

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

# Automatic Image Annotation by Sequentially Learning from Multi-Level Semantic Neighborhoods

HOUJIE LI<sup>1</sup>, WEI LI<sup>2,3</sup>, HONGDA ZHANG<sup>2</sup>, XIN HE<sup>1</sup>, MINGXIAO ZHENG<sup>2</sup>, AND HAIYU SONG<sup>2,3</sup>.

<sup>1</sup>School of Information and Communication Engineering, Dalian Minzu University, Dalian 116600, China

<sup>2</sup>School of Computer Science and Engineering, Dalian Minzu University, Dalian 116600, China

<sup>3</sup>College of Traffic, Jilin University, Changchun 130022, China

Corresponding author: Haiyu Song (shy@dlmu.edu.cn).

This work was supported in part by the Liaoning Natural Science Foundation Projects under Grant 2019-ZD-0182 and 20180550625, in part by the National Science Foundation Projects under Grant 62072152, and in part by the National Statistical Science Research Project under Grant 2019LY12

**ABSTRACT** Automatic image annotation is a key technology in image understanding and pattern recognition, and is becoming increasingly important in order to annotate large-scale images. In the past decade, the nearest neighbor model-based AIA methods have been proved to be the most successful in all classical models. This model has four major challenges including semantic gap, label-imbalance, wider range labels, and weak-labeling. In this paper, we propose a novel annotation model based on three-pass KNN (k-Nearest Neighbor) to address the aforementioned challenges. The key idea is to identify appropriate neighbors at each pass KNN. In the first pass KNN, we identify the several most relevant categories based on label feature rather than visual feature as traditional models. In the second pass KNN, we determine the relevant images based on multi-modal (visual and textual label) embedding features. As the test image has not been annotated with any label, we propose a pre-annotation strategy before image annotation to improve the semantic level. In the third pass KNN, we capture relevant labels from semantically and visually similar images and propagate them to the given unlabeled image. In contrast with traditional nearest neighbor based methods, our method can inherently alleviate the problems of semantic gap, label-imbalance, and wider range labels. In addition, to alleviate the issue of weak-labeling, we propose label refinement for training images. Extensive experiments on three classical benchmark datasets demonstrate that the proposed method significantly outperforms the state-of-the-art in terms of per-label and per-image metrics.

**INDEX TERMS** Automatic image annotation, semantic gap, nearest neighbor, weak-labeling.

## I. INTRODUCTION

WITH the prevalence of digital photography and social networks in our daily lives, billions of images are generated and shared on the Internet. Users have access to a flood of images, making it a challenge to retrieve and manage the ones they care about from this vast ocean of available visual data [1]. Automatic image annotation (AIA) is the task of automatically assigning several textual labels to a given image based on its semantics. Recently, the AIA has been an active research topic in the fields of computer vision and machine learning due to its great potential applications in image retrieval, image classification, image understanding, and image management [2]–[4].

In the past 20 years, a considerable amount of research

effort has been made to devise automatic image annotation models. Some representative AIA approaches have been proposed and great achievements have been made, such as MBRM [5], JEC [6], 2PKNN [2], and D<sup>2</sup>IA [7]. Recently, significant advances have been achieved on large-scale image recognition tasks [8], with deep learning models such as Convolutional Neural Network (CNN) and Generative Adversarial Network (GAN). In comparison with image recognition, image annotation is a more challenging task, since it is a multi-label multi-class classification problem [8], instead of a single-label multi-class classification problem as in image recognition. The four most difficult challenges of AIA are semantic gap, label-imbalance, weak-labeling, and wider range labels [2]. The semantic gap is the semantic

difference between image low-level features represented by machines and high-level human perceptions used to perceive the image. The label-imbalance problem means there exists a high variance among the number of images corresponding to different labels, and this problem is quite common when the size of a dataset or label vocabulary is large. Today, the label-imbalance is a pending issue. Weak-labeling means that manual annotations are noisy, irrelevant, or incomplete. Weak-labeling will also cause poor-labeling. The wider range labels mean that labels in image annotation can refer to a much wider and more diverse range of concepts or drastically different levels of abstraction, such as concrete visual objects (cat, train), scenes (beach, city), amorphous background elements (sky, grass), or abstract concepts (scary, serene) [1], [9]. The traditional CNNs, designed for single-label image classification, are thus unsuitable for image annotation task because they fail to provide rich representations at different abstraction scales.

During the past decade, there have been several efforts for addressing such issues as MBRM, JEC, CNN+WARP [10], LTN [11], TagProp [12], 2PKNN, CCA-KNN, D<sup>2</sup>IA, and CNN-RNN [13]. Benefitting from deep learning features (such as CNN and GAN), most models based on deep learning can reduce the semantic gap, although the issue has not been resolved thoroughly. In single-label image classification, deep learning features are able to address the issue of the semantic gap. CNN has shown great performance as general feature representations for object recognition applications. However, for multi-label images that contain multiple objects from different categories, scale and location, global CNN features are not optimal. Some models close to success (TagProp, KCCA-2PKNN, SKL-CRM, SVM-DMBRM [14], and 2PKNN) are able to address the semantic gap problem with computationally expensive metric learning. Some models based on nearest neighbors, i.e. 2PKNN, can alleviate the label-imbalance by paying more attention to rare labels, which always improve the performance of the low frequency words by sacrificing high frequency words. Some researches attempt to provide multi-level deep features to provide high-quality image features suitable for image annotation [9], [15]. Nevertheless, the most important challenge, i.e. weak-labeling problem, has never been tackled explicitly. In summary, the AIA is still a difficult and challenging task [2].

In this paper, we propose a novel image annotation method based on nearest neighbors to address the problems above mentioned. Different from the existing methods based on nearest neighbors, our method sequentially learns from multi-level semantic neighborhoods rather than a single neighborhood as the existing methods. First, we complete labels to the training dataset by propagating from neighbors to overcome the problem of the weak-labeling. Second, we divide the training images which have similar textual label features into a semantic neighbor group by N-cut algorithm. Each group is considered as a category. Third, we pick up top N similar images from the three most similar categories according to their similarities in KCCA space, which is trained by

visual features and completed labels of the training image dataset. The selected neighbors form a category-level similar neighborhood. Fourth, pick up top M similar images from category-level similar neighborhood according to visual similarity, and propagate labels from category-level and object-level similar neighbors by using their similarity as weight.

Our main contributions are as follows: (1) We proposed a novel annotation method based on three-pass KNN, which can accurately capture the relevant categories, relevant images, and appropriate labels for a given test image. (2) We proposed multi-level semantic neighborhoods rich representations at different abstraction scales, suitable for image annotation tasks. (3) We proposed a pre-annotation strategy for the unlabeled test image to perform multi-modal image retrieval so as to reduce the semantic gap. (4) We proposed label refinement for training images based on their textual label similarities to alleviate the issue of weak-labeling. (5) Our proposed model can alleviate the problem of the label-imbalance by enhancing rare labels without sacrificing frequent labels. Extensive experiments are carried out on three benchmark datasets: Corel5k, ESP Game, and IAPR TC-12. The experimental results demonstrate that our model outperforms the state-of-the-art alternatives.

## II. RELATED WORKS

A large number of automatic image annotation models have been developed. In this section, we briefly review the closely related and representative works.

### A. GENERATIVE MODELS

The generative model-based AIA methods are quite popular in the early 21st century, and great achievements have been made. In the model training stage, the generative model aims at learning a joint distribution between visual and label features so that the learned model can predict the conditional probability of labels for features of a test image. The generative models focus on maximizing the generative likelihood of image features and labels. The generative models used for AIA mainly consist of relevance models, mixture models, and topic models. The representative models include CMRM, CRM, MBRM, and PLSA-WORDS. These models are usually expensive or require simplifying assumptions that can be suboptimal for predictive performances [16].

### B. DISCRIMINATIVE MODELS

Discriminative models consider image annotation as a multi-label multi-class classification problem. These models regard each label as an independent class. A separate classifier is trained for each label with visual features of training images, and then the trained classifier can predict particular labels for a given test image. Most of discriminative models are based on support vector machine (SVM) or its variants [17]. The other representative models include SML [18], DMBRM, and SVM-DMBRM. The SML model learns class-specific distributions for each label. The SML model treats the image annotation as a multi-classification problem and learns a

class-specific distribution for each label [18]. The SVM-DMBRM model is a hybrid method that takes full advantages of the merits of both generative and discriminative models [14]. While SVM tries to solve the weak-labeling issue, DMBRM strives to solve the class-imbalance issue. However, these multi-label classification approaches are unscalable to a large number of labels [17].

### C. NEAREST NEIGHBOR BASED MODELS

The nearest neighbor based models have become a more and more widely used method for AIA due to their effectiveness and simplicity. These models explore the visual similarities between a test image and training images, and finally assign labels to the test image by sorting the scores of neighbors of visually similar images [19]. The representative models based on nearest neighbor include JEC [6], TagProp [12], and 2PKNN [2].

The Joint Equal Contribution (JEC) model is one of the most classical nearest-neighbor models [6]. The JEC model utilizes various low-level image features and a simple combination of basic distance measures to find the nearest neighbors of a given image. It creates a family of very simple and intuitive baseline methods for image annotation [20]. Guillaumin et al. presented the TagProp [12] method, which learns the weight of each feature group and uses the label relevance prediction to annotate images [2]. The TagProp promotes rare labels and penalizes frequent ones by training a logistic model, which alleviates the class-imbalance problem. Its great achievement largely benefits from metric learning.

The 2-pass k-nearest neighbor (2PKNN) model represents a classical solution to solve problems related to the label-imbalance and the weak-labeling [2]. It is a two-pass variant of the traditional KNN. In the first pass, the image-label similarity is used, while image-image similarity is used in the second pass. It identifies the k most similar semantic neighbor images for each label in the vocabulary. Due to its successfully solving the label-imbalance problem, the 2PKNN makes great achievements and is still one of the most influential image annotation approaches.

### D. DEEP NEURAL NETWORK BASED MODELS

Recently, Convolutional Neural Networks (CNNs) have shown great performance in many computer vision tasks (i.e. image recognition) by extracting end-to-end feature vectors from original images [10], [21]–[24]. Most of the deep learning-based AIA approaches are based on convolutional neural network (CNN) [10], [11], [24], [25], and features are always extracted from pre-trained AlexNet, VGGNet network [11], [16], [24], [25], and ResNet.

The conventional deep networks can be subjected to the decayed performance if we have insufficient training examples. Shu proposed weakly-shared Deep Transfer Networks (DTNs) to mitigate the problem by bringing in rich labels from the text domain [26]. The proposed model can translate cross-domain information from text to image. Tang proposed a novel generalized deep transfer networks (DTNs), capa-

ble of transferring label information across heterogeneous domains, textual domain to visual domain. The proposed framework has the ability to adequately mitigate the problem of insufficient training images by bringing in rich labels from the textual domain [27].

The CNN+WARP [10], proposed by Jia (the creator of Caffe), is the first attempt to leverage CNN features to solve the image annotation task. The CNN+WARP adopts a weighted approximate ranking loss function for training to promote image annotation performances. The VCCA [28], an image annotation method based on a deep multi-view learning model, extends linear CCA to nonlinear observation models parameterized by deep neural networks.

Different from these classical end-to-end deep learning features, Yu proposed to extract middle-level features from a deep learning model to accurately depict semantic concepts for image annotation [15]. This model can improve annotation performance with expensive time and space cost. Recently, the CNN-RNN (convolutional and recurrent neural networks) encoder-decoder architectures are jointly adopted to image understanding, where the CNN subnetwork encodes the input pixels of images into visual features, and the RNN subnetwork decodes the visual feature into a label prediction path [29], [30]. The CNN-RNN model uses the output fusion to merge CNN output features and RNN output [29]. The D<sup>2</sup>IA, image annotation method based on generative adversarial network (GAN) model, aims to create semantically relevant, yet distinct and diverse labels [7].

MangoNet [31] is a novel deep learning based image annotation model that combines co-attention mechanism and graph convolutional network (GCN). It explores image neighbors by measuring their metadata similarities and utilizes a graph network to model the correlations between the target image and its neighbors. To accurately capture the visual clues from the neighborhood, a co-attention mechanism is introduced to leverage the visual attention within the neighborhood. However, the GCN model increases the space and time complexities, which will be unfavorable to apply to large-scale databases.

Despite their relative success, most of deep learning based models suffer from the single abstraction level. As we knew, the labels in AIA are always much wider and more diverse range of visual objects or abstract concepts with different abstraction levels, while the ones in image recognition (classification) are always concrete visual objects with the same level. Most CNN models, designed originally for the image classification task, are unsuitable for the image annotation task because they fail to provide rich representations at different abstraction scales. As a result, only CNN features could alleviate the problem of the semantic gap in AIA rather than solve the problem thoroughly.

## III. OUR PROPOSED METHOD

### A. OUR FRAMEWORK

To resolve problems of the weak-labeling, models close to success focus on metric learning (such as TagProp and SVM-

DMBRM), which always require computationally expensive metric learning approaches. 2PKNN propose a novel two-pass KNN solution to address the issue of the label-imbalance by considering the image-label similarity and image-image similarity in the two passes, respectively. Although 2PKNN significantly improves the per-label performance, it always sacrifices high frequent labels. In fact, 2PKNN could not improve the annotation performance as shown by per-label evaluation metrics.

Figure 1 shows our proposed framework for automatic image annotation. The proposed framework is composed of two main components, i.e. training and testing processes. The training process includes label refinement, category generation, and KCCA model training and embedding features.

To resolve the problems of the weak-labeling and the label-imbalance, we propose a novel image annotation method based on nearest neighbors. In contrast with traditional NN based models and classical 2PKNN which need computationally expensive metric learning, we propose a novel and simple label refinement to address the weak-labeling. Rather than in traditional visual feature space, our proposed method refines labels for all training images in the label feature space, which can inherently address the problem of the semantic gap. Our proposed method divides the images which have similar features in the label feature space into a semantic neighbor group called a category. Our proposed method maps visual feature vectors extracted by deep learning architecture (pre-trained VGG-16), and refines label vectors to a common feature space by the KCCA model. New visual features and new label features are used in the test stage.

Similar to 2PKNN, our proposed method is a three-pass variant of the traditional KNN. Given a test image, in the first KNN, our method aims to find the  $K1$  most relevant categories based on label features. In the second KNN, we select the  $K2$  most similar images from each relevant category and combine them into a single neighborhood including relevant images. These relevant images are similar to the test image in both embedding visual feature and embedding label feature. In the third KNN, based on two steps, we find the  $K3$  most visually similar images and propagate a fixed number of labels to the test image according to their original visual feature similarities.

## B. LABEL REFINEMENT

To alleviate the shortcoming of the weak-labeling, most methods devise sophisticated models with expensive time and space cost in annotation process. Tang proposed a novel tri-clustered tensor completion framework to collaboratively explore these three kinds of information to improve the performance of social image tag refinement [32]. Tang proposed a novel Social anchor-Unit GrAph Regularized Tensor Completion (SUGAR-TC) method to efficiently refine the tags of social images, which is insensitive to the scale of data [33].

Our method compensates some appropriate labels for each training image based on its neighbors' labels. To start with,

our method directly completes labels for each training image in the label feature space. More specifically, if the associated textual labels associated with an image ( $I_i$ ) can be considered as another modal feature for the image, the textual feature vector of the image can be represented as follows:

$$t_i = [P(w_1 | I_i), \dots, P(w_k | I_i), \dots, P(w_M | I_i)] \quad (1)$$

where  $M$  is the volume of labels (e.g. 260 for Corel5k),  $P(w_k | I_i) = \phi(w_k \in W_i)$  denotes the presence/absence of label  $w_k$  in the label set  $W_i$  of image  $I_i$ , with  $P(w_k | I_i)$  being 1 if the image  $I_i$  has been manually annotated with the corresponding  $k$ -th label and 0 otherwise.

For each training image, if the number of its initial labels is smaller than the target number ( $M$ ), we can compensate several labels to the image by propagating labels from the neighbourhood in the label feature space. The similarity measure between the image  $I_i$  and the image  $I_j$  is based on  $L2$  distance, as follows:

$$\text{sim}_{\text{text}}(I_i, I_j) = \exp(-\text{Dist}_{L2}(t_i, t_j)) \quad (2)$$

We choose  $K$  neighbor training images in the textual label feature space for each current image  $I_i$  and rank the labels for image  $I_i$  according to their probability scores of:

$$P(I_i | w_k) = \sum_{j=1}^K \text{sim}_{\text{text}}(I_i, I_j) \times \sigma(w_k \in W_i) \quad (3)$$

where  $\text{sim}_{\text{text}}(I_i, I_j)$  is the textual similarity between  $I_i$  and  $I_j$  (as shown in Equation 2). Based on probability theory, the probability of assigning a label  $w_k$  to  $I_i$  can be defined the posterior probability as follows:

$$P(w_k | I_i) = \frac{P(I_i | w_k)P(w_k)}{P(I_i)} \quad (4)$$

where  $P(w_k)$  is the probability of the label  $w_k$ . The best label for the test image  $I_i$  will be given by the following:

$$y^* = \arg \max_k P(w_k | I_i) \quad (5)$$

We consider the top  $M - |I_i|$  labels as refined labels for the current image  $I_i$ ,  $|I_i|$  is the number of original manual labels associated with the image  $I_i$ . In addition, the posterior probability of specific label  $w_i$ , computed by Equation 4, is considered as its confidence in the image's label vector. After label refinement, several zero elements of the textual label feature vector of the image (as shown in Equation 1) are replaced by non-zero probability scores of corresponding labels. As a result, the refined label vector is real-valued rather than discrete (or binary) as the original label feature.

In the training stage, the proposed method divides the images which have similar refined label features into the same category by using k-means algorithm. The center of a category ( $k$ ) is defined as the mean of all images' label features in this category, denoted as:

$$C_k = \frac{1}{N_k} \sum_{i=1}^{N_k} t_i \quad (6)$$

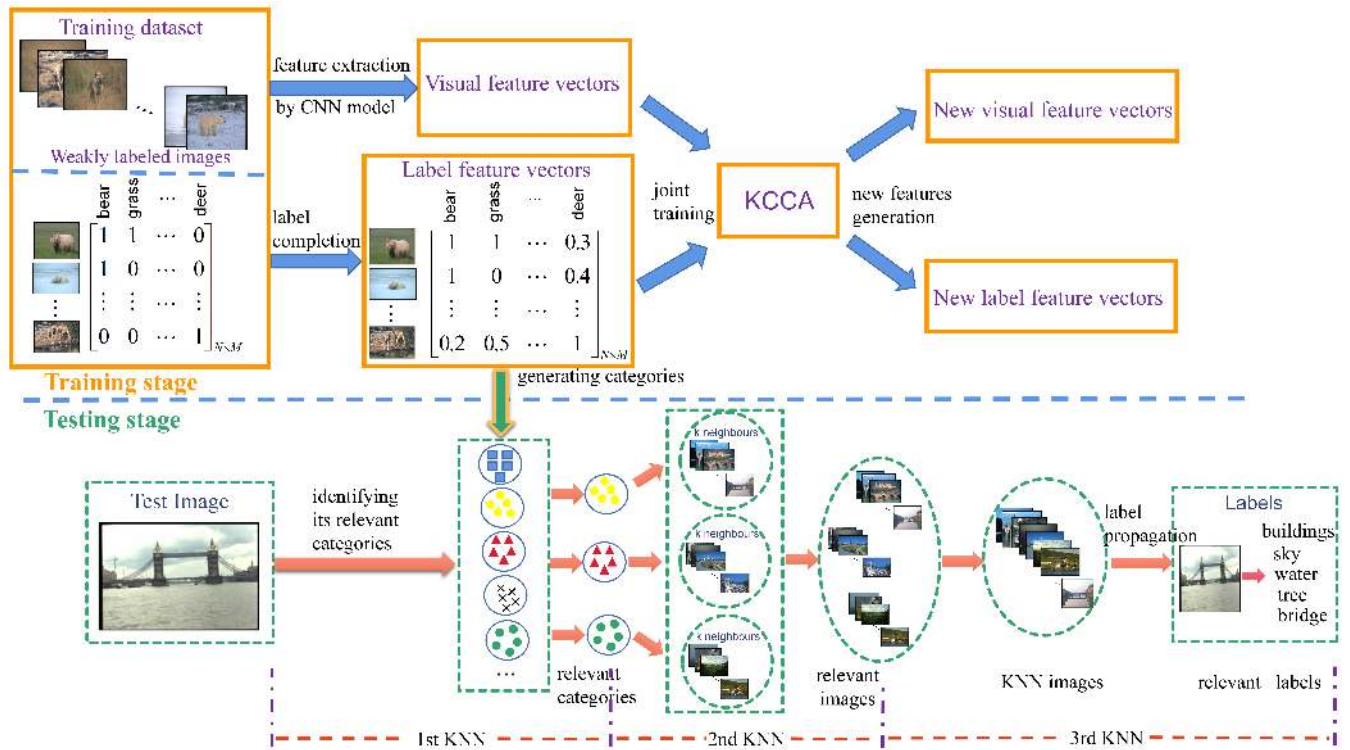


FIGURE 1. The proposed annotation framework.

### C. FEATURE EXTRACTION AND REPRESENTATION

The feature vectors in our method are different at different stages. The original visual feature is extracted by pre-trained architecture (VGG-16). We also consider the label information associated with the image as its another modal feature.

To promote the semantic level of the image feature vector, our model utilizes semantic embedding to properly map refined labels and visual features to a meaningful semantic space by using kernelized canonical correlation analysis (KCCA). In the training stage, the KCCA model can be learned from the original visual feature and the refined label feature. Then, the learned KCCA model can map the original visual feature and the textual label feature to a common meaningful semantic space, where the (embedding) new visual feature and the (embedding) new label feature can be generated.

Given the two views (visual modality and textual label modality) of the images, a common representation can be constructed by KCCA model. KCCA seeks to utilize images consisting of paired views to simultaneously find projections from each feature space so that the correlation between the projected representations is maximized. For given  $N$  training pairs of visual and refined label features  $\{(v_1, t_1), \dots, (v_i, t_i), \dots, (v_N, t_N)\}$ , the idea is to simultaneously find directions  $w_v$  and  $w_t$  that maximize the correlation of the projections of  $\phi_v$  onto  $w_v$  and  $\phi_t$  onto  $w_t$  [16], [24]. The  $\phi_v$  and  $\phi_t$  mapping is achieved using kernel function  $K_v(v_i, v_j) = \phi_v(v_i)^\top \phi_v(v_j)$  and  $K_t(t_i, t_j) = \phi_t(t_i)^\top \phi_t(t_j)$ . Thus, KCCA is to search for

solutions of  $w_v$  and  $w_t$  as a linear combination of the training data:

$$w_v = \sum_{i=1}^N \alpha_i \phi_v(v_i) \quad (7)$$

$$w_t = \sum_{i=1}^N \beta_i \phi_t(t_i) \quad (8)$$

The objective of KCCA is thus to identify the weights  $\alpha, \beta \in R^N$  that maximize the objective function [24]:

$$\alpha^*, \beta^* = \arg \max_{\alpha, \beta} \frac{\alpha^\top K_v K_t \beta}{\sqrt{\alpha^\top K_v^2 \alpha \beta^\top K_t^2 \beta}} \quad (9)$$

where  $K_v$  and  $K_t$  denote the  $N \times N$  visual and textual kernel matrices over a sample of  $N$  pairs, respectively. The solution yields top  $M$  eigenvectors  $A_x = [\alpha_1 \dots \alpha_M]$  and  $B_y = [\beta_1 \dots \beta_M]$  which form the projection matrix. Given any image, we can project its visual feature  $v$  onto  $A_x$ , and its textual label feature  $t$  onto  $B_y$ . The new embedding feature can be defined as follows:

$$v^{KCCA} = (v - \mu_v) A_x \quad (10)$$

$$t^{KCCA} = (t - \mu_t) B_y \quad (11)$$

**Input:** 1)  $Q$ : the binary image-label matrix,  $Q \in B^{N \times M}$ ,  $N$  is the number of the training image dataset,  $M$  is the number of labels in the image dataset. 2)  $trainingImgSet$ : the training image dataset.

- 1: assign  $Q$  to  $NQ$ ,  $NQ \in R^{N \times M}$ ,  $NQ$  is a real-valued image-label matrix
- 2: **for**  $I_i$  in  $trainingImgSet$  **do**
- 3:   set  $sim_{vector} = \phi$
- 4:   set  $neighborhood = \phi$
- 5:   **for**  $I_j$  in  $trainingImgSet$  **do**
- 6:     compute  $sim = sim_{text}(I_i, J)$  with Equation (2)
- 7:     set  $sim_{vector}(j) = sim$
- 8:   **end for**
- 9:   sort  $sim_{vector}$  in descending order
- 10:   assign top  $K$  elements of  $sim_{vector}$  to  $sim_{KNN}$ , and their corresponding images to the neighborhood
- 11:   **for**  $k$  in  $[1, M]$  **do**
- 12:     compute  $P(w_k | I)$  with Equation (4)
- 13:   **end for**
- 14:   sort  $P(w_k | I)$  in descending order
- 15:   assign the largest  $(5-|I_i|)$  probability scores of  $P(w_m | I_i)$  to  $NQ(i)$ .
- 16: **end for**=0

**Output:**  $NQ^{N \times M}$ : the refined image-label matrix.

#### D. LABEL PROPAGATION BASED ON MULTI-LEVEL SEMANTIC NEIGHBORHOODS

The image annotations (labels) always cover drastically different levels of abstraction semantic concepts including image categories, scenes, abstract concepts, and concrete visual objects [9]. Currently, most methods based on the nearest neighbor model measure the similarities between the test image and training images only based on single-level visual or semantic features, which fail to provide rich representations at different abstraction scales and could not depict multi-level semantic concepts. Consequently, many noisy images, their content irrelevant to the test image, are considered as neighbors and involved to labels propagation based on neighbors. Since noisy images may worsen the annotation performance, it is necessary to get rid of them. In contrast with traditional annotation models based on single neighborhood directly selected from the whole training image dataset, we propose a novel annotation method based on multi-level semantic neighbors.

Given an unlabeled test image, we proposed a pre-annotation strategy before image annotation. The pre-annotation strategy assumes several labels by propagating visual neighbors' labels to the test image by weighted KNN, whose weights are visual similarities between the image and neighbors. Similar to the label refinement method (Equation 2-5), we can assign 5 labels to the test image. The only difference between the pre-annotation and the label refinement is that the pre-annotation step determines neighbors based on original visual features instead of textual label features.

We use the pre-annotation labels as the test image's labels in the following process until the final annotation labels are predicted.

Our proposed method includes three KNN steps. First, our proposed model computes image-category similarities to determine certain categories semantically relevant to the test image. The image-category similarity is the similarities of label vectors between the test image and all categories' centre. If the image-category similarity is larger than the specified threshold, the category is considered as relevant one. Second, compute multi-modal (including new visual feature and new label feature) image-image similarity to capture visually and semantically similar neighbors in each relevant category. Third, combine KNN images of all relevant categories into a single neighbourhood set including relevant images. Finally, assign the  $N$  most relevant labels to the test image based on visual similarities between the test image and relevant images.

In the first KNN, given a test image, our goal is to select the  $K1$  most relevant categories. We can define the similarity between a test image ( $I$ ) and a category ( $m$ ) as follows:

$$sim_{text}(I, C_m) = exp(-Dist_{L2}(t_i, C_m)) \quad (12)$$

where  $C_m$  is the center of the category  $m$ . We consider the  $K1$  most similar categories as relevant categories. In the second KNN, our goal is to pick the  $K2$  most similar images from each relevant category to combine relevant images, which are visually and semantically similar to the test image. We define the multi-modal similarity between the test image ( $I$ ) and training image ( $J$ ) as follows:

$$sim_{multi-modal}(I, J) = \theta \times \cos(v_i^{KCCA}, v_j^{KCCA}) + (1 - \theta) \times \cos(t_i^{KCCA}, t_j^{KCCA}) \quad (13)$$

where  $v_i^{KCCA}$  and  $v_j^{KCCA}$  are embedding visual features of image  $I$  and image  $J$ , respectively,  $t_i^{KCCA}$  and  $t_j^{KCCA}$  are their embedding textual label features. Finally, we can obtain KNN images from each category and regard them as relevant images.

In the third KNN, our goal is to find the  $K3$  most visual similar images and assign their labels to the test image. To focus on depicting local visual features, we measure image similarity based on the original visual feature rather than embedding one, whose metric function is cosine similarity. We choose  $K3$  neighbor training images in original visual feature space for each current image  $I_i$  and rank the labels for  $I_i$  according to their probability scores of:

$$P(I_i | w) = \sum_{j=1}^{K3} sim_{vis}(I_i, I_j) \times P(w | I_j) \quad (14)$$

where  $sim_{vis}(I_i, I_j) = \cos(v_i, v_j)$ ,  $v_i$  and  $v_j$  are original feature vectors, while  $P(w | I_j)$  is a refined label feature. All labels' probability scores for the image  $I_i$ , ( $P(w | I_i)$ ), can be predicted with Equation 4. After a group of images are automatically annotated, we regularize these probability scores.

First, the probability score for each image  $I_i$  is regularized using row-normalization as follows:

$$P(w_k | I_i) = \frac{P(w_k | I_i)}{\max_k P(w_k | I_j)} \quad (15)$$

Then, probability scores of for the group images are regularized using column-normalization as follows:

$$P(w_k | I_i) = \frac{P(w_k | I_i)}{\max_j P(w_k | I_j)} \quad (16)$$

At last, the final annotations can be selected with Equation (5).

## IV. EXPERIMENT

### A. DATASETS

We conducted our experiments on three benchmark datasets including Corel5k, ESP Game, and IAPR TC-12. The images in these datasets are of various categories such as natural scene, game, sketches, personal photos and so on, which makes the annotation a challenging task.

Corel5K is the first and also the most widely used dataset for evaluating image annotation. It was first used by Duygulu et al. in 2002, and since then it has become a de facto evaluation benchmark for comparing the annotation performance [2], [34]. It consists of 4500 training images and 499 test images. Each image is either  $192 \times 128$  or  $128 \times 192$  pixels. Each image is annotated with up to 5 words (labels), with 3.5 labels on average from a dictionary of 260 labels.

ESP Game dataset was published by von Ahn and Dabbish in 2004. The dataset consists of 18689 training images and 2081 test images. Each image is manually annotated with up to 15 labels, with 4.7 labels on average from a dictionary of 268 labels. The dataset images are annotated by game player using an online game. The two mutually unknown players are required to predict the same keyword(s) to score points for a randomly given image, which makes this dataset quite challenging and diverse.

IAPR TC-12 dataset was introduced by Grubinger for cross-lingual information retrieval in 2007. Each image is initially associated with a long description. The English nouns extracted from the descriptions by Makadia [4], [12] are treated as annotations. The dataset consists of 17665 training images and 1962 test images. Each images is  $480 \times 360$  or  $360 \times 480$  pixels. Each image is manually annotated up to 23 labels, with 5.7 labels on average from a dictionary of 291 labels. The dataset has been widely used for evaluating image annotation models.

The famous large-scale datasets include NUS-WIDE [35] and Microsoft COCO(MS-COCO). There are many works on NUS-WIDE [29], [31], and all of them remove some noisy tags and images to obtain clean image dataset. However, the refined image datasets are different. Therefore, we conduct experiments only on large-scale dataset MS-COCO. The MS-COCO dataset is used for image recognition, segmentation, and captioning. It contains 123 thousand images of 170,339 user provided noisy tags and 80 expert-provided ground truth

labels. Following previous works [12], [36], we only keep 1,000 frequent tags and remove the images without any expert label, which leaves us with 123,286 images including 82,782 training images and 40,504 test images; each image being annotated with 2.9 labels on average. The refined MS-COCO dataset is the same as some research works [12], [36].

### B. EVALUATION METRICS

The per-label evaluation metrics have been widely used to evaluate image annotation approaches in the past two decades. Today, the per-label evaluation metrics have been considered as standard evaluation metrics. The per-label evaluation metrics include precision, recall, and F1-measure. For each label, per-label precision is defined as the number of images correctly predicted over the total number of images predicted with this label, and per-label recall is defined as the number of images correctly predicted over the total number of images having this label in its ground-truth or manual annotations. These values are averaged over all the labels in the vocabulary to get average (percentage) per-label precision ( $P_L$ ) and average per-label recall ( $R_L$ ) respectively. From these scores, we compute the average per-label F1-measure ( $F1_L$ ), which is the harmonic mean of  $P_L$  and  $R_L$ . The per-label precision is defined as,

$$P_L = \frac{TP}{TP + FP} \quad (17)$$

where  $TP$  is the number of images that contain the label in manual annotations and are correctly predicted the label by annotation model.  $FP$  is the number of the images that do not contain the label and are incorrectly predicted the label.  $TP + FP$  equals to the total number of images predicted the label by model. The per-label recall is defined as,

$$R_L = \frac{TP}{TP + FN} \quad (18)$$

where  $FN$  is the number of the images that contain the label in manual annotations and are not predicted the label by the model.  $TP + FN$  equals to the total number of the images containing the label in the manual annotations. F1-measure combines  $P$  with  $R$ , indicating the integrated result.

F1-measure is used for comprehensive performance evaluation by combing precision and recall. The per-label F1-measure is defined as,

$$F1_L = \frac{2 \times P_L \times R_L}{P_L + R_L} \quad (19)$$

We also consider the  $N+$  metric, which counts how many labels in the vocabulary are correctly predicted for at least one on test images.

Besides per-label metrics, more and more researchers adopt per-image metrics (also including precision, recall, and F1-measure) to evaluate annotation performance [9]–[11], [20], [30], [37], the per-label metrics are biased toward infrequent labels because making them correct could have a very significant impact on final accuracy [10]. These values are averaged over all the images in the test dataset to get

average (percentage) per-image precision ( $P_I$ ) and average per-image recall ( $R_I$ ), respectively. The per-image precision is defined as,

$$P_I = \frac{TP}{TP + FP} \quad (20)$$

where  $TP$  is the number of the labels that are contained in the image and are correctly predicted the label by annotation model.  $FP$  is the number of the labels that are not contained in the image and are incorrectly predicted the label.  $TP+FP$  equals to the total number of labels that are predicted by the model. The per-image recall is defined as,

$$R_I = \frac{TP}{TP + FN} \quad (21)$$

where  $FN$  is the number of the labels that are contained in the image and are not predicted the label by the model.  $TP + FN$  equals to the total number of the labels that is contained in the image. F1-measure combines  $P_I$  with  $R_I$ , indicating the integrated result. The per-image F1-measure ( $F1_I$ ), the harmonic mean of  $P_I$  and  $R_I$ . The per-image F1-measure is defined as,

$$F1_I = \frac{2 \times P_I \times R_I}{P_I + R_I} \quad (22)$$

The mean average precision (MAP) is a widely used metric in the field of image retrieval [11], [12], [38]. The MAP includes per-label MAP ( $MAP_L$ ) and per-image MAP ( $MAP_I$ ), which take into account all labels for every image, and evaluate the full ranking.  $MAP_L$  measures image-ranking quality corresponding to labels, while  $MAP_I$  measures label-ranking quality corresponding to images. MAP measures the full ranking of images instead of only the top labels for each image as traditional evaluation metrics [11]. Therefore,  $MAP_L$  is less noisy and preferable to other per-label metrics. Recently, more and more works use MAP as image annotation evaluation metrics [12], [39]–[41]. To more comprehensively evaluate annotation performance, we also use  $MAP_L$  and  $MAP_I$  as supplementary evaluation metrics for image annotation.

### C. IMPLEMENTATION DETAILS

For a fair comparison, visual features of all methods except MBRM are extracted from the same deep learning network architecture (VGG-16), while MBRM is performed using a handcraft feature due to the model itself. For the PLSA method, we first extract convolutional features from Conv5\_2 of VGG-16, and generate a 1000-dimension visual feature vector for each image by the k-means algorithm. For other methods, we use FC7 of VGG-16 to extract 4096-dimensional vector as a visual feature vector. The VGG-16 network used in this paper is pre-trained on the ImageNet2012 dataset [22] without retraining or fine-tuning on target datasets to demonstrate our model generality.

For nearest-neighbor based models, the number of nearest neighbors  $K$  is set to the optimum value for each model, such as JEC, TagProp, and ours setting as 15, while 2PKNN as 3.  $K$  of Equation 3 is set to 100.  $\theta$  of Equation 13 is set to 0.8.

The neighbor number  $K$  of our three-pass KNN is set to 3, 30, and 30, respectively.

### D. RESULTS AND COMPARISON

For a fair comparison, we carry out our experiments on the same three benchmark datasets (Corel5k, ESP Game and IAPR TC-12) and predict a fixed length of annotations (five labels) for each test image. We compare our method and some representative methods using per-label metrics (precision, recall, F1-measure), per-image metrics, and MAP. Furthermore, we use the hybrid F1-measure (called H-F1) combining per-label F1-measure and per-image F1-measure with the harmonic mean [9]. We compare our method with state-of-the-art models, including classical probabilistic model MBRM, classical topic model PLSA-WORDS, classical CCA model CCA-KNN, two nearest-neighbor models JEC and TagProp, classical discriminative model SVM-DMBRM [14], and the state-of-the-art nearest neighbour based model 2PKNN. We also compare with GAN based D<sup>2</sup>IA annotation method. Only part of metrics of CCA-KNN, SVM-DMBRM and D<sup>2</sup>IA are compared in the following, whose performances are quoted from [7], [14], [24].

The experiment results on Corel5k, ESP Game, and IAPR TC-12 are summarized in Tables 1, 2, and 3, respectively. From Tables 1-3, we can see that our proposed method significantly outperforms all methods but D<sup>2</sup>IA on three benchmark datasets in terms of almost all metrics. Our performance improvement largely benefits from label refinement, multi-level semantic neighborhood.

To further evaluate the annotation performance, we varied the number of annotation labels from 2 to 20 and compared our method with competitive methods. Both per-label and per-image precision-recall curves of MBRM, JEC, TagProp, PLSA-WORDS, 2PKNN, and our method are visualized in Figure 2 based on three benchmark datasets. Both per-image and per-label precision/recall values are the mean values calculated over all the test images and all the labels, respectively. As can be seen from Figure 2, our model remarkably outperforms the others for almost any number of annotation labels. These again confirm the effectiveness of our method.

To compare with deep learning based image annotation models on large-scale datasets, we also carry out our experiments on MS-COCO dataset and predict a fixed length of annotations (three labels) for each test image. We compare our method with state-of-the-art models, including CNN+WARP, CNN-RNN, and MangoNet. We annotate images based on multi-modal (deep visual features and textual tag features) embedding features mapped by KCCA. The experiment results on MS-COCO is summarized in Tables 4. As can be seen from Table 4, our method significantly outperforms the other methods (non-deep as well as deep learning based methods) on large-scale datasets in terms of most evaluation metrics, which mainly benefits from high-level semantic features and accurate neighbors. MangoNet outperforms our method in terms of precision or recall metrics, which might largely benefit from the co-attention and GCN model, as it



TABLE 1. Performance evaluation on Corel5k dataset

Model	Visual	P <sub>L</sub>	R <sub>L</sub>	F1 <sub>L</sub>	N+	P <sub>I</sub>	R <sub>I</sub>	F1 <sub>I</sub>	H-F1	MAP <sub>L</sub>	MAP <sub>I</sub>
MBRM [5]	HC	24	25	24.90	122	32	45	37.10	29.50	26.31	44.23
JEC [6]	CNN	30	33	31.38	137	41	58	48.12	38.07	35.32	53.69
TagProp [12]	CNN	31	40	34.97	149	44	61	50.96	41.48	38.01	59.67
PLSA-WORDS [16]	CNN	20	30	23.87	129	35	50	41.07	30.20	27.43	46.03
2PKNN [2]	CNN	38	46	41.67	179	38	54	44.85	43.20	46.73	48.95
CCA-KNN [24]	CNN	39	51	44.20	192	-	-	-	-	-	-
SVM-DMBRM [14]	CNN	42	45	43.45	186	-	-	-	-	-	-
Ours	CNN	<b>44</b>	<b>55</b>	<b>49.18</b>	<b>198</b>	<b>47</b>	<b>67</b>	<b>55.21</b>	<b>52.02</b>	<b>50.02</b>	<b>64.95</b>

TABLE 2. Performance evaluation on ESP Game dataset

Model	Visual	P <sub>L</sub>	R <sub>L</sub>	F1 <sub>L</sub>	N+	P <sub>I</sub>	R <sub>I</sub>	F1 <sub>I</sub>	H-F1	MAP <sub>L</sub>	MAP <sub>I</sub>
MBRM [5]	HC	18	19	18.80	209	25	28	26.83	21.89	20.34	28.11
JEC [6]	CNN	32	23	26.95	228	35	39	36.87	31.14	21.05	39.99
TagProp [12]	CNN	36	28	31.61	234	38	42	39.87	35.26	25.82	40.73
PLSA-WORDS [16]	CNN	20	24	21.78	201	25	27	25.94	22.68	19.00	27.93
2PKNN [2]	CNN	33	34	33.45	255	35	39	37.03	35.15	30.70	40.99
D <sup>2</sup> IA [7]	GAN	31	<b>49</b>	38.10	-	-	-	-	-	-	-
CCA-KNN [24]	CNN	44	32	37.05	254	-	-	-	-	-	-
SVM-DMBRM [14]	CNN	<b>51</b>	26	35.44	251	-	-	-	-	-	-
Ours	CNN	49	35	<b>40.89</b>	<b>259</b>	<b>45</b>	<b>50</b>	<b>47.32</b>	<b>43.87</b>	<b>37.92</b>	<b>53.35</b>

TABLE 3. Performance evaluation on IAPR TC-12 dataset

Model	Visual	P <sub>L</sub>	R <sub>L</sub>	F1 <sub>L</sub>	N+	P <sub>I</sub>	R <sub>I</sub>	F1 <sub>I</sub>	H-F1	MAP <sub>L</sub>	MAP <sub>I</sub>
MBRM [5]	HC	24	23	23.54	223	29	27	28.06	25.53	25.27	28.35
JEC [6]	CNN	34	22	26.30	218	43	41	41.88	32.31	26.26	46.75
TagProp [12]	CNN	43	35	38.83	257	48	46	47.05	42.55	39.76	53.40
PLSA-WORDS [16]	CNN	23	25	23.72	207	33	32	32.19	27.31	21.99	33.47
2PKNN [2]	CNN	49	32	38.74	274	42	41	41.92	40.26	39.09	47.78
D <sup>2</sup> IA [7]	GAN	33	<b>45</b>	37.73	-	-	-	-	-	-	-
CCA-KNN [24]	CNN	41	34	37.17	273	-	-	-	-	-	-
SVM-DMBRM [14]	CNN	<b>58</b>	27	36.84	268	-	-	-	-	-	-
Ours	CNN	51	37	<b>42.79</b>	<b>278</b>	<b>52</b>	<b>50</b>	<b>50.78</b>	<b>46.44</b>	<b>41.88</b>	<b>58.20</b>

TABLE 4. Performance evaluation on MS-COCO dataset












Model	Visual	P <sub>L</sub>	R <sub>L</sub>	F1 <sub>L</sub>	P <sub>I</sub>	R <sub>I</sub>	F1 <sub>I</sub>	H-F1	MAP <sub>L</sub>	MAP <sub>I</sub>
JEC [6]	CNN	49.12	41.35	44.90	49.39	61.10	54.63	49.29	41.63	67.34
TagProp [12]	CNN	56.43	56.15	56.29	56.78	69.59	62.54	59.25	63.14	77.67
2PKNN [2]	CNN	62.21	45.81	52.76	51.27	61.88	56.08	54.37	55.66	69.20
CNN+WARP [10]	CNN	57.09	55.31	56.19	57.54	70.03	63.18	59.48	58.11	78.93
CNN-RNN [29]	CNN	66.00	55.60	60.40	69.20	66.40	67.80	63.89	-	-
MangoNet [31]	GCN+Co-attention	<b>87.10</b>	57.90	67.60	<b>89.50</b>	61.90	73.20	70.29	<b>77.50</b>	84.30
Ours	CNN	76.11	<b>70.41</b>	<b>73.15</b>	68.35	<b>82.01</b>	<b>74.56</b>	<b>71.81</b>	73.37	<b>86.54</b>

can capture high-quality visual features and accurately model the correlations between each target image and its neighbors by metadata neighborhood graph. As far as the most important metric per-label F1 and comprehensive metric H-F1 are concerned, our method generally outperforms MangoNet.

The main reason of our proposed model performance improvement can be summarized as follows 1) We propose label refinement to alleviate the weak-labeling. 2) We address the issues of semantic gap and different levels of abstraction by our proposed multi-level semantic neighborhoods.

3) Our method outperforms most methods in terms of per-label metrics by a large margin, which mainly contributes to our addressing the issue of label-imbalance. In contrast to the traditional NN models that pay more attention to frequent labels, our method pays more attention to the same category and relevant images rather than rare labels, which gives equal importance to all labels of the relevant images. As a consequence, our method can improve the annotation performance of infrequent labels without sacrificing frequent labels, thus improving performance in both per-label and per-

TABLE 5. Example images and annotation predicted by various methods

Test Image	Manual Label	JEC	2PKNN	Ours
	tree, flowers, garden, tulip	flowers, <i>house</i> , garden, <i>window</i> , tree	<i>blooms</i> , flowers, garden, <i>house</i> , <i>window</i>	flowers, tulip, garden, tree, <i>landscape</i>
	water, hills, coast, lighthouse	water, <i>sky</i> , coast, <i>island</i> , <i>mountain</i>	lighthouse, coast, <i>island</i> , water, <i>rocks</i>	coast, lighthouse, <i>waves</i> , hills, water
	forest, cat, tiger, bengal	cat, tiger, forest, bengal, <i>water</i>	tiger, cat, forest, bengal, <i>rocks</i>	cat, forest, tiger, bengal, <i>head</i>
	couple, man, people, red, shirt, woman	<i>white</i> , <i>blue</i> , woman, <i>sky</i> , <i>yellow</i>	<i>beard</i> , <i>group</i> , <i>glasses</i> , <i>arrow</i> , man	man, people, shirt, <i>smile</i> , woman
	brown, horse, man	man, horse, <i>black</i> , woman, <i>white</i>	horse, <i>grass</i> , <i>hat</i> , man, <i>animal</i>	horse, <i>grass</i> , man, brown, <i>tail</i>
	man, show, tie, tv	man, <i>picture</i> , tie, tv, <i>hair</i>	tie, tv, man, <i>picture</i> , show	tie, tv, show, man, <i>suit</i>
	fruit, man, table	<i>tree</i> , <i>people</i> , <i>house</i> , table, <i>lawn</i>	table, <i>couch</i> , <i>restaurant</i> , <i>house</i> , cup	<i>cloth</i> , table, <i>chair</i> , fruit, man
	dog, grass, horse, landscape, mountain, people	<i>man</i> , <i>front</i> , <i>wall</i> , woman, <i>clothes</i>	house, dog, <i>tourist</i> , <i>front</i> , <i>man</i>	<i>cape</i> , horse, <i>meadow</i> , people, dog
	bottle, hand, man	man, table, <i>glass</i> , bottle, woman	<i>glass</i> , hand, man, bottle, <i>tee-shirt</i>	<i>couple</i> , bottle, man, <i>glass</i> , hand
	city, cloud, green, sky, tree	<i>white</i> , green, sky, tree, cloud	cloud, sky, green, tree, <i>white</i>	sky, cloud, <i>mountain</i> , <i>hill</i> , <i>lake</i>
	floor, front, jersey, team, wall	front, wall, jersey, floor, team	jersey, team, floor, front, wall	floor, team, jersey, front, <i>skirt</i>

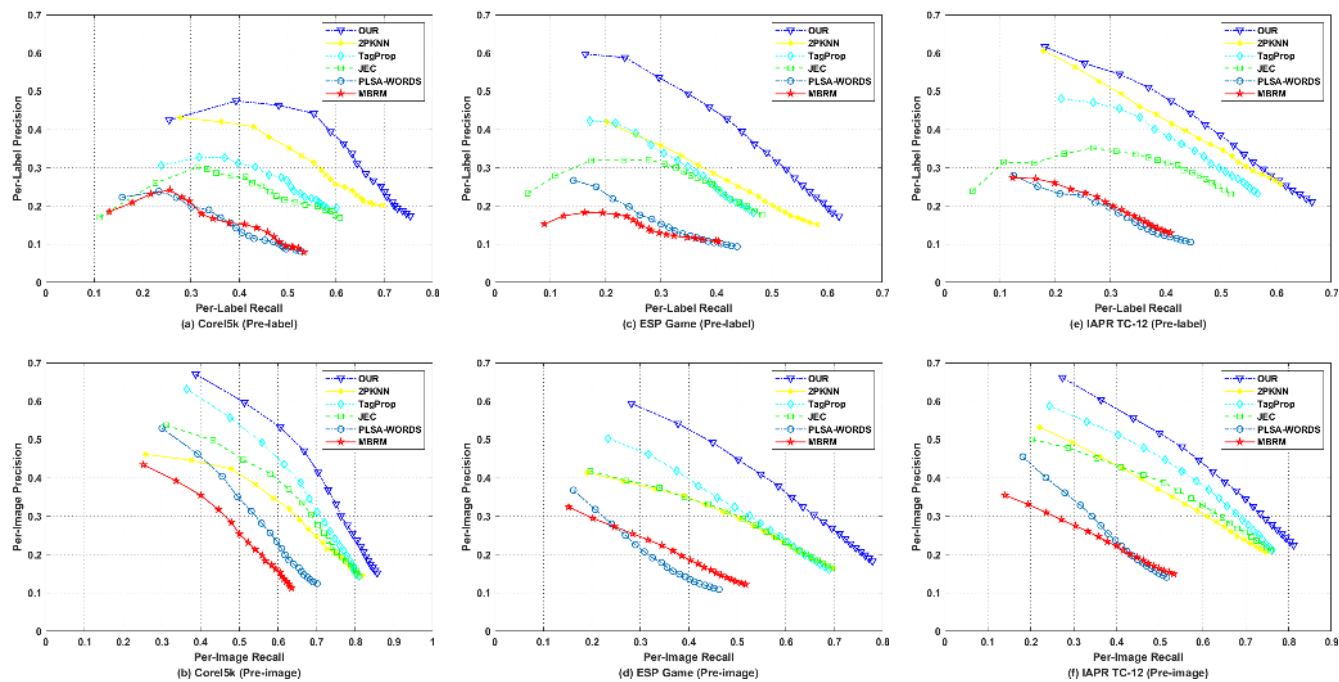


FIGURE 2. Precision-Recall curves on three datasets

image metrics.

### E. QUALITATIVE ANALYSIS

Table 5 shows several examples of annotations produced by JEC, 2PKNN and our method on the three datasets. The example images in the first three rows are from Core5k, the second three rows from ESP Game, the third three rows from IAPR TC-12, and the last two images are from ESP Game and IAPR TC-12. As for most images, we can see that our method can correctly predicate the ground-truth annotations, although there are some extra labels. By checking the extra labels (with blue font), we find that most of them are all consistent with the content of the images but not included in ground-truth labels. Our method can consider category-level, semantic information, and visual information in different steps so as to find visually and semantically similar images and predicate the correct annotation labels.

As for the tenth image, any method has not correctly predicted the ground-truth annotations. Both JEC and 2PKNN improperly predict “white”, while our proposed method improperly predict “mountain”, “hill” and “lake”. As for JEC and 2PKNN, the relevance between the test image and training images completely depend on their visual similarities. Hence, they can predict “white” according to the visual feature of the sky. As for our proposed method, it first identify the image as scene category, and predict “mountain”, “hill”, and “lake”. As for the eleventh image, only our method improperly predict “skirt” rather than “wall”, while the others correctly predicated the ground-truth annotations. This is possibly because our method tends to the foreground object rather than the background.

### F. EFFICIENCY ANALYSIS

To verify the efficiency of the proposed model, as shown in Table 6, we compare the time costs among MBRM, JEC, TagProp, 2PKNN, PLSA, and our model. The experiments are mainly performed using Matlab on a computer of Intel Corei7-9750H CPU with 2.6GHz and 16 GB RAM, running Windows 10 OS, but some components of Tagprop is based on C language. As shown in Table 6, the time costs of two models (TagProp and PLSA) can be separated the training stage and the testing stage, while those of JEC, LL-PLSA, and 2PKNN can not be separated.

As shown in Table 6, our proposed model can dramatically reduce time cost in contrast to any other model. The time costs of annotating all test images on Core5k by our model is 0.57 seconds. The time costs of annotating all test images on Core5k by MBRM, JEC, and 2PKNN are 34.28, 5.09 and 9.57 seconds, respectively. The time costs of training model from training images on Core5k for TagProp and PLSA are 43.74 and 17.61 seconds, respectively, while the time costs of annotating all test images are 0.49 and 4.34 seconds, respectively. The experiments on ESP Game and IAPR TC-12 show similar results.

In contrast to all nearest-neighbor models, the time overhead of our model is proportional to the number and the size of categories rather than the size of the entire training image database. Mostly, the number of categories is small and constant; therefore, our time cost is much smaller than others. All in all, compared with other nearest-neighbor models, our proposed model is more fit to real-world online image repository or large-scale social image database.

TABLE 6. Time costs of various models (in seconds)

Model	Corel5k	ESP Game	IAPR TC-12
MBRM [5]	34.28	512.21	678.72
JEC [6]	5.09	9.97	9.56
TagProp [12]	43.74+0.49	161.41+2.01	164.63+2.13
2PKNN [2]	9.57	113.81	111.01
PLSA [16]	17.61+4.34	66.73+26.62	205.92+23.59
Ours	0.57	4.89	4.41

## V. CONCLUSION AND FUTURE WORK

We present a novel image annotation based on multi-level semantic neighborhood. Our proposed method has several advantages. 1) To our knowledge, this is the first published work that proposes a pre-annotation strategy to determine the test image's category for promoting semantic level. 2) Our proposed method refines labels before image annotation for alleviating the issue of weak-labeling. 3) Our proposed method is based on multi-level semantic neighborhoods, which can provide rich representations at different abstraction scales. As a consequence, this model is suitable for image annotation task because it can address the issue of wider range labels. 4) Our proposed method is a three-pass variant of the traditional KNN, with each pass using a different feature vector. Our method can find visually and semantically similar neighbor images, which can reduce the semantic gap and improve the performance. 5) In contrast to the traditional NN models paying more attention to frequent labels and classical 2PKNN paying more attention to rare labels, our method can improve performance in both per-label and per-image metrics.

Extensive experiments demonstrate that our method can achieve significantly outperforms competitive methods in terms of almost all evaluation metrics. Even though nearest neighbors based annotation models are concept-clear, structure-intuitive, and effective, there are several shortcomings. First, these methods will be time-consuming and space-consuming if the number of the training image dataset is huge. Second, the performance of nearest neighbor model-based AIA methods may be influenced by the size of training datasets.

In the future, we will explore a new modeling strategy based on the existing model combining the merits of discriminative and generative models so as to further reduce modeling complexity. In addition, we are interested in exploring the new technology in attention models into feature extraction and Graph Neural Network into representation learning of multi-modal information.

## REFERENCES

- [1] J. Wang, A. Gilbert, B. Thomee, and M. Villegas, "Automatic image annotation at imageclef," in *Information Retrieval Evaluation in a Changing World*. Springer, 2019, pp. 251–273.
- [2] Y. Verma and C. Jawahar, "Image annotation by propagating labels from semantic neighbourhoods," *International Journal of Computer Vision*, vol. 121, no. 1, pp. 126–148, 2017.
- [3] P. Bhagat and P. Choudhary, "Image annotation: Then and now," *Image and Vision Computing*, vol. 80, pp. 1–23, 2018.
- [4] Y. Ma, Y. Liu, Q. Xie, and L. Li, "Cnn-feature based automatic image annotation method," *Multimedia Tools and Applications*, vol. 78, no. 3, pp. 3767–3780, 2019.
- [5] S. Feng, R. Manmatha, and V. Lavrenko, "Multiple bernoulli relevance models for image and video annotation," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, vol. 2. IEEE, 2004, pp. II–II.
- [6] A. Makadia, V. Pavlovic, and S. Kumar, "Baselines for image annotation," *International Journal of Computer Vision*, vol. 90, no. 1, pp. 88–105, 2010.
- [7] B. Wu, W. Chen, P. Sun, W. Liu, B. Ghanem, and S. Lyu, "Tagging like humans: Diverse and distinct image annotation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7967–7975.
- [8] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," *Machine learning*, vol. 85, no. 3, p. 333, 2011.
- [9] Y. Niu, Z. Lu, J.-R. Wen, T. Xiang, and S.-F. Chang, "Multi-modal multi-scale deep learning for large-scale image annotation," *IEEE Transactions on Image Processing*, vol. 28, no. 4, pp. 1720–1731, 2018.
- [10] Y. Gong, Y. Jia, T. Leung, A. Toshev, and S. Ioffe, "Deep convolutional ranking for multilabel image annotation," *arXiv preprint arXiv:1312.4894*, 2013.
- [11] J. Johnson, L. Ballan, and L. Fei-Fei, "Love thy neighbors: Image annotation by exploiting image metadata," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4624–4632.
- [12] A. Dutta, Y. Verma, and C. Jawahar, "Automatic image annotation: the quirks and what works," *Multimedia Tools and Applications*, vol. 77, no. 24, pp. 31 991–32 011, 2018.
- [13] T. Tesan, P. Coscia, and L. Ballan, "A cnn-rnn framework for image annotation from visual cues and social network metadata," *arXiv preprint arXiv:1910.05770*, 2019.
- [14] V. N. Murthy, E. F. Can, and R. Manmatha, "A hybrid model for automatic image annotation," in *Proceedings of International Conference on Multimedia Retrieval*, 2014, pp. 369–376.
- [15] Y. Ning, S. Hai-yu, S. Dong-yang, W. Peng-jie, and Y. Jin-xin, "Image annotation based on middle-layer convolution features of deep learning," *Journal of Graphics*, vol. 40, no. 5, p. 872, 2019.
- [16] T. Uricchio, L. Ballan, L. Seidenari, and A. Del Bimbo, "Automatic image annotation via label transfer in the semantic space," *Pattern Recognition*, vol. 71, pp. 144–157, 2017.
- [17] Q. Cheng, Q. Zhang, P. Fu, C. Tu, and S. Li, "A survey and analysis on automatic image annotation," *Pattern Recognition*, vol. 79, pp. 242–259, 2018.
- [18] G. Carneiro, A. B. Chan, P. J. Moreno, and N. Vasconcelos, "Supervised learning of semantic classes for image annotation and retrieval," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 3, pp. 394–410, 2007.
- [19] C. Xu, Y. Dai, R. Lin, and S. Wang, "Social image refinement and annotation via weakly-supervised variational auto-encoder," *Knowledge-Based Systems*, vol. 192, p. 105259, 2020.
- [20] H. Song, P. Wang, J. Yun, W. Li, B. Xue, and G. Wu, "A weighted topic model learned from local semantic space for automatic image annotation," *IEEE Access*, vol. 8, pp. 76 411–76 422, 2020.
- [21] M. Koskela and J. Laaksonen, "Convolutional network features for scene recognition," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 1169–1172.
- [22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [23] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," in *International conference on machine learning*. PMLR, 2014, pp. 647–655.

- [24] V. N. Murthy, S. Maji, and R. Manmatha, "Automatic image annotation using deep learning representations," in *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, 2015, pp. 603–606.
- [25] M. Zang, D. Wen, K. Wang, T. Liu, and W. Song, "A novel topic feature for image scene classification," *Neurocomputing*, vol. 148, pp. 467–476, 2015.
- [26] X. Shu, G.-J. Qi, J. Tang, and J. Wang, "Weakly-shared deep transfer networks for heterogeneous-domain knowledge propagation," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 35–44.
- [27] J. Tang, X. Shu, Z. Li, G.-J. Qi, and J. Wang, "Generalized deep transfer networks for knowledge propagation in heterogeneous domains," 2016.
- [28] W. Wang, X. Yan, H. Lee, and K. Livescu, "Deep variational canonical correlation analysis," *arXiv preprint arXiv:1610.03454*, 2016.
- [29] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, "Cnn-rnn: A unified framework for multi-label image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2285–2294.
- [30] F. Liu, T. Xiang, T. M. Hospedales, W. Yang, and C. Sun, "Semantic regularisation for recurrent image annotation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2872–2880.
- [31] J. Zhang, Q. Wu, J. Zhang, C. Shen, and J. Lu, "Mind your neighbours: Image annotation with metadata neighbourhood graph co-attention networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2956–2964.
- [32] J. Tang, X. Shu, G.-J. Qi, Z. Li, M. Wang, S. Yan, and R. Jain, "Tri-clustered tensor completion for social-aware image tag refinement," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 8, pp. 1662–1674, 2016.
- [33] J. Tang, X. Shu, Z. Li, Y.-G. Jiang, and Q. Tian, "Social anchor-unit graph regularized tensor completion for large-scale image retagging," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 8, pp. 2027–2034, 2019.
- [34] P. Duygulu, K. Barnard, J. F. de Freitas, and D. A. Forsyth, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," in *European conference on computer vision*. Springer, 2002, pp. 97–112.
- [35] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "Nus-wide: a real-world web image database from national university of singapore," in *Proceedings of the ACM international conference on image and video retrieval*, 2009, pp. 1–9.
- [36] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: Lessons learned from the 2015 mscoco image captioning challenge," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 4, pp. 652–663, 2016.
- [37] D. Putthividhy, H. T. Attias, and S. S. Nagarajan, "Topic regression multimodal latent dirichlet allocation for image annotation," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 3408–3415.
- [38] X. Li, T. Uricchio, L. Ballan, M. Bertini, C. G. Snoek, and A. D. Bimbo, "Socializing the semantic gap: A comparative survey on image tag assignment, refinement, and retrieval," *ACM Computing Surveys (CSUR)*, vol. 49, no. 1, pp. 1–39, 2016.
- [39] Z. Li, J. Tang, and T. Mei, "Deep collaborative embedding for social image understanding," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 9, pp. 2070–2083, 2018.
- [40] N. A. Tu, K. U. Khan, and Y.-K. Lee, "Featured correspondence topic model for semantic search on social image collections," *Expert Systems With Applications*, vol. 77, pp. 20–33, 2017.
- [41] H. Song, J. Yun, H. Li, M. Zheng, J. Yao, H. Lv, and A. Fang, "An efficient and effective model based on mean positive examples for social image annotation," *IEEE Access*, vol. 8, pp. 210 695–210 708, 2020.



HOUJIE LI Received the B.S. and M.S. degrees in communication and information system from Jilin University, in 2001, 2004, and the Ph.D. degree in signal and information processing from Dalian University of Technology, in 2019. He is currently an associate professor with the School of Information and Communication Engineering, Dalian Minzu University, China. His current research interests include Image Processing, computer vision, and machine learning.



WEI LI Received the B.S. and M.S. degrees in computer software and theory from Jilin University, in 2002, 2005, and the Ph.D. degree in traffic information engineering and control from Jilin University, in 2021. She is currently a lecturer with the School of Computer Science and Engineering, Dalian Minzu University, China. Her current research interests include image understanding, computer vision, and machine learning.



HONGDA ZHANG is currently pursuing the M.S. degree with the School of Information and Communication Engineering, Dalian Minzu University, China. His current research interests include computer vision, image processing, and machine learning.



XIN HE is currently pursuing the M.S. degree with the School of Information and Communication Engineering, Dalian Minzu University, China. Her current research interests include computer vision, image processing, and machine learning.



MINGXIAO ZHENG is currently pursuing the B.S. degree with the School of Computer Science and Engineering, Dalian Minzu University, China. His current research interests include computer vision, image processing, and deep learning.



HAIYU SONG Received the B.S.,M.S.,and Ph.D. degrees in computer software and theory from Jilin University, in 1996, 2003, and 2012, respectively. He is currently a professor with the School of Computer Science and Engineering, Dalian Minzu University, China. His current research interests include image understanding, computer vision, and machine learning.

...