frac{\lambda_G}{2}\left\|G - Q_P V_{PW} Q_W^T\right\|^2 \\ &+ \frac{\lambda_H}{2}\left\|H - Q_F V_{FW} Q_W^T\right\|^2 + \frac{\alpha}{2}Tr(Q_P^T L_P Q_P) \\ &+ \frac{\beta}{2}Tr(Q_F^T L_F Q_F) + \frac{\gamma}{2}Tr(Q_W^T L_W Q_W), \\ s.t.\ & Q_P \geq 0, \quad Q_F \geq 0,\ Q_W \geq 0,\ V_{PF} \geq 0, \\ & V_{PW} \geq 0, \quad V_{FW} \geq 0 \end{aligned} \tag{1}$$

where $\|.\|$ denotes the Frobenius norm, $\lambda_G, \lambda_H, \alpha, \beta, \gamma$ are the parameters that balance the reconstruction error of TG-CMTF in the first three terms and the remaining terms. The first term is designed for factorizing the image-feature matrix of all views, which effectively induces a low-dimensional representation of the latent space. This term helps to find the low-rank image specific latent matrix and feature specific latent matrix. The second term is for factorizing the image-to-label relation matrix, and the third term is for the feature-to-label relation matrix. The last three terms are the graph regularizations to utilize the intratype relations. Cofactorizing each relation helps to find the more interpretable new latent representations.

In this equation, α is used to tune the smoothness of the images. β controls the consistency of all the views and γ retains the label smoothness. The three graph Laplacian terms preserve the local geometric information of images, features and labels. To minimize this objective function in Eq. (1), we can rewrite it as follows

$$\begin{aligned} \min\ & L = \frac{1}{2}(XX^T - 2XQ_P V_{PF}^T Q_P^T + Q_P V_{PF} Q_F^T Q_F V_{PF}^T Q_P^T) \\ &+ \frac{\lambda_G}{2}(GG^T - 2GQ_W V_{PW}^T Q_P^T + Q_P V_{PW} Q_W^T Q_W V_{PW}^T Q_P^T) \\ &+ \frac{\lambda_H}{2}(HH^T - 2HQ_W V_{FW}^T Q_F^T + Q_F V_{FW} Q_W^T Q_W V_{FW}^T Q_F^T) \\ &+ \frac{\alpha}{2}Tr(Q_P^T L_P Q_P) + \frac{\beta}{2}Tr(Q_F^T L_F Q_F) + \frac{\gamma}{2}Tr(Q_W^T L_W Q_W), \\ s.t.\ & Q_P \geq 0, \quad Q_F \geq 0,\ Q_W \geq 0,\ V_{PF} \geq 0, \\ & V_{PW} \geq 0,\ V_{FW} \geq 0 \end{aligned} \tag{2}$$

There is not a straightforward method for solving the above equation by a closed solution, because it is not a convex function with respect to the six variables together. However, it is a convex function about each variable separately. We optimize this function by an alternating scheme and solve it with respect to one variable while fixing the other variables. Repeat this process until convergence.

IV. IMAGE ANNOTATION VIA TG-CMTF

In this section, we study how to optimize Eq. (1) using an alternating scheme. As is known, the objective function in Eq. (1) is a nonconvex function with respect to all the variables together, but it is convex separately. Therefore, we optimize one variable by fixing the others and repeating this procedure until convergence.

A. THE MULTIPLICATIVE UPDATING RULES

To obtain the multiplicative updating rules, we iteratively solve these variables one by one. In general, the proposed method is a 3-factor NMF. It is worth noting that the factors Q_P, Q_F and Q_W in the 3-factor NMF can be dealt with in the same way as in 2-factor NMF [23]. The key difference is the factors in the middle, such as V_{PF}, V_{PW} and V_{FW}. These rules are obtained in the following.

1) COMPUTATION OF Q_P

Solve Eq. (1) for Q_P is equivalent to optimizing the following equation

$$\begin{aligned}
&L(Q_P)\\
&= \frac{1}{2}\left\|X - Q_P V_{PF} Q_F^T\right\|^2 + \frac{\lambda_G}{2}\left\|G - Q_P V_{PW} Q_W^T\right\|^2\\
&\quad + \frac{\alpha}{2} Tr(Q_P^T L_P Q_P)\\
&= \frac{1}{2}(XX^T - 2XQ_F V_{PF}^T Q_P^T + Q_P V_{PF} Q_F^T Q_F V_{PF}^T Q_P^T)\\
&\quad + \frac{\lambda_G}{2}(GG^T - 2GQ_W V_{PW}^T Q_P^T + Q_P V_{PW} Q_W^T Q_W V_{PW}^T Q_P^T)\\
&\quad + \frac{\alpha}{2} Tr(Q_P^T L_P Q_P)
\end{aligned}$$

By using the Karush-Kuhn-Tucker complementary condition (KKTCC)[56] for the nonnegativity of Q_P and setting $\frac{\partial L(Q_P)}{\partial Q_P} = 0$, we obtain the following updating formula

$$\begin{aligned}
&[-XQ_F V_{PF}^T + Q_P V_{PF} Q_F^T Q_F V_{PF}^T - \lambda_G G Q_W V_{PW}^T\\
&\qquad + \lambda_G Q_P V_{PW} Q_W^T Q_W V_{PW}^T + \alpha(L_P Q_P)]_{ij}(Q_P)_{ij} = 0
\end{aligned}$$

Similar to the previous work [57] which introduced $L_U = L_U^+ + L_U^- M_{ij}^+ = (|M_{ij}| + M_{ij})/2$, and $M_{ij}^- = (|M_{ij}| - M_{ij})/2$, we obtain the following multiplicative updating rule (3), as shown at the bottom of this page.

2) COMPUTATION OF Q_F

Solving the objective function in Eq. (1) with respect to Q_F is equivalent to optimizing the following problem:

$$\begin{aligned}
L(Q_F) = \frac{1}{2}\left\|X - Q_P V_{PF} Q_F^T\right\|^2 + \frac{\lambda_H}{2}\left\|H - Q_F V_{FW} Q_W^T\right\|^2\\
+ \frac{\beta}{2} Tr(Q_F^T L_F Q_F).
\end{aligned}$$

Calculate the partial derivative of Eq. (3) and use the KKTCC condition to obtain

$$\begin{aligned}
&[-X^T Q_P V_{PF} + Q_F V_{PF}^T Q_P^T Q_P V_{PF} - \lambda_H H Q_W V_{FW}^T\\
&\qquad + \lambda_H Q_F V_{FW} Q_W^T Q_W V_{FW}^T + \beta L_F Q_F]_{ij}(Q_F)_{ij} = 0
\end{aligned}$$

Since L_F may take any signs, the following equation will hold

$$\begin{aligned}
&[-X^T Q_P V_{PF} + Q_F V_{PF}^T Q_P^T Q_P V_{PF} - \lambda_H H Q_W V_{FW}^T\\
&+ \lambda_H Q_F V_{FW} Q_W^T Q_W V_{FW}^T + \beta L_F^+ Q_F - \beta L_F^- Q_F]_{ij}(Q_F)_{ij} = 0
\end{aligned}$$

Solving it for Q_F, we obtain the updating rule with respect to Q_F as (4), shown at the bottom of this page.

3) COMPUTATION OF Q_W

Similarly, we fix the other variables and obtain the following equation

$$\begin{aligned}
L(Q_W) = \frac{\lambda_G}{2}\left\|G - Q_P V_{PW} Q_W^T\right\|^2 + \frac{\lambda_H}{2}\left\|H - Q_F V_{FW} Q_W^T\right\|^2\\
+ \frac{\gamma}{2} Tr(Q_W^T L_W Q_W).
\end{aligned}$$

Calculating the partial derivative of $L(Q_W)$ and setting it to zero, the following equation holds

$$\begin{aligned}
&[-\lambda_G G^T Q_P V_{PW} + \lambda_G Q_W V_{PW}^T Q_P^T Q_P V_{PW} - \lambda_H H^T Q_F V_{FW}\\
&\quad + \lambda_H Q_W V_{FW}^T Q_F^T Q_F V_{FW} + \gamma L_W^+ Q_W - \gamma L_W^- Q_W]_{ij}(Q_W)_{ij}\\
&= 0
\end{aligned}$$

Then, we solve for Q_W and obtain the updating rule for Q_W (5), as shown at the bottom of this page.

4) COMPUTATION OF V_P, V_F and V_W

Similar to the previous steps, the solution of middle factors V_P, V_F and V_W are analogous. We define the following functions

$$\begin{aligned}
L(V_{PF}) &= \frac{1}{2}\left\|X - Q_P V_{PF} Q_F^T\right\|^2\\
&= \frac{1}{2}(XX^T - 2XQ_P V_{PF}^T Q_P^T + Q_P V_{PF} Q_F^T Q_F V_{PF}^T Q_P^T)\\
L(V_{PW}) &= \frac{\lambda_G}{2}\left\|G - Q_P V_{PW} Q_W^T\right\|^2\\
&= \frac{\lambda_G}{2}(GG^T - 2GQ_W V_{PW}^T Q_P^T\\
&\quad + Q_P V_{PW} Q_W^T Q_W V_{PW}^T Q_P^T)\\
L(V_{FW}) &= \frac{\lambda_H}{2}\left\|H - Q_F V_{FW} Q_W^T\right\|^2\\
&= \frac{\lambda_H}{2}(HH^T - 2HQ_W V_{FW}^T Q_F^T\\
&\quad + Q_F V_{FW} Q_W^T Q_W V_{FW}^T Q_F^T)
\end{aligned}$$

Then, we calculate the first derivatives of these equations. To preserve the negativity of V_{PF}, V_{PW} and V_{FW}, we use the KKTCC and set the partial derivatives of the three equations

$$(Q_P)_{ij} \leftarrow (Q_P)_{ij}\sqrt{\frac{[XQ_F V_{PF}^T + \lambda_G G Q_W V_{PW}^T + \alpha L_P^- Q_P]_{ij}}{[Q_P V_{PF} Q_F^T Q_F V_{PF}^T + \lambda_G Q_P V_{PW} Q_W^T Q_W V_{PW}^T + \alpha L_P^+ Q_P]_{ij}}} \tag{3}$$

$$(Q_F)_{ij} \leftarrow (Q_F)_{ij}\sqrt{\frac{[X^T Q_P V_{PF} + \lambda_H H Q_W V_{FW}^T + \beta L_F^- Q_F]_{ij}}{[Q_F V_{PF}^T Q_P^T Q_P V_{PF} + \lambda_H Q_F V_{FW} Q_W^T Q_W V_{FW}^T + \beta L_F^+ Q_F]_{ij}}} \tag{4}$$

$$(Q_W)_{ij} \leftarrow (Q_W)_{ij}\sqrt{\frac{[\lambda_G G^T Q_P V_{PW} + \lambda_H H^T Q_F V_{FW} + \gamma L_W^- Q_W]_{ij}}{[\lambda_G Q_W V_{PW}^T Q_P^T Q_P V_{PW} + \lambda_H Q_W V_{FW}^T Q_F^T Q_F V_{FW} + \gamma L_W^+ Q_W]_{ij}}} \tag{5}$$

to zero. By solving these equations with respect to the three variables, we obtain

$$(V_{PF})_{ij} \leftarrow (V_{PF})_{ij}\sqrt{\frac{[Q_P^T X Q_P]_{ij}}{[Q_P^T Q_P V_{PF} Q_F^T Q_F]_{ij}}} \tag{6}$$

$$(V_{PW})_{ij} \leftarrow (V_{PW})_{ij}\sqrt{\frac{[Q_P^T G Q_W]_{ij}}{[Q_P^T Q_P V_{PW} Q_W^T Q_W]_{ij}}} \tag{7}$$

$$(V_{FW})_{ij} \leftarrow (V_{FW})_{ij}\sqrt{\frac{[Q_F^T H Q_W]_{ij}}{[Q_F^T Q_F V_{FW} Q_W^T Q_W]_{ij}}} \tag{8}$$

The successive updates lead the object function in Eq. (1) to converge to a local minimum.

B. THE CONVERGENCE OF TG-CMTF

There are 6 factor variables in the objective function. With these six variables together, the objective function in Eq. (1) is not a strictly convex function; however, it is convex with these variables separately. Therefore, the minimization of Eq. (1) can be achieved by the alternative multiplicative updating rules. In the following, we prove the convergence of Eq. (1). In the proof, the auxiliary function technique [58] is employed. We begin with the definition and lemma.

Definition 1: If $K(U, U^t)$ is an auxiliary function of $S(U)$, the following three conditions should be satisfied:

$$\begin{aligned} S(U^t) &\geq K(U^t, U^t) \\ K(U^t, U^t) &\geq K(U^{t+1}, U^t) \\ K(U^{t+1}, U^t) &\geq S(U^{t+1}) \end{aligned}$$

Lemma 1: If $K(U, U^t)$ is an auxiliary function of $S(U)$, then $S(U)$ converges under the following update: $U^{t+1} = \arg\min K(U, U^t)$, where U^t is the t^{th} iteration of U.

According to the above definition and lemma, we derive the following theorem.

Theorem 1: The updating rule in Eq. (3) leads the objective function in Eq. (1) to converge to a local minimum with respect to Q_P.

To prove Theorem 1, the following lemma used in previous works is required.

Lemma 2 [59]: For any symmetric matrices $A \in \Re_+^{n\times n}$, $B \in \Re_+^{k\times k}$, and any matrices $M \in \Re_+^{n\times k}$ and $M' \in \Re_+^{n\times k}$, the following inequality holds:

$$\sum_{ij} \frac{(AM'B)_{ij} M_{ij}^2}{M'_{ij}} \geq Tr(M^T A M B)$$

Proof: The objective function in Eq. (1.2) with respect to Q_P is equal to

$$\begin{aligned} S(Q_P) = &-Tr(X Q_F V_{PF}^T Q_P^T) + Tr(\frac{1}{2} Q_P V_{PF} Q_F^T Q_F V_{PF}^T Q_P^T) \\ &- Tr(\lambda_G G Q_W V_{PW}^T Q_P^T) \\ &+ Tr(\frac{\lambda_G}{2} Q_P V_{PW} Q_W^T Q_W V_{PW}^T Q_P^T) \\ &+ \frac{\alpha}{2} Tr(Q_P^T L_P^+ Q_P) - \frac{\alpha}{2} Tr(Q_P^T L_P^- Q_P) \end{aligned}$$

We find an auxiliary function for $S(Q_P)$ as follows

$$\begin{aligned} &K(Q_P, Q'_P) \\ &= -\sum_{ij} ((X Q_P V_{PF}^T)_{ij} (Q_P)'_{ij} (1 + \log \frac{(Q_P)_{ij}}{(Q_P)'_{ij}}) \\ &\quad + \frac{1}{2} \sum_{ij} \frac{(Q'_P V_{PF} Q_F^T Q_F V_{PF}^T)_{ij} (Q_P)_{ij}^2}{(Q_P)'_{ij}} \\ &\quad - \lambda_G \sum_{ij} (G Q_W V_{PW}^T)_{ij} (Q_P)'_{ij} (1 + \log \frac{(Q_P)_{ij}}{(Q_P)'_{ij}}) \\ &\quad + \frac{\lambda_G}{2} \sum_{ij} \frac{(Q'_P V_{PW} Q_W^T Q_W V_{PW}^T)_{ij} (Q_P)_{ij}^2}{(Q_P)'_{ij}} \\ &\quad + \frac{\alpha}{2} \sum_{ij} \frac{(L_P^+ (Q_P)')_{ij} (Q_P)_{ij}^2}{(Q_P)'_{ij}} \\ &\quad - \frac{\alpha}{2} \sum_{ijk} (L_P^-)_{jk} (Q_P)'_{ji} (Q_P)'_{ki} (1 + \log \frac{(Q_P)_{ji} (Q_P)_{ki}}{(Q_P)'_{ji} (Q_P)'_{ki}}) \end{aligned}$$

To prove $K(Q_P, Q'_P)$ is an auxiliary function of $S(Q_P)$, we can prove it satisfies three conditions in Definition 1. First, the equality $K(Q_P, Q'_P) = S(Q_P)$ holds when $Q'_P - Q_P$. Second, the inequality holds $K(Q_P, Q'_P) \geq S(Q_P)$. This is because (a) due to $z \leq 1 + \log z$, $\forall z > 0$, let $z = (Q_P)_{ik}/(QP)'_{ik}$, then we obtain the following inequalities

$$\begin{aligned} &Tr(X Q_F V_{PF}^T Q_P^T) \\ &\quad \geq \sum_{ij} ((X Q_P V_{PF}^T)_{ij} (Q_P)'_{ij} (1 + \log \frac{(Q_P)_{ij}}{(Q_P)'_{ij}}) \\ &Tr(G Q_W V_{PW}^T Q_P^T) \\ &\quad \geq \sum_{ij} (G Q_W V_{PW}^T)_{ij} (Q_P)'_{ij} (1 + \log \frac{(Q_P)_{ij}}{(Q_P)'_{ij}}) \\ &Tr(Q_P^T L_P^- Q_P) \\ &\quad \geq \sum_{ijk} (L_P^+)_{jk} (Q_P)'_{ji} (Q_P)'_{ki} (1 + \log \frac{(Q_P)_{ji} (Q_P)_{ki}}{(Q_P)'_{ji} (Q_P)'_{ki}}) \end{aligned}$$

(b) by applying Lemma 2, we obtain the following inequalities

$$Tr(Q_P V_{PF} Q_F^T Q_F V_{PF}^T Q_P^T) \leq \sum_{ij} \frac{(Q'_P V_{PF} Q_F^T Q_F V_{PF}^T)_{ij} (Q_P)_{ij}^2}{(Q_P)'_{ij}}$$

$$Tr(Q_P^T L_P^+ Q_P) \leq \sum_{ij} \frac{(L_P^+ (Q_P)')_{ij} (Q_P)_{ij}^2}{(Q_P)'_{ij}}$$

Summing the above bounds, we obtain $K(Q_P, Q'_P) \geq S(Q_P)$. Thus, the conditions in Definition 1 are all satisfied.

We then obtain the minimum value of $K(Q_P, Q'_P)$ according to Lemma 1. The partial derivation of $K(Q_P, Q'_P)$ with respect to Q_P is:

$$\begin{aligned} &\frac{\partial K(Q_P, Q'_P)}{\partial Q_P} \\ &= -(X Q_P V_{PF}^T)_{ij} \frac{(Q_P)'_{ij}}{(Q_P)_{ij}} + \frac{(Q'_P V_{PF} Q_F^T Q_F V_{PF}^T)_{ij} (Q_P)_{ij}}{(Q_P)'_{ij}} \\ &\quad - \lambda_G (G Q_W V_{PW}^T)_{ij} \frac{(Q_P)'_{ij}}{(Q_P)_{ij}} \end{aligned}$$

$$+\lambda_G \frac{(Q'_P V_{PW} Q_W^T Q_W V_{PW}^T)_{ij}(Q_P)_{ij}}{(Q_P)'_{ij}}$$
$$+\alpha \frac{(L_P^+(Q_P)')_{ij}(Q_P)_{ij}}{(Q_P)'_{ij}} - \alpha(L_P^-)_{ij}(Q_P)'_{ij}\frac{(Q_P)'_{ij}}{(Q_P)_{ij}}$$

By setting $\frac{\partial K(Q_P, Q'_P)}{\partial Q_P} = 0$ and solving for $(Q_P)_{ij}$, the minimum is $(Q_P)_{ij}$, as shown at the bottom of this page.

According to Lemma 1, let $(Q_P)_{ij}^{t+1} = (Q_P)_{ij}$ and $(Q_P)'_{ij} = (Q_P)_t$, we have Eq. (3). Thus, Eq. (1) decreases monotonically and converges to a local minimum, and we have proven Theorem 1.

Similar to Theorem 1, we can derive the following theorems:

Theorem 2: The updating rule in Eq. (4) leads the objective function in Eq. (1) to converge to a local minimum with respect to Q_F.

Theorem 3: The updating rule in Eq. (5) leads the objective function in Eq. (1) to converge to a local minimum with respect to Q_W.

Noting the symmetry of Q_P, Q_F and Q_W in Eq. (1), the proofs of Theorem 2 and 3 are analogous to Theorem 1. These theorems guarantee the objective function in Eq. (1) under each updating rule always decreases until convergence.

C. IMAGE ANNOTATION VIA TG-CMTF

As is known, NMTF [23] was proposed to cluster the rows and columns of the input relationship matrix simultaneously. It achieves this by factorizing the input data matrix into three nonnegative matrices. This work applies the NMTF to solve the image annotation problem. First, it factorizes the image-feature matrix into three low-rank matrices. The rows of the image-to-feature matrix indicate the images in the dataset, and the columns can be interpreted as the multiview features of each image. Simultaneously, it factorizes the image-to-label matrix and feature-to-label matrix into two low-dimension matrices. The three relational matrices impact each other. Additionally, using the label related matrices helps in discovering each visual latent factor better. Thereby, it can derive a more interpretable feature factor matrix from the factorization of the image-to-feature matrix, and obtain an approximation of the image-to-label matrix from the second factorization. Moreover, this approximation can be used to refine the results of the first factorization.

The above factorization step aims to discover the multilatent spaces, such as the latent image factor matrix and latent feature matrix. The image annotation task can be achieved by label prediction for the testing images. We use X_{test}^{nTr+i} to denote the feature representation of the i^{th} image. Let $P = \{p_1, p_2, \ldots, p_m\}$ be the image set, and $W = \{w_1, w_2, \ldots, w_n\}$ be the labels set. In the training set, each image is related to several labels. We use $\mathrm{T} = \{(p_1, L_1), \ldots, (p_m, L_m)\}$ to denote the image-labels pair. We model the image annotation by the conditional probabilities of image A.

$$P(w_i/A) = \frac{P(A/w_i)P(w_i)}{P(A)}$$

where $P(w_i)$ denotes the prior probability of the label w_i, and $P(A/w_i)$ is the conditional probability of image A given a label w_i, which models the feature distribution of image A. Given the test image B, we obtain the best label set by the following function

$$w^* = \arg\max_i P(w_i/B) \tag{9}$$

After learning all the parameters in Eq. (1), we finish the image annotation by Algorithm 1.

Algorithm 1 Image Annotation via TG-CMTF

Input: image-label matrix G with labeled images, image-to-image similarity matrix W_P and feature-to-feature similarity matrix W_F, label-to-label cooccurrence matrix X, and image-to-label matrix G, image-to-feature matrix X, feature-to-label matrix H, loss error $\varepsilon \geq 0$, regularization parameters $\lambda_G, \lambda_H, \alpha, \beta, \gamma$, number of images m, number of total labels n, and number of latent features k;

Output: Predict the label set matrix T^{pre} for the test image set P_{test}.

Initialize: $Q_P \geq 0, Q_F \geq 0, V_{PF} \geq 0, Q_W \geq 0, V_{PW} \geq 0, V_{FW} \geq 0$

1: Extract multiview visual features for each image in the dataset;
2: Construct image-to-feature relation matrix X;
3: Construct image-to-label matrix G;
4: Construct feature-to-label relation matrix H;
5: Construct label cooccurrence matrix W^L;
6: Construct image-to-image similarity matrix W^P;
8: Construct feature-to-feature similarity matrix W^F;
9: Initialize $Q_P \geq 0, Q_F \geq 0, V_{PF} \geq 0$ and $Q_W \geq 0, V_{PW} \geq 0, V_{FW} \geq 0$, randomly;
10: **while** the loss error of Eq. (1) $> \varepsilon$ **do**
11: $t := t + 1$;
12: update Q_P^{t+1} according to Eq. (3);
13: update Q_F^{t+1} according to Eq. (4);
14: update Q_W^{t+1} according to Eq. (5);
15: update V_{PF}^{t+1} according to Eq. (6);
16: update V_{PW}^{t+1} according to Eq. (7);
17: update V_{FW}^{t+1} according to Eq. (8);
18: **end while**
19: Input the new learned low-rank representations to predict the labels by Eq. (9);
20: Return a tag list of top 5 tags with the largest 5 values for each test image.

$$(Q_P)_{ij} \leftarrow (Q_P)'_{ij}\sqrt{\frac{[XQ_F V_{PF}^T + \lambda_G G Q_W V_{PW}^T + \alpha L_P^- Q_P]_{ij}}{[Q_P V_{PF} Q_F^T Q_F V_{PF}^T + \lambda_G Q_P V_{PW} Q_W^T Q_W V_{PW}^T + \alpha L_P^+ Q_P]_{ij}}}.$$

Algorithm 1 summarizes the process of the proposed image annotation method TG-CMTF, which uses interrelationships and intrarelationships among the three types of images, features and labels for annotation. First, it extracts multiview features by various types of methods. After that, image-to-feature, feature-to-label and feature-to-feature relationships are constructed. Then, it builds the image-to-label, image-to-image and label-to-label relationships. After that, the labels of the test image are predicted by Eq. (9). In this algorithm, steps 10–18 are the multiplicative updating process. Repeating these steps, the objective function will converge.

D. TIME COMPLEXITY OF TG-CMTF

In this subsection, we study the computational complexity of the proposed algorithm. The complexity is denoted as big O. There are two major costs in the proposed algorithm. The first part constructs the multiple relation matrices. The second part updates the iterative multiplicative rules. There are six relation matrices that need to be built. Before that, the multiview visual features should be extracted, including CNN features. Since extracting the CNN features is a truly time-consuming issue, we do that using a predefined neural network offline. Other low-level features can also be extracted offline. Therefore, we do not consider the cost time of feature extraction. To reduce the running time, we also built the six matrices offline. We mainly consider the multiplicative updating time. Because the sizes of the image-label matrix and the label-to-label matrix are much smaller than the image-to-feature matrix, the key cost of the optimization is the factorization of the image-to-feature matrix. We assume the multiplicative updates stop after t' iterations. Thus, the time cost of the multiplicative updates is $O(t'(mkv + 2mkn))$, where m is the number of images, n is the number of labels, and v is the number of features. The construction of the image-to-image similarity matrix costs m^2. Building the multiple feature graph costs m^2, and the label-to-label cooccurrence graph spends n^2. Therefore, the overall time complexity of Algorithm 1 is approximate to $O(t'(mkv + 2mkn) + 2m^2 + n^2)$. Because k, v, n are all much smaller than m, and the complexity can be rewritten as $O(t'm + 2m^2)$.

V. EXPERIMENTS AND EVALUATIONS

In this section, we investigate the effectiveness of the proposed method by comparing it with other multilabel approaches. Furthermore, we analyze the results and show the influence of related parameters used in this paper.

A. DATASETS

To evaluate the effectiveness of the proposed method, we conduct a set of experiments on three standard publicly available image annotation datasets: Corel5K [60], ESP game [61] and IAPRTC12 [62]. We summarize the characteristics of these datasets in Table 1.

Corel5K: It has been the most important benchmark for the image annotation field. There are 5,000 manually annotated images as the training set, and a fixed set of 499 images as the test set. The images are annotated with 1 to 5 keywords. The average tag per image (TPI) is 3.4.

TABLE 1. Statistics of datasets.

Name	Number of images	Vocabulary size	TPI	Train	Test
Corel5k	5,000	260	3.522	4,500	500
IAPR TC12	19,627	291	5.723	17,665	1,962
ESP	20,700	268	4.762	18,689	2,081

TABLE 2. Statistics of features.

Feature name	Dimension
Gist	512
RGB	4096
HSV	4096
LAB	4096
Harris-SIFT	1000
Dense SIFT	1000
Harris Hue	100
Dense Hue	100
RGB H3V1	5184
HSV V3H1	5184
LAB V3H1	5184
Harris-SIFT V3H1	3000
Dense SIFT V3H1	3000
Harris Hue V3H1	300
Dense Hue V3H1	300
CNN	4096

ESP Game: This dataset is generated from an online game, which is also widely used in multilabel annotation tasks. This dataset consists of more images, including logos and personal photos. There are 18,689 training images and 2,081 testing images in this dataset, and the TPI is 4.7.

IAPR TC12: This dataset was first used for cross-language information retrieval. It contains 19,627 images, including sports, people, cities and other contemporary scenes. We use 1,962 images as the testing set and the rest as the training set. Its TPI is 5.7.

In the experiments, we use 15 publicly available [39] features to construct related matrices. The multiview features are shown in Table 2.

B. EVALUATION MEASURES

We evaluate the proposed method with standard performance measures, which have been used widely in many annotation works [39], [45]. In the image annotation domain, precision and recall for fixed annotation length are the most popular measures. Thus, each image is annotated with the n most relevant labels. In the experiments, we annotate each image with the 5 most relevant keywords. Then, we calculate the mean precision(P) and recall(R) over keywords. As the harmonic mean of recall and precision, the F1 score is computed, which is more reliable. N plus measures the nonzero recall value.

This indicates the number of labels that are correctly assigned to one image, which is particularly useful in the case of class-imbalance. We define the precision, recall and F1 score as follows:

$$precision(w_i) = \frac{N_{correct}}{N_{labeled}}, \quad recall(l_i) = \frac{N_{correct}}{N_{all}}$$
$$F_1 - score(w_i) = 2\frac{\Pr ecision(w_i) \times \mathrm{Re}\ call(w_i)}{\Pr\ ecision(w_i) + \mathrm{Re}\ call(w_i)} \quad (10)$$

where w_i is the i^{th} label. $N_{correct}$ denotes the correctly annotated images, $N_{labeled}$ represents the correctly annotated images relevant to the ground-truth annotations, and N_{all} is the total number of images to be labeled. We choose cross-validation on 10 sets of random samples to validate the experiments.

C. COMPARISON WITH OTHER APPROACHES

To evaluate the efficiency of the proposed method, we choose 10 typical algorithms as the benchmark baselines, then compare them with the proposed method. The compared methods are summarized as follows

. **JEC** [37] uses a greedy algorithm to transfer labels from their visually similar neighbors. It extracts multiple features of images and treats each feature equally.

. **FastTag** [63] explores two simple linear mappings that are coregularized in a joint convex loss function and then extends the classic metric learning algorithm for multilabel prediction.

. **GENMF** [64] presents a semisupervised framework based on graph embedding and multiview nonnegative matrix factorization for multilabel image annotation. This method first constructs a graph embedding term in the multiview NMF for labels. Then, it fused the multiview features and reduced the dimensions based on the multiview NMF algorithm. After that, labels are predicted by using the new features through a KNN-based algorithm.

. **NMF-KNN** [45] presents a weighted extension of multiview nonnegative matrix factorization, which imposes consensus constraints on the coefficient matrices across different features. The introduction of the weighted matrices alleviates the issue of dataset imbalance.

. **TagProp** [39] learns a weighted nearest neighbor model and annotates the labels using label relevance prediction. It maximizes the likelihood of a probabilistic model and learns rank-based weight.

. **2PKNN** [40] uses a two-level multilabel metric learning method to learn the weights of each feature and the weight of each element of a single feature vector. It is a variant of the k-nearest neighbor algorithm with two-step. The first step utilizes the image-to-label relation to address the label-imbalance problem, while the second step solves the weak-labeling issue by using an image-to-image relation.

. **MvNMF-DK** [46] uses multiple features, which are in a large variety of dimensions. The goal of this method is to find the best k for each feature experimentally by testing different values for the number of basis vectors and then

TABLE 3. Parameters used for experiments.

Parameters notation	Description	Parameter setting
λ_G	Weight of image-to-label relation information	0.1
λ_H	Weight of feature-to-label relation information	10
α	Weight of image-image visual-based similarity information	90
β	Weight of label-label semantic cooccurrence information	70
γ	Weight of label-to-label information	200
k	Dimension of latent features	50

improves the image annotation performance. However, high-dimensional data usually have a more complex latent space, which increases the computational cost.

. **MLDL** [65] exploits the label information by dictionary learning both in the input space and the output space. It simultaneously explores a label consistency regularization and partial-identical label embedding in multilabel dictionary learning.

. **CCA-KNN** [66] explores a k-nearest-neighbor-based canonical correlation analysis (CCA) method, which not only utilizes the convolutional neural network visual features but also explicitly incorporates the word embedding semantic information.

. **KCCA-LP** [42] proposed a label propagation framework based on kernel canonical correlation analysis (KCCA). This method fully utilizes the correlation of visual and textual features by a semantic embedding.

To validate the proposed method, we deployed several experiments on three multilabel datasets. The settings of these methods are determined by their papers or their codes. Before evaluating the effectiveness with other methods, we first tune the parameters. To achieve the best results, we use the best combination of parameters. Table 3 lists the parameters used in the experiments. All the experiments are conducted on a computer with an Intel Core i7, 3.6 GHz CPU and 16 GB of memory.

D. EXPERIMENTAL RESULTS AND DISCUSSION

In this subsection, we show the quantitative evaluation of the proposed method compared with other methods for the three datasets. The qualitative results are shown in Table 4. According to the definition of Eq. (10), the F1 score is the harmonic mean of recall and precision. Thus, the analysis of F1 is more reliable than recall or precision. Therefore, we mainly analyze the F1 score for these methods.

First, we evaluate the performance of the proposed method with JEC, FastTag, GENMF, NMF-KNN, TagProp and 2PKNN on the Corel5k dataset. Among these methods, except for FastTag, the remaining methods are all KNN-based methods. According to the results in Table 4, we observe that the proposed method achieves the best F1 score for the Core5K dataset. NMF-KNN attains the second-best result t, and JEC performs worst among these methods in this dataset.

TABLE 4. Experimental results for the three datasets.

Methods	Corel5k				IAPR TC12				ESP			
	R	P	F1	N^+	R	P	F1	N^+	R	P	F1	N^+
JEC	32	27	29.3	139	29	28	28.5	250	25	22	23.4	224
FastTag	43	32	37	166	26	47	34	280	22	46	30	247
GENMF	39	38	39.2	168	-	-	-	-	-	-	-	-
NMF-KNN	**56**	38	45.2	150	-	-	-	-	26	33	29.0	238
TagProp	42	33	36.9	160	34	45	39	260	27	39	31.9	239
2PKNN+ML	46	44	45	191	37	54	43.9	278	27	53	35.7	252
MvNMF-DK	47.5	44	45.6	197	38.5	49.4	43.3	281	31.4	43.7	36.7	254
MLDL	49	**45**	47	198	40	**56**	**47**	**282**	31	**56**	40	259
CCA-KNN	52	42	46	**201**	38	45	41	260	**36**	46	41	**278**
KCCA-LP	-	-	-	-	34	44	38.3	250	34	38	35.8	240
Proposed	51.4	43.7	**47.2**	189	**42.1**	51.8	46.4	278	35.6	55.8	**42.9**	276

As shown in Table 4, the proposed method has a 2% gain compared with NMF-KNN and a 17.9% improvement compared with JEC. The NMF-KNN method extracts the multiview features of images and formulates the tag feature as a single view of the image feature, then factorizes the multiple matrices to find the basis and the coefficient matrices for these images. Finally, it predicts the images by the learned matrices and the neighbors of the test images, then predicts the labels for the query images. This method discards the differences of these different views of features. Additionally, the features it extracts are all handcrafted features. This needs more prior knowledge. Nevertheless, the proposed method considers more relations among features, labels and semantic concepts. It also uses the CNN features as a view of the features. It helps to describe the images more understandably from the visual content viewpoint. 2PKNN+ML uses metric learning to learn the different weights for each feature, which improves the performance. However, it is slightly worse than NMF-KNN. GENMF performs worse than the 2PKNN method. GENMF combines the multiview NMF and KNN-based annotation methods. It only considers the semantic similarity between images. 2PKNN makes use of both the image-to-label and image-to-image similarities. FastTag performs better than JEC but worse than GENMF. Aiming to reduce the computational cost by the KNN-based methods, FastTag explores two simple linear mappings in a joint convex loss function. It can effectively deal with sparse tagged training data and rare tags. However, compared to GENMF, its goal is to learn a model that can infer the complete tags from image visual features. It does not consider the multiple information from images and labels.

Then, the proposed method is compared with MvNMF-DK, MLDL, CCA-KNN and KCCA-LP on the Corel5k dataset. According to the obtained results in Table 4, we can see that our method achieves the best results among these methods. MLDL achieves the second position, and CCA-KNN is slightly worse than MLDL. However, CCA-KNN obtains the best N plus score. Among these methods, MLDL is a dictionary-based multilabel method. It also achieves the second-best result among all the methods in the Corel5k dataset under the F1 measure because it utilizes not only label consistency and label embedding in the method, but it proves the importance of the label information in the image annotation. Our method also considers the label information. CCA-KNN is the third-best among all the methods in this dataset because of the combination of CNN visual features and the canonical correlation analysis. It maximizes the advantages of CNN features and word embedding, which not only uses the visual features but also the textual features. However, our method explores more useful information for the annotation.

Moreover, to evaluate the effectiveness of the proposed method, we conduct a set of experiments on the other public datasets IAPR TC12 and ESP. As shown in Table 4, we can see that MLDL attains the first position under the recall and F1 score measures for the IAPR TC12 dataset. It performs slightly better than the proposed method. MLDL also achieves the best precision and N plus. JEC performs worst among all the methods. This method extracts visual features and weights each feature equally. It cannot deal with the label-imbalance problem. FastTag presents a better performance in the F1 score than JEC. TagProp and KCCA-LP are both better than FastTag since FastTag only utilizes the visual features. TagProp uses multiple features and the metric learning to address the image annotation problem, which deals with the label-imbalance problem. KCCA-LP uses the CNN features and CCA to transfer the label. It considers more information from images and labels. 2PKNN, MvNMF-DK and CCA-KNN perform better than TagProp in this dataset. This is similar in the Corel5k dataset. It is worth noting that in the ESP dataset, CCA-KNN achieves the best N plus and the second-best F1 score, which is slightly worse than ours. It also has the best recall among all the methods. It is interesting that NMF-KNN performs worse than TagProp and FastTag, which means that NMF-KNN performs better in small datasets, such as Corel5k. It does not scale well in the larger datasets. Due to the smaller TPI than IAPR TC12, these methods achieve smaller recall than the recall in the IAPR TC12 dataset. Corel5k has fewer labels per image, and thus the recall can be improved in this dataset. For the proposed method, there is

an interesting result in that the Corel5k has the best F1 score. The reasons are due to the following two aspects: first, the use of multiple information from images, features and labels help to reduce the semantic gap, and thus it has rather good recall and precision in this dataset; second, to solve the image annotation problem, TG-CMTF takes advantage of the collective matrix factorization to use the intertype and intratype relationships to find more interpretable low-rank representations, which is beneficial to the annotation performance.

E. PARAMETER TUNING

The proposed method has 6 parameters: λ_G, λ_H, α, β, λ, and latent feature dimension k, where λ_G and λ_H control the contributions of the latent feature of the image-to-label relation and the latent feature of the feature-to-label relation, respectively. α, β and λ weight the contributions of image-to-image, feature-to-feature and label-to-label relations, respectively. We study the impact of these parameters one by one. When analyzing one parameter, we fix the others.

1) SENSITIVITY ANALYSIS OF λ_G

Parameter $\lambda_G$is used to control the contribution of image-to-label to the loss function. We vary the value of λ_G by fixing the other parameters as $\lambda_H = 1, \alpha = 100, \beta = 100, k = 40$ and $\gamma = 100$. We search the parameter λ_G within the set {0, 0.01, 0.1, 1, 10, 30, 50, 70, 90, 100, 200, 300, 500, 800, 1000}. In this way, we ensure that both the image-to-label and feature-to-label relation can impact the objective function; additionally, we weight the other intertype relation equally. We conduct these experiments on the Corel5k dataset. The results are shown in Figure 3(a).

As shown in Figure 3(a), we can see that when the other parameters are fixed, the performance of the proposed method increases with the increase in λ_G. When λ_G increases to a certain extent, F1 cannot be improved but decreases sharply. In this figure, when $0 \leq \lambda_G \leq 0.1$, F1 increases continuously, and when $1 \leq \lambda_G$, F1 decreases slowly because λ_G adjusts the contribution of image-to-label factorization. When $\lambda_G = 0$, we do not use the image-to-label relation, the performance cannot reach a satisfactory level because the image-to-label relation provides the semantic mapping from images to labels, which can help to bridge the semantic gap. When the ratio increases to 1, the image-to-label relation provides more useful information. When λ_G is considerably larger, the image-to-label relation leads to noisy information for the annotation. As a result, the F1 score decreases sharply. To obtain an optimal result, we set $\lambda_G = 0.1$.

2) SENSITIVITY ANALYSIS OF λ_H

Similar to λ_G, we study the impact of λ_H by fixing the other parameters and search it within the set {0, 0.01, 0.1, 1, 10, 30, 50, 70, 90, 100, 200, 300, 500, 800, 1000}. The other parameters are set to $\lambda_G = 0.1, \alpha = 100, \beta = 100$, and $\gamma = 100$, and the number of latent factors is set to $k = 40$. The results are shown in Figure 3(b). As shown in this figure, when the value of parameter $\lambda_H = 0$, the performance is not satisfactory because we do not use the feature-to-label relation at all. When λ_H increases, F1 also increases. However, when it increases to 10, the performance is stable. It is worth noting that when $\lambda_H \geq 10$, F1 is better than that of $\lambda_H \leq 10$. This indicates that a larger λ_H can result in a better F1. However, to some extent, λ_H can obtain a stable result. This may be due to the following reasons: (1) λ_H is used to regularize the sensitivity of feature-to-label relation factorization. When this parameter is small, the less information can be provided, while it increases, the importance of feature-to-label relation also increases, which helps the first factorization in the objective function to find more useful low-dimension representations and helps to connect the feature to the label also narrow the semantic gap. (2) The first term in the objective function provides some feature information. This parameter helps to provide more feature information from the viewpoint of the semantic concept. Thus, to some extent, it is enough to find the latent feature factors for the factorizations. In this paper, to obtain the best result, we set $\lambda_H = 10$.

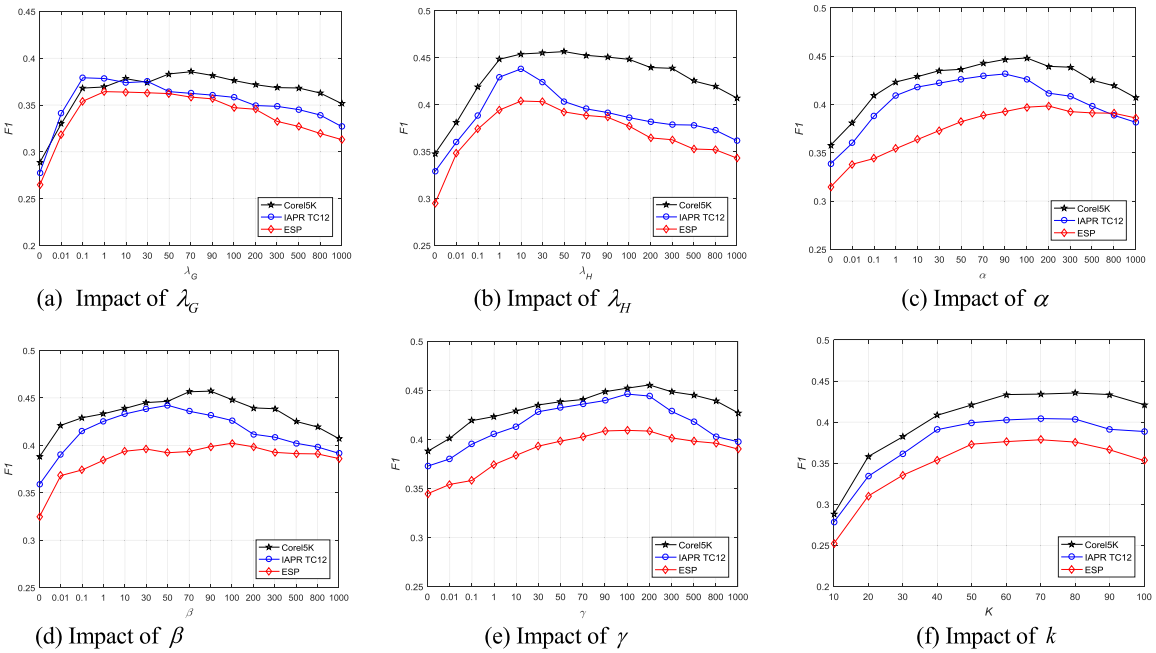
3) SENSITIVITY ANALYSIS OF α

Regularization parameter α weights the importance of the image-to-image relation to the objective function in Eq. (1). In this paper, the image-to-image relation provides visual similarity information for the factorizations. To study this parameter, we vary the value of α by fixing other parameters to $\lambda_G = 0.1$, $\lambda_H = 10$, $\beta = 100$, , $\lambda = 100$ and $k = 40$. In this study, we set the latent feature factor dimension to $k = 40$. The results are shown in Figure 3(c).

As shown in Figure 3(c), the F1 curve first increases and later becomes stable as α increases. As we know, α controls the weights on the image-to-image similarity information. If $\alpha = 0$, we do not use this type of information at all. When α increases, the performance also increases. This helps the factorizations to find more visual-based similar images. However, if this parameter continues to increase, visual similarity information is enough, and the performance becomes stable. Additionally, when the parameter is too large, the information that can be provided is limited. The performance cannot be improved significantly, but the computational cost and noisy information increase. This is shown in Figure 3(c). We can observe that when $\alpha = 70 \sim 90$, F1 achieves the best result and becomes stable. When $\alpha \geq 90$, F1 decreases slowly. To obtain an optimal result, we set $\alpha = 90$.

4) SENSITIVITY ANALYSIS OF β

We use β to control the importance of intra-feature relationships. As a regularization parameter, this parameter can help the factorizations to find the most feature similarity latent factors. We search within the following set {0, 0.01, 0.1, 1, 10, 30, 50, 70, 90, 100, 200, 300, 500, 800, 1000} and set other parameters to $\lambda_G = 0.1, \lambda_H = 10, \alpha = 90, \gamma = 100$ and $k = 40$. The results are shown in Figure 3(d). We can observe from this figure that when $\beta > 0$, the F1 score improves more significantly than that of $\beta = 0$. This proves the importance



(a) Impact of λ_G (b) Impact of λ_H (c) Impact of α

(d) Impact of β (e) Impact of γ (f) Impact of k

FIGURE 3. Sensitivity of parameters.

of feature similarity in the matrix factorizations, which can help the factorizations to find more interpretable information, especially for the image-to-feature factorization. This similarity can help the first term in the objective function to find more useful feature specific latent factors. When β continues to increase, the F1 score increases. The performance becomes stable when β increases to the range of $70 \sim 90$and then F1 increases very slowly and even decreases. Thus, we set $\beta = 70$ for a tradeoff.

5) SENSITIVITY ANALYSIS OF γ

γ controls the impact of the label-label cooccurrence information on the objective function. To determine the most appreciable γ, we fix $\lambda_G = 0.1, \lambda_H = 10, \alpha = 90, \beta = 70$ and $K = 40$ according to the previous study. Then, we search the parameter γ within the set {0, 0.01, 0.1, 1, 10, 30, 50, 70, 90, 100, 200, 300, 500, 800, 1000} and show the results in Figure 3(e).

As shown in Figure 3(e), we observe that when $\gamma = 0$, the performance is poor, which means the label-to-label semantic information is not used at all. When γ increases, F1 improves obviously. We determine that the reason is that, with γ increasing, increasing semantic information can be provided by the label graph. This can help the factorizations to build a connect among image, label and feature, which provides the method the high-level features of images. The larger γ is, the more semantic information can be provided. Nevertheless, to some extent, the semantic information becomes saturated; thus, F1 is hard to improve. Therefore, too much information may result in a decrease in performance. In Figure 6, when $\gamma \geq 200$, F1 starts to decrease. To achieve better performance, we set $\lambda = 200$in the experiments.

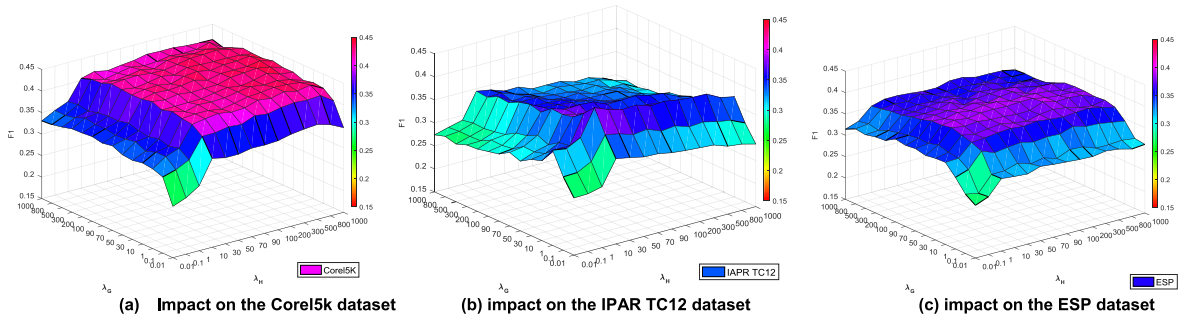
6) IMPACT OF THE NUMBER OF LATENT FEATURES k

In the proposed method, the input matrix is decomposed into three low-rank matrices. As is known, the higher the dimensionality of the low-dimensional representation, the better the approximation is. However, if the dimension is too large, it may cause more computational cost. Additionally, the dimensionality increases to some extent, the information that can be extracted, and the performance is hard to improve. To determine an appreciable dimensionality, we conduct a set of experiments fixing the parameters to $\lambda_G = 0.1, \lambda_H = 10, \alpha = 90, \beta = 70$ and $\gamma = 200$. We vary the value of k within {10, 20, 30, 40, 50, 60, 70, 80, 90, 100}. The results are shown in Figure 3(f).

It is worth noting that with the dimensionality of the latent features increasing, F1 also increases. This result agrees with the above assumption since the larger the latent features, the more information that can be represented by the latent features. The approximation will be much closer to the original matrix. However, when the dimensionality becomes large enough, no significant improvement is achieved. This is because existing latent features represent the useful information very well. Too much information may cause noise to interrupt the performance. In our empirical study, when $K \geq 50$, the performance improves slowly. Considering the efficiency of the computational cost, we set $k = 50$ in our experiments.

7) IMPACT OF COMBINATION OF IMAGE-TO-LABEL AND FEATURE-TO-LABEL RELATIONS

The parameter λ_G denotes the importance of the image-to-label relation, while λ_H weights the feature-to-label relation. To use both of the relations, the two parameters should be set with proper values. To determine the best combi-



(a) Impact on the Corel5k dataset (b) impact on the IPAR TC12 dataset (c) impact on the ESP dataset

FIGURE 4. Impact of λ_G and λ_H on the three datasets.

nation of these two types of information, we iteratively adjust one of the parameters to a fixed value within the set $\{0.01, 1, 10, 30, 50, 70, 90, 100, 200, 300, 500, 800, 1000\}$. Additionally, we set the other parameters as $\alpha = 90$, $\beta = 70$, $\gamma = 200$ and $k = 50$. The results are shown in Figure 4 (a), (b) and (c).

As shown in the figure, we observe that the optimal value can be achieved when the region is approximately at $\lambda_G = 0.1 \sim 1$ and $\lambda_H = 1 \sim 10$ for these datasets. This result indicates that the feature-to-label relation is more important than the image-to-label relation, which proves the importance of the feature-to-label connection because the feature-to-label is the only mapping from image features to labels.

From these figures, we also find that when $\lambda_G = 1$ and $\lambda_H = 10$, the performance achieves the best performance and then improves slowly. As is known, too large parameters result in more computational cost. Thus, for a tradeoff, we set $\lambda_G = 1, \lambda_H = 10$.

VI. CONCLUSION

In this paper, we investigate image annotation with a new framework, TG-CMTF. This framework utilizes multiple relationships from images, labels and features by a graph regularized collective matrix tri-factorization. It not only explores interrelations such as multiview features of images, image-to-feature and image-to-label but also the intrarelations as intraimages, intrafeatures and intralabels by three graph regularized terms. To narrow the semantic gap between low-level features and high-level semantic concepts, multiview low-level features and convolutional neural networks features are both extracted. Then, the image-to-feature relations are factorized to propagate the labels from labeled images to unlabeled images. This process explores the latent features between images and multiview features, which can be used to propagate the labels. Simultaneously, image-to-label and feature-to-label matrices are also factorized, which can help to find more meaningful latent features among images-features-labels. Moreover, the correlations among labels and the similarities among images are also introduced. These relations help to maintain the consistencies among different views and labels. Furthermore, such information improves the performance of image annotation. The promising results have proven the benefits of this framework for image annotation.

In the future, some improvements can be achieved by the following techniques. First, the multiplicative updating rules can be accelerated by a Newton updating solution. Second, deep learning-based multiview features and a word embedding model will be combined to improve the annotation performance.

REFERENCES

[1] G. Carneiro, A. B. Chan, P. J. Moreno, and N. Vasconcelos, "Supervised learning of semantic classes for image annotation and retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 3, pp. 394–410, Mar. 2007.

[2] X. Gao, X. Gao, X. Gao, and X. Gao, "Large sparse cone non-negative matrix factorization for image annotation," *ACM Trans. Intell. Syst. Technol.*, vol. 8, no. 3, p. 37, 2017.

[3] V. Maihami and F. Yaghmaee, "Automatic image annotation using community detection in neighbor images," *Phys. A, Stat. Mech. Appl.*, vol. 507, pp. 123–132, Oct. 2018.

[4] L. Sun, H. Ge, S. Yoshida, Y. Liang, and G. Tan, "Support vector description of clusters for content-based image annotation," *Pattern Recognit.*, vol. 47, no. 3, pp. 1361–1374, 2014.

[5] N. Zhou, W. K. Cheung, G. Qiu, and X. Xue, "A hybrid probabilistic model for unified collaborative and content-based image tagging," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 7, pp. 1281–1294, Jul. 2011.

[6] B. Demir and L. Bruzzone, "A novel active learning method in relevance feedback for content-based remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2323–2334, May 2015.

[7] L. Cao, J. Luo, L. Feng, and T. S. Huang, "Heterogeneous feature machines for visual recognition," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 1095–1102.

[8] F. Wu, Y. Han, T. Qi, and Y. Zhuang, "Multi-label boosting for image annotation by structural grouping sparsity," in *Proc. 18th Int. Conf. Multimedea*, Oct. 2010, pp. 15–24.

[9] H. Ma, J. Zhu, M. R.-T. Lyu, and I. King, "Bridging the semantic gap between image contents and tags," *IEEE Trans. Multimedia*, vol. 12, no. 5, pp. 462–473, Aug. 2010.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. 25th Int. Conf. Neural Inf. Process. Syst.*, Dec. 2012, pp. 1097–1105.

[11] Y. Ma, Y. Liu, Q. Xie, and L. Li, "CNN-feature based automatic image annotation method," *Multimedia Tools Appl.*, vol. 78, no. 3, pp. 3767–3780, 2019.

[12] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2014, pp. 512–519.

[13] R. Jin, J. Y. Chai, and L. Si, "Effective automatic image annotation via a coherent language model and active learning," in *Proc. ACM Int. Conf. Multimedia*, Oct. 2004, pp. 892–899.

[14] A. Budanitsky and G. Hirst, "Evaluating wordnet-based measures of lexical semantic relatedness," *Comput. Linguistics*, vol. 32, no. 1, pp. 13–47, 2006.

[15] H. Wang, H. Huang, and C. Ding, "Image annotation using bi-relational graph of images and semantic labels," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 793–800.

[16] Y. Verma and C. V. Jawahar, "Image annotation by propagating labels from semantic neighbourhoods," *Int. J. Comput. Vis.*, vol. 121, no. 1, pp. 126–148, 2016.
[17] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, Oct. 1999.
[18] H. Liu, Z. Wu, X. Li, D. Cai, and T. S. Huang, "Constrained nonnegative matrix factorization for image representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1299–1311, Jul. 2012.
[19] L. Jing, C. Zhang, and M. K. Ng, "SNMFCA: Supervised NMF-based image classification and annotation," *IEEE Trans. Image Process.*, vol. 21, no. 11, pp. 4508–4521, Nov. 2012.
[20] F. A. González, J. C. Caicedo, O. Nasraoui, and J. Ben-Abdallah, "NMF-based multimodal image indexing for querying by visual example," in *Proc. ACM Int. Conf. Image Video Retr.*, Jul. 2010, pp. 366–373.
[21] R. Merris, "Laplacian matrices of graphs: A survey," *Linear Algebra Appl.*, vols. 197–198, no. 2, pp. 143–176, 1994.
[22] D. Cai, X. He, J. Han, and T. S. Huang, "Graph regularized nonnegative matrix factorization for data representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1548–1560, Aug. 2011.
[23] C. Ding, T. Li, W. Peng, and H. Park, "Orthogonal nonnegative matrix t-factorizations for clustering," in *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2006, pp. 126–135.
[24] L. Wu, R. Jin, and A. K. Jain, "Tag completion for image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 716–727, Mar. 2013.
[25] G. Nasierding and A. Z. Kouzani, "Empirical study of multi-label classification methods for image annotation and retrieval," in *Proc. Int. Conf. Digit. Image Comput., Techn. Appl.*, Dec. 2010, pp. 617–622.
[26] M. Lienou, H. Maitre, and M. Datcu, "Semantic annotation of satellite images using latent Dirichlet allocation," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 1, pp. 28–32, Jan. 2010.
[27] D. Putthividhy, H. T. Attias, and S. S. Nagarajan, "Topic regression multi-modal latent Dirichlet allocation for image annotation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3408–3415.
[28] D. Tian and Z. Shi, "Automatic image annotation based on Gaussian mixture model considering cross-modal correlations," *J. Vis. Commun. Image Represent.*, vol. 44, pp. 50–60, Apr. 2017.
[29] J. Tian, Y. Huang, Z. Guo, X. Qi, Z. Chen, and T. Huang, "A multi-modal topic model for image annotation using text analysis," *IEEE Signal Process. Lett.*, vol. 22, no. 7, pp. 886–890, Jul. 2015.
[30] V. Lavrenko, R. Manmatha, and J. Jeon, "A model for learning the semantics of pictures," in *Proc. Adv. Neural Inf. Process. Syst.*, 2004, pp. 553–560.
[31] Y. Liu, R. Jin, and L. Yang, "Semi-supervised multi-label learning by constrained non-negative matrix factorization," in *Proc. Nat. Conf. Artif. Intell. 18th Innov. Appl. Artif. Intell. Conf.*, Boston, MA, USA:, Jul. 2006, pp. 421–426.
[32] F. Monay and D. Gatica-Perez, "PLSA-based image auto-annotation: Constraining the latent space," in *Proc. 12th Annu. ACM Int. Conf. Multimedia*, Oct. 2004, pp. 348–351.
[33] Y. Liu, K. Wen, Q. Gao, X. Gao, and F. Nie, "SVM based multi-label learning with missing labels for image annotation," *Pattern Recognit.*, vol. 78, pp. 307–317, Jun. 2018.
[34] Y. Verma and C. V. Jawahar, "Exploring SVM for image annotation in presence of confusing labels," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, Sep. 2013, pp. 1–11.
[35] R. Hong, M. Wang, Y. Gao, D. Tao, X. Li, and X. Wu, "Image annotation by multiple-instance learning with discriminative feature mapping and selection," *IEEE Trans. Cybern.*, vol. 44, no. 5, pp. 669–680, May 2014.
[36] M.-L. Zhang and L. Wu, "Lift: Multi-label learning with label-specific features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 1, pp. 107–120, Jan. 2015.
[37] A. Makadia, V. Pavlovic, and S. Kumar, "A new baseline for image annotation," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 316–329.
[38] L. Xi, R. Liu, L. Fei, and Q. Cao, "Graph-based dimensionality reduction for KNN-based image annotation," in *Proc. 21st Int. Conf. Pattern Recognit.*, Nov. 2012, pp. 1253–1256.
[39] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid, "TagProp: Discriminative metric learning in nearest neighbor models for image auto-annotation," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep./Oct. 2010, pp. 309–316.
[40] Y. Verma and C. V. Jawahar, "Image annotation using metric learning in semantic neighbourhoods," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 836–849.
[41] Y. Ma, Q. Xie, Y. Liu, and S. Xiong, "A weighted KNN-based automatic image annotation method," in *Neural Computing and Applications*. London, U.K.: Springer-Verlag, 2019, pp. 1–12.
[42] T. Uricchio, L. Ballan, L. Seidenari, and A. Del Bimbo, "Automatic image annotation via label transfer in the semantic space," *Pattern Recognit.*, vol. 71, pp. 144–157, Nov. 2017.
[43] G. Cao, A. Iosifidis, K. Chen, and M. Gabbouj, "Generalized multi-view embedding for visual recognition and cross-modal retrieval," *IEEE Trans. Cybern.*, vol. 48, no. 9, pp. 2542–2555, Sep. 2018.
[44] J. Gao, J. Han, J. Liu, and C. Wang, "Multi-view clustering via joint nonnegative matrix factorization," in *Proc. 13th SIAM Int. Conf. Data Mining*, May 2013, pp. 252–260.
[45] M. M. Kalayeh, H. Idrees, and M. Shah, "NMF-KNN: Image annotation using weighted multi-view non-negative matrix factorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 184–191.
[46] R. Rad and M. Jamzad, "Image annotation using multi-view non-negative matrix factorization with different number of basis vectors," *J. Vis. Commun. Image Represent.*, vol. 46, pp. 1–12, Jul. 2017.
[47] Y. Luo, T. Liu, D. Tao, and C. Xu, "Multiview matrix completion for multilabel image classification," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2355–2368, Aug. 2015.
[48] J. Farquhar, D. Hardoon, H. Meng, J. S. Shawe-Taylor, and S. Szedmák, "Two view learning: SVM-2K, Theory and Practice," in *Proc. Adv. Neural Inf. Process. Syst.*, 2005, pp. 355–362.
[49] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Comput.*, vol. 16, no. 12, pp. 2639–2664, 2004.
[50] J. Kludas, E. Bruno, and S. Marchand-Maillet, "Information fusion in multimedia information retrieval," in *Adaptive Multimedia Retrieval: Retrieval, User, and Semantics*, vol. 4918. Berlin, Germany: Springer, 2007, pp. 147–159.
[51] L. Zheng, S. Wang, L. Tian, F. He, Z. Liu, and Q. Tian, "Query-adaptive late fusion for image search and person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1741–1750.
[52] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.
[53] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
[54] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, Apr. 2015, pp. 1–14.
[55] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 580–587.
[56] S. Boyd, L. Vandenberghe, and L. Faybusovich, "Convex optimization," *IEEE Trans. Autom. Control*, vol. 51, no. 11, p. 1859, Nov. 2006.
[57] Q. Gu and J. Zhou, "Co-clustering on manifolds," in *Proc. ACM SIGKDD 15th Int. Conf. Knowl. Discovery Data Mining*, Jul. 2009, pp. 359–368.
[58] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2000, pp. 556–562.
[59] Z. Yang and E. Oja, "Linear and nonlinear projective nonnegative matrix factorization," *IEEE Trans. Neural Netw.*, vol. 21, no. 5, pp. 734–749, May 2010.
[60] P. Duygulu, K. Barnard, J. F. G. de Freitas, and D. A. Forsyth, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," in *Proc. Eur. Conf. Comput. Vis.*, 2002, pp. 97–112.
[61] L. von Ahn and L. Dabbish, "Labeling images with a computer game," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, Apr. 2004, pp. 319–326.
[62] M. Grubinger, P. Clough, H. Müller, and T. Deselaers, "The IAPR TC-12 benchmark: A new evaluation resource for visual information systems," in *Proc. Int. Workshop OntoImage*, vol. 2, May 2006, pp. 1–55.
[63] M. Chen, A. Zheng, and K. Q. Weinberger, "Fast image tagging," in *Proc. Int. Conf. Int. Conf. Mach. Learn.*, 2013, pp. 1–9.
[64] H. Ge, Z. Yan, J. Dou, Z. Wang, and Z. Wang, "A semisupervised framework for automatic image annotation based on graph embedding and multiview nonnegative matrix factorization," *Math. Probl. Eng.*, vol. 2018, Jun. 2018, Art. no. 5987906.
[65] X.-Y. Jing, F. Wu, Z. Li, R. Hu, and D. Zhang, "Multi-label dictionary learning for image annotation," *IEEE Trans. Image Process.*, vol. 25, no. 6, pp. 2712–2725, Jun. 2016.
[66] V. N. Murthy, S. Maji, and R. Manmatha, "Automatic image annotation using deep learning representations," in *Proc. 5th ACM Int. Conf. Multimedia Retr.*, Jun. 2015, pp. 603–606.