# Multi-View Image Classification With Visual, Semantic and View Consistency

Chunjie Zhang, Jian Cheng, and Qi Tian, *Fellow, IEEE*

***Abstract*— Multi-view visual classification methods have been widely applied to use discriminative information of different views. This strategy has been proven very effective by many researchers. On the one hand, images are often treated independently without fully considering their visual and semantic correlations. On the other hand, view consistency is often ignored. To solve these problems, in this paper, we propose a novel multi-view image classification method with visual, semantic and view consistency (VSVC). For each image, we linearly combine multi-view information for image classification. The combination parameters are determined by considering both the classification loss and the visual, semantic and view consistency. Visual consistency is imposed by ensuring that visually similar images of the same view are predicted to have similar values. For semantic consistency, we impose the locality constraint that nearby images should be predicted to have the same class by multi-view combination. View consistency is also used to ensure that similar images have consistent multi-view combination parameters. An alternative optimization strategy is used to learn the combination parameters. To evaluate the effectiveness of VSVC, we perform image classification experiments on several public datasets. The experimental results on these datasets show the effectiveness of the proposed VSVC method.**

***Index Terms*— Multi-view learning, image classification, visual consistency, semantic consistency, view consistency.**

## I. Introduction

IMAGE classification has been widely explored in recent years. It aims to accurately classify an image based on its visual content. Multi-view-based image classification by jointly exploring the discriminative information of different views has been proven very effective [1]–[8].

Manuscript received June 22, 2018; revised December 7, 2018, January 29, 2019, March 13, 2019, and May 12, 2019; accepted August 5, 2019. Date of publication August 16, 2019; date of current version September 25, 2019. This work was supported in part by the Fundamental Research Funds for the Central Universities under Grant 2019RC040, in part by the National Science Foundation of China (NSFC) under Grant 61872362, and in part by the Beijing Municipal Science and Technology Commission under Grant Z181100008918012. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yun Fu. *(Corresponding author: Chunjie Zhang.)*

C. Zhang is with the Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China (e-mail: cjzhang@bjtu.edu.cn).

J. Cheng is with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 250014, China, with the University of Chinese Academy of Sciences, Beijing 100049, China, and also with the Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Beijing 250014, China (e-mail: jcheng@nlpr.ia.ac.cn).

Q. Tian is with the Department of Computer Sciences, The University of Texas at San Antonio, TX 78249 USA (e-mail: qitian@cs.utsa.edu).

Digital Object Identifier 10.1109/TIP.2019.2934576

One problem with multi-view-based classification methods is that different images [9]–[11] are often treated equally and independently without considering the internal relationships of images. Some images are relatively more difficult to classify than are others, even if multi-view information is considered [12]. To alleviate this problem, the use of exemplar classifier [13]–[16] and sub-classes of images [17]–[20] has become popular. These methods divide images of the same class into several sub-classesfor classifier training. However, images are still treated independently.

Locality information has been proven very effective for modelling the correlations of images [21]–[28]. Many methods assume that nearby images or features should be predicted to have similar values. Although this approach is very effective, this information is often ignored when modelling multi-view correlations of images. This is because images have varied neighbours in different views. We believe that this information should also be used for multi-view classification.

Additionally, due to visual polysemy, only using visual similarity is not enough for reliable image classification. Applications of semantic correlations have also been widely studied [10], [14], [29]–[37]. Semantic information can be obtained manually by humans or mined from the Internet. Although this strategy is effective, it is labour-intensive and is easily contaminated by noise. Furthermore, domain variance also hinders performance. Instead, it is more plausible to use the training images. For each image, we view the joint predictions of multi-views as its semantics. Similar images should have similar semantics when learning multi-view combination. The information obtained from multiple views is also combined to boost classification performance [2], [38]–[43]. This strategy tries to automatically learn optimal multi-view combination parameters. One problem with this strategy is that images are often treated individually, leaving the view consistency information unconsidered. Similar images should also be combined with consistent views.

In this paper, we propose a novel multi-view image classification method using visual, semantic and view consistency (VSVC). For each image, we linearly combine the outputs of multiple views for classification. The combination parameters are determined by optimizing over the classification loss along with visual, semantic and view consistency. For visual consistency, we use visual similarity to guide the smoothness of predicted values. The semantic consistency is achieved by ensuring that nearby images should be predicted to have the same semantics by multi-view combination.

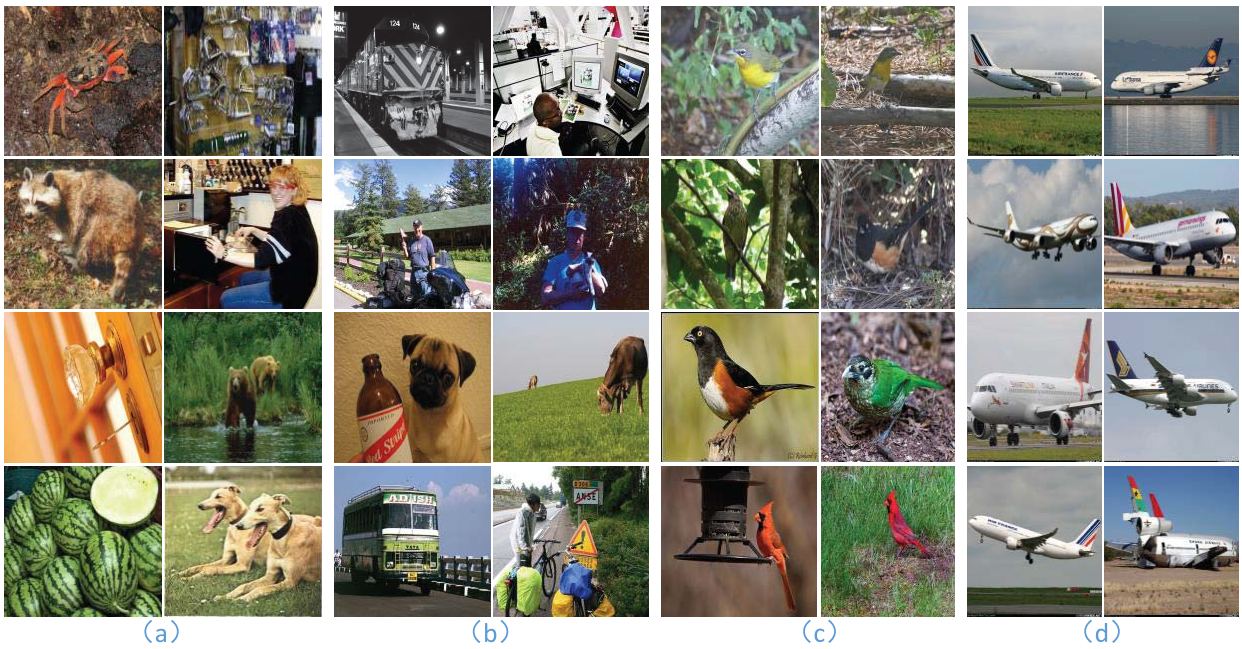


Fig. 1. Example images of the (a) Caltech-256 dataset, (b) PASCAL VOC 2012 dataset, (c) CUB-200-2011 dataset, and (d) FGVC-aircraft dataset.

The view consistency is imposed as the smoothness of combination parameters of different views. We conduct image classification experiments on several public datasets, and the experimental results demonstrate the effectiveness of the proposed VSVC method. Figure 1 shows the flowchart of the proposed VSVC method.

The novelty of this paper is twofold:

- First, instead of only using visual consistency, we jointly use visual consistency in a single view and semantic consistency and view consistency in multi-view for multi-view classification.
- Second, the proposed method can use state-of-the-art pre-learned classifiers as multi-views to improve image classification performance.

The proposed method differs from traditional manifold regularization-based methods in two aspects. On the one hand, instead of only using visual consistency, we also impose semantic consistency over multi-views. On the other hand, VSVC uses multi-view correlation by using view consistency and jointly learning visual, semantic and view correlations. The proposed VSVC method is more general and universal and can be combined with various pre-learned classifiers.

The rest of this paper is organized as follows. Related studies are described in Section II. The details of the proposed VSVC method are illustrated in Section III. Experimental results and analysis are given in Section IV. Finally, Section V concludes.

## II. Related Work

Information extracted from multiple views has been widely used for various applications. Shi *et al.* [1] targeted the multimedia problem with multi-view sparse feature selection. Zhang *et al.* [2] shared labels among different views for image classification. Zhang and Zheng [3] used multi-view hashing to search images in a semi-supervised way. Tao *et al.* [4] used adaptive regression for multi-view semi-supervised classification. Wang *et al.* [5] used a structured low-rank constraint for spectral clustering with multi-view information. Shen *et al.* [6] searched cross-view information for label prediction. Wu *et al.* [7] targeted the graph classification problem by multiple-structure learning, while Peng *et al.* [8] used multi-view boosting with information propagation. The use of information from multiple views could improve classification performance.

Although multi-view-based methods have been proven very effective, they have ignored various difficult aspects of images. It would be more effective to consider each image individually or similar images of the same class jointly.Zhang *et al.* [12] used the correlations of exemplar classifiers for semantic modelling of images. Li *et al.* [13] used low-rank exemplar SVM classifiers for domain adaptation, while Zhang *et al.* [14] used image-level information. Hui and Sankaranarayanan [15] used virtual exemplars for reflectance estimation, while Zhu *et al.* [16] explored semantic features for image and video stylization. The use of low-rank correlations has also been widely explored [17]–[19]. Cai *et al.* [20] used multi-view information for heterogeneous image feature combination.

Manifold smoothness has often been assumed by ensuring that visually similar images have similar semantics [21]–[28]. Wang *et al.* [21] tried to learn a marginalized denoising dictionary using a locality constraint. Zhang *et al.* [22] used the search results for image classification with neighbour similarity propagation. Kwitt *et al.* [23] classified scene images on the semantic similarity graph. Gao *et al.* [25] imposed a locality constraint in the sparse coding process, while Li and Fu [26] used a low-rank constraint to learn robust subspaces. Ding *et al.* [27] transferred missing information via a low-rank constraint. Boiman *et al.* [28] studied the nearest neighbour of local features for image classification without training classifiers. However, these methods often ignored the correlations of

images with different views. The single view and multi-view as well as view combination information should be jointly considered to further improve classification performance.

Semantic-based methods have also been widely used [10], [14], [29]–[37]. Karpathy and Li [29] generated image descriptions by visual and semantic alignment. Zhang *et al.* [10] used the contextual relationships of the exemplar classifier for discriminative image representation and classification. Instead of using implicitly generated semantics, Farhadi *et al.* [30] used human-defined attributes for object representation. Zhang *et al.* [14] used the hierarchical structure information to classify each image using both visual and semantic similarities. Li *et al.* [31] used Internet images to generate an object bank. However, these images were often contaminated with noisy information. Zhang *et al.* [32] used scale and class consistency to generate codebooks. Xu *et al.* [33] tried to adapt information from other domains for fine-grained classification using the information from the Internet. Zhang *et al.* [34] jointly considered the object, contextual and background information for classification. To cope with the shortage of labelled images, Zhang *et al.* [36] used unlabelled images by ensuring prediction consistency. Tang *et al.* [37] proposed multi-view-based support vector machines.

Semantic correlation has been very useful for classification. However, for multi-view image classification, the view information is also very important and has often been ignored by researchers [38]–[43]. Nie *et al.* [38] combined multi-view physician attributes for health analysis. Li *et al.* [39] used low-rank embedding to combine multi-view correlations. Ma *et al.* [40] explored the navigation problem with multi-views. Chang *et al.* [41] combined semantic representation with bi-level representation, while Yu *et al.* [42] ranked images with deep multimodal distance learning. Takahashi *et al.* [43] used audio information to analyse videos.

Many methods have also been proposed to improve classification performance with various classifiers [44]–[82], e.g., AlexNet [44], VGG [45], GoogleNet [46], and ResNet [47]. Zhang *et al.* [52] generated many codebooks with low-rank sparse coding. Wang *et al.* [55] used the locality constraint for sparse coding. Chatfield *et al.* [56] re-implemented many methods to evaluate their details. Wei *et al.* [57] explored multi-label image classification with a CNN. Jointly combining multiple data could eventually improve performance. A bilinear convolutional neural network was proposed by Lin *et al.* [61] to use the spatial layouts of images. Zhang *et al.* [62] encoded the spatial information, while Cui *et al.* [63] used human-labelled data. Jaderberg *et al.* [64] proposed the spatial transformer networks, while Moghimi *et al.* [65] boosted a number of networks. The combination of deep networks has also been explored [67]–[70]. Researchers have also used location information of objects for classification [72], [74] with the sparsity constraint [71]. Meyer *et al.* [73] combined neighbour information, while Wang *et al.* [75] combined multi-view clues in an unsupervised way. Researchers also proposed many view combination strategies for various visual applications [76]–[84].

TABLE I
SYMBOLS AND THEIR CORRESPONDING DESCRIPTIONS USED IN THIS PAPER

| Symbol | Description |
|---|---|
| $V$ | view number |
| $N$ | number of training images |
| $\boldsymbol{x}_n^v$ | representation of $n$-th image in $v$-th view |
| $y_n$ | label of the $n$-th image |
| $f^v(*)$ | outputs of the $n$-th view classifier |
| $\alpha_n^v$ | linear combination parameter for $n$-th image and $v$-th view |
| $\boldsymbol{\alpha}_n$ | combination parameter for $n$-th image |
| $\boldsymbol{\alpha}$ | combination parameter matrix |
| $\boldsymbol{F}(\boldsymbol{x_n})$ | vector form of predicted value of multi-views |
| $\widehat{y}_n$ | predicted values for the $n$-th image |
| $\boldsymbol{\alpha}_n^T$ | transpose of $\boldsymbol{\alpha}_n$ |
| $\ell(*,*)$ | classification loss |
| $\|*\|$ | Euclidean distance |
| $\beta_1$ | weighting parameter for single view |
| $\sigma$ | scaling parameter of locality information |
| $M_v$ | number of nearby images for the $v$-th view |
| $\beta_2$ | weighting parameter for multi-view |
| $\beta_3$ | weighting parameter for view combination |
| $\beta_4$ | regularization parameter for $\alpha_n^v$ |
| $M_{it}$ | maximum iteration number |
| $\widetilde{\boldsymbol{x}}_t^v$ | $v$-th view representation of testing image |
| $\widetilde{M}_v$ | number of nearby images of $\widetilde{\boldsymbol{x}}^v$ |
| $\widetilde{\boldsymbol{\alpha}}_t$ | multi-view combination parameter for testing image $\widetilde{\boldsymbol{x}}_t$ |
| $\widetilde{y}_t$ | predicted class of testing image $\widetilde{\boldsymbol{x}}_t$ |
| $\boldsymbol{\gamma}_v$ | parameter of classifier $f^v$ |

## III. IMAGE CLASSIFICATION WITH VISUAL, SEMANTIC AND VIEW CONSISTENCY

In this section, we give the details of the proposed multi-view image classification method using visual, semantic and view consistency.

### A. Linear Multi-View Combination

Considering the two-class classification problem as an example, suppose that we have a total of $V$ views with images in the $v$-th view denoted by $(\boldsymbol{x}_n^v, y_n), n = 1\ldots, N$. $N$ is the number of images, $\boldsymbol{x}_n^v$ is the visual representation of the $n$-th image in the $v$-th view, and $y_n$ is the label of the $n$-th image. The extension to multi-class is straightforward. In this paper, views refer to visual representations generated by various state-of-the-art deep convolutional neural network-based methods [44]–[47]. Note that image representations generated using various visually based transformations [48] can also be used. Classifier $f^v(*)$ is the single-view classifier with the corresponding image representations.

To use the discriminative information of $V$ views, we linearly combine the outputs of $f^v(*), v = 1, \ldots, V$ to predict the class of image $\boldsymbol{x}_n^v$ as

$$\widehat{y}_n = \sum_{v=1}^{V} \alpha_n^v f^v(\boldsymbol{x}_n^v) \tag{1}$$

where $\widehat{y}_n$ is the predicted label of image $\boldsymbol{x}_n^v$.

Let $\boldsymbol{\alpha}_n = [\alpha_n^1; \ldots; \alpha_n^V]$ and $\boldsymbol{F}(\boldsymbol{x_n}) = [f^1(\boldsymbol{x}_n^1); \ldots; f^V(\boldsymbol{x}_n^V)]$, where $\alpha_n^v, v = 1, \ldots, V$ are the linear combination parameters of the $n$-th image that can be

learned by minimizing the classification error as follows:

$$(\boldsymbol{\alpha}_n, \boldsymbol{F}(*)) = argmin_{(\boldsymbol{\alpha}_n, \boldsymbol{F}(*))} \ell(\boldsymbol{\alpha}_n^T \boldsymbol{F}(\boldsymbol{x_n}), y_n) \tag{2}$$

where $\boldsymbol{\alpha}_n^T$ is the transpose of $\boldsymbol{\alpha}_n$. The classification loss is denoted by $\ell(*, *)$. A quadratic hinge loss is used:

$$\ell(\boldsymbol{\alpha}_n^T \boldsymbol{F}(\boldsymbol{x_n}), y_n) = max^2(0, 1 - \boldsymbol{\alpha}_n^T \boldsymbol{F}(\boldsymbol{x_n}) y_n) \tag{3}$$

We use the quadratic hinge loss because it is differentiable and is often used for classification tasks.

### *B. Visual, Semantic and View Consistency for Classification*

Eq. 2 ignores visual consistency. The values predicted for visually similar images should be similar in each view. Hence, we add a visual consistency constraint in a single view to Eq. 2 as

$$\begin{aligned}(\boldsymbol{\alpha}_n, \boldsymbol{F}(*)) = & \, argmin_{(\boldsymbol{\alpha}_n, \boldsymbol{F}(*))} \ell(\boldsymbol{\alpha}_n^T \boldsymbol{F}(\boldsymbol{x_n}), y_n) \\ & + \beta_1 \sum_{v=1}^{V} \sum_{m=1}^{M_v} \| f^v(\boldsymbol{x}_n^v) - f^v(\boldsymbol{x}_m^v) \|^2 e^{-\|\boldsymbol{x}_n^v - \boldsymbol{x}_m^v\|^2/\sigma}\end{aligned} \tag{4}$$

where $\beta_1$ is the weighting parameter of the visual consistency constraint in a single view. Parameter $\sigma$ is the scaling parameter that controls the influence of view consistency. $M_v$ is the number of nearby images for the $v$-th view. The neighbours are selected using Euclidean distance. In this paper, we simply set $M_v$ to 5 consistently with [55].

In addition, the predicted values (semantics) of visually nearby images should be similar when multi-view combination is used. We similarly add a semantic consistency constraint to Eq. 4 as follows:

$$\begin{aligned}(\boldsymbol{\alpha}_n, \boldsymbol{F}(*)) = & \, argmin_{(\boldsymbol{\alpha}_n, \boldsymbol{F}(*))} \ell(\boldsymbol{\alpha}_n^T \boldsymbol{F}(\boldsymbol{x_n}), y_n) \\ & + \sum_{v=1}^{V} \sum_{m=1}^{M_v} (\beta_1 \| f^v(\boldsymbol{x}_n^v) - f^v(\boldsymbol{x}_m^v) \|^2 e^{-\|\boldsymbol{x}_n^v - \boldsymbol{x}_m^v\|^2/\sigma} \\ & + \beta_2 \| \boldsymbol{F}(\boldsymbol{x_n}) - \boldsymbol{F}(\boldsymbol{x_m}) \|^2 e^{-\|\boldsymbol{x}_n^v - \boldsymbol{x}_m^v\|^2/\sigma})\end{aligned} \tag{5}$$

where $\beta_2$ is the weighting parameter of the semantic consistency constraint.

Moreover, the view similarity of nearby images should also be combined. We add a view consistency constraint to Eq. 5 as follows:

$$\begin{aligned}(\boldsymbol{\alpha}, \boldsymbol{F}(*)) = & \, argmin_{(\boldsymbol{\alpha}, \boldsymbol{F}(*))} \sum_{n=1}^{N} (\ell(\boldsymbol{\alpha}_n^T \boldsymbol{F}(\boldsymbol{x_n}), y_n) \\ & + \sum_{v=1}^{V} \sum_{m=1}^{M_v} (\beta_1 \| f^v(\boldsymbol{x}_n^v) - f^v(\boldsymbol{x}_m^v) \|^2 e^{-\|\boldsymbol{x}_n^v - \boldsymbol{x}_m^v\|^2/\sigma} \\ & + \beta_2 \| \boldsymbol{F}(\boldsymbol{x_n}) - \boldsymbol{F}(\boldsymbol{x_m}) \|^2 e^{-\|\boldsymbol{x}_n^v - \boldsymbol{x}_m^v\|^2/\sigma} \\ & + \beta_3 \| \alpha_n^v - \alpha_m^v \|^2 e^{-\|\boldsymbol{x}_n^v - \boldsymbol{x}_m^v\|^2/\sigma}))\end{aligned} \tag{6}$$

where $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_1; \ldots; \boldsymbol{\alpha}_N]$. $\beta_3$ is the weighting parameter of view consistency. Furthermore, we add a regularization term to the linear combination parameter $\boldsymbol{\alpha}_n^v$. Accordingly, let $\boldsymbol{x}_n = [\boldsymbol{x}_n^1; ..; \boldsymbol{x}_n^V]$; then, the overall objective function can be written as

$$\begin{aligned}(\boldsymbol{\alpha}, \boldsymbol{F}(*)) = & \, argmin_{(\boldsymbol{\alpha}, \boldsymbol{F}(*))} \sum_{n=1}^{N} (\ell(\boldsymbol{\alpha}_n^T \boldsymbol{F}(\boldsymbol{x_n}), y_n) \\ & + \sum_{v=1}^{V} \sum_{m=1}^{M_v} (\beta_1 \| f^v(\boldsymbol{x}_n^v) - f^v(\boldsymbol{x}_m^v) \|^2 e^{-\|\boldsymbol{x}_n^v - \boldsymbol{x}_m^v\|^2/\sigma} \\ & + \beta_2 \| \boldsymbol{F}(\boldsymbol{x_n}) - \boldsymbol{F}(\boldsymbol{x_m}) \|^2 e^{-\|\boldsymbol{x}_n^v - \boldsymbol{x}_m^v\|^2/\sigma} \\ & + \beta_3 \| \alpha_n^v - \alpha_m^v \|^2 e^{-\|\boldsymbol{x}_n^v - \boldsymbol{x}_m^v\|^2/\sigma}) + \beta_4 \|\boldsymbol{\alpha}_n\|^2)\end{aligned} \tag{7}$$

### *C. Optimization*

We can perform optimization in Eq. 7 to determine the optimal $\boldsymbol{\alpha}$ and $\boldsymbol{F}(*)$. However, it is very hard to jointly optimize over $\boldsymbol{\alpha}$ and $\boldsymbol{F}(*)$. It is more feasible to alternatively optimize over $\boldsymbol{\alpha}/\boldsymbol{F}(*)$ while keeping $\boldsymbol{F}(*)/\boldsymbol{\alpha}$ fixed.

If $\boldsymbol{F}(*)$ is fixed, Eq. 7 is equivalent to

$$\begin{aligned}\boldsymbol{\alpha} = & \, argmin_{\boldsymbol{\alpha}} \sum_{n=1}^{N} (\ell(\boldsymbol{\alpha}_n^T \boldsymbol{F}(\boldsymbol{x_n}), y_n) \\ & + \beta_3 \sum_{v=1}^{V} \sum_{m=1}^{M_v} \| \alpha_n^v - \alpha_m^v \|^2 e^{-\|\boldsymbol{x}_n^v - \boldsymbol{x}_m^v\|^2/\sigma} \\ & + \beta_4 \|\boldsymbol{\alpha}_n\|^2)\end{aligned} \tag{8}$$

However, jointly optimizing over all $N$ images is still very difficult. We can simplify and iteratively optimize over each image by keeping the other combination parameters fixed. In this way, Eq. 8 can be optimized as follows:

$$\begin{aligned}\boldsymbol{\alpha}_n = & \, argmin_{\boldsymbol{\alpha}_n} \ell(\boldsymbol{\alpha}_n^T \boldsymbol{F}(\boldsymbol{x_n}), y_n) \\ & + \beta_3 \sum_{v'=1}^{v-1} \sum_{m=1}^{M_{v'}} \| \alpha_n^{v'} - \alpha_m^{v'} \|^2 e^{-\|\boldsymbol{x}_n^{v'} - \boldsymbol{x}_m^{v'}\|^2/\sigma} \\ & + \beta_3 \sum_{m=1}^{M_v} \| \alpha_n^v - \alpha_m^v \|^2 e^{-\|\boldsymbol{x}_n^v - \boldsymbol{x}_m^v\|^2/\sigma} \\ & + \beta_3 \sum_{v'=v+1}^{V} \sum_{m=1}^{M_{v'}} \| \alpha_n^{v'} - \alpha_m^{v'} \|^2 e^{-\|\boldsymbol{x}_n^{v'} - \boldsymbol{x}_m^{v'}\|^2/\sigma} \\ & + \beta_4 \|\boldsymbol{\alpha}_n\|^2, \quad \forall n = 1, \ldots, N\end{aligned} \tag{9}$$

This problem can be solved over each view while keeping the combination parameters of other views fixed. In this way, the second term, the fourth term and part of the fifth term are fixed which have no influences on the optimization. Let

$$\begin{aligned}L(\alpha_n^v) = & \, max^2(0, \epsilon - \alpha_n^v f^v(\boldsymbol{x}_n^v)) \\ & + \beta_3 \sum_{m=1}^{M_v} \| \alpha_n^v - \alpha_m^v \|^2 e^{-\|\boldsymbol{x}_n^v - \boldsymbol{x}_m^v\|^2/\sigma} + \beta_4 \|\alpha_n^v\|^2\end{aligned} \tag{10}$$

where $\epsilon = 1 - \sum_{i=1, i \neq v}^{V} \alpha_n^i f^i(\boldsymbol{x}_n^i)$. Note that $\alpha_m^v$ is fixed while optimizing over $\alpha_n^v$ in Eq. 10. We can calculate the derivative of $L(\alpha_n^v)$ as

$$\begin{aligned}\frac{\partial L(\alpha_n^v)}{\partial \alpha_n^v} = & \, 2(\beta_3 \sum_{m=1}^{M_v} (\alpha_n^v - \alpha_m^v) e^{-\|\boldsymbol{x}_n^v - \boldsymbol{x}_m^v\|^2/\sigma} \\ & - max(0, \epsilon - \alpha_n^v f^v(\boldsymbol{x}_n^v) y_n) f^v(\boldsymbol{x}_n^v) y_n + \beta_4 \alpha_n^v)\end{aligned} \tag{11}$$

If $\boldsymbol{\alpha}$ is fixed, Eq. 7 can be rewritten as

$$\begin{aligned}\boldsymbol{F}(*) = argmin_{\boldsymbol{F}(*)} \sum_{n=1}^{N} (\ell(\boldsymbol{\alpha}_n^T \boldsymbol{F}(\boldsymbol{x_n}), y_n) \\ + \sum_{v=1}^{V} \sum_{m=1}^{M_v} (\beta_1 \| f^v(\boldsymbol{x}_n^v) - f^v(\boldsymbol{x}_m^v) \|^2 e^{-\|\boldsymbol{x}_n^v - \boldsymbol{x}_m^v\|^2/\sigma} \\ + \beta_2 \| \boldsymbol{F}(\boldsymbol{x_n}) - \boldsymbol{F}(\boldsymbol{x_m}) \|^2 e^{-\|\boldsymbol{x}_n^v - \boldsymbol{x}_m^v\|^2/\sigma}) \end{aligned} \tag{12}$$

This problem can also be iteratively optimized over classifiers of each view as follows:

$$f^v(*) = argmin_{f^v(*)} \mathcal{L}(f^v(*)) \forall v = 1, \ldots, V \tag{13}$$

with

$$\begin{aligned}\mathcal{L}(f^v(*)) = \sum_{n=1}^{N} (\ell(\alpha_n^v f^v(\boldsymbol{x}_n^v) + \Omega, y_n) \\ + \sum_{m=1}^{M_v} (\beta_1 \| f^v(\boldsymbol{x}_n^v) - f^v(\boldsymbol{x}_m^v) \|^2 e^{-\|\boldsymbol{x}_n^v - \boldsymbol{x}_m^v\|^2/\sigma} \\ + \beta_2 \| f^v(\boldsymbol{x}_n^v) - f^v(\boldsymbol{x}_m^v) + \Phi \|^2 e^{-\|\boldsymbol{x}_n^v - \boldsymbol{x}_m^v\|^2/\sigma}) \\ \Phi = \sum_{i=1, i\neq v}^{V} f^i(\boldsymbol{x}_n^i) - f^i(\boldsymbol{x}_m^i) \\ \Omega = \sum_{i=1, i\neq v}^{V} \alpha_n^i f^i(\boldsymbol{x}_n^i) \end{aligned} \tag{14}$$

In this paper, we use the sigmoid classifier as $f^v(*)$:

$$f^v(\boldsymbol{x}) = \frac{1}{1 + e^{-\boldsymbol{\gamma}_v^T \boldsymbol{x}}}, \quad \forall v = 1, \ldots, V \tag{15}$$

and its optimization is quite straightforward. Parameter $\boldsymbol{\gamma}_v$ can be optimized with

$$\frac{\partial \mathcal{L}(f^v(*))}{\partial \boldsymbol{\gamma}_v} = \frac{\partial \mathcal{L}(f^v(*))}{\partial f^v(*)} \frac{\partial f^v(*)}{\partial \boldsymbol{\gamma}_v} \tag{16}$$

Once the optimal $\boldsymbol{\alpha}$ and $\boldsymbol{F}(*)$ have been learned, we can predict the classes of images accordingly.

### *D. Image Class Prediction*

For each testing image, we learn the combination parameter to predict its class using Eq.1 with fixed $\boldsymbol{F}(*)$. The combination parameter can be learned similarly to Eq.9 without considering the quadratic hinge loss term $\ell(*,*)$. Let $\widetilde{\boldsymbol{x}}_t^v$ be the representation of one testing image of the $v$-th view, $v = 1, \ldots, V$; the multi-view combination parameter $\widetilde{\boldsymbol{\alpha}}_t = [\widetilde{\alpha}_t^1; ..; \widetilde{\alpha}_t^v; ..; \widetilde{\alpha}_t^V]$ can be learned by solving

$$\widetilde{\boldsymbol{\alpha}}_t = argmin_{\widetilde{\boldsymbol{\alpha}}_t} \beta_4 \| \widetilde{\boldsymbol{\alpha}}_t \|^2 + \beta_3 \sum_{v=1}^{V} \sum_{m=1}^{\widetilde{M}_v} \| \widetilde{\alpha}_t^v - \alpha_m^v \|^2 e^{-\|\widetilde{\boldsymbol{x}}_t^v - \boldsymbol{x}_m^v\|^2/\sigma} \tag{17}$$

where $\widetilde{M}_v$ is the number of nearby images of $\widetilde{\boldsymbol{x}}_t^v$. The image class $\widetilde{y}_t$ can then be predicted as

$$\widetilde{y}_t = \sum_{v=1}^{V} \widetilde{\alpha}_t^v f^v(\widetilde{\boldsymbol{x}}^v) \tag{18}$$

**Algorithm 1** Steps of the Proposed Multi-View Combination-Based Image Classification Method With Visual, Semantic and View Consistency

**Input:**

Parameters $\beta_1$, $\beta_2$, $\beta_3$, $\beta_4$, $\sigma$; training images $(\boldsymbol{x}_n^v, y_n)$ of multi-views; maximum number of iterations $M_{it}$; testing image $\widetilde{\boldsymbol{x}}_t$ of multi-views $\widetilde{\boldsymbol{x}}_t^v$, $v = 1, ..., V$.

**Output:**

The predicted class $\widetilde{y}_t$ of testing image $\widetilde{\boldsymbol{x}}_t$.

1: **for** i=1:$M_{it}$
2: Determine the optimal $\boldsymbol{\alpha}$ with $\boldsymbol{F}(*)$ fixed by solving Eq. 8. This optimization can be performed iteratively for each image with Eq. 9;
3: Determine the optimal $\boldsymbol{F}(*)$ with $\boldsymbol{\alpha}$ fixed by solving Eq. 12. This optimization can be performed iteratively for each view with Eq. 13;
4: **end for**.
5: Determine the optimal combination parameter $\widetilde{\boldsymbol{\alpha}}_t$ of testing image $\widetilde{\boldsymbol{x}}_t$ by solving Eq.15;
6: Predict the testing image's class using Eq. 16;
7: **return** The predicted class $\widetilde{y}_t$.

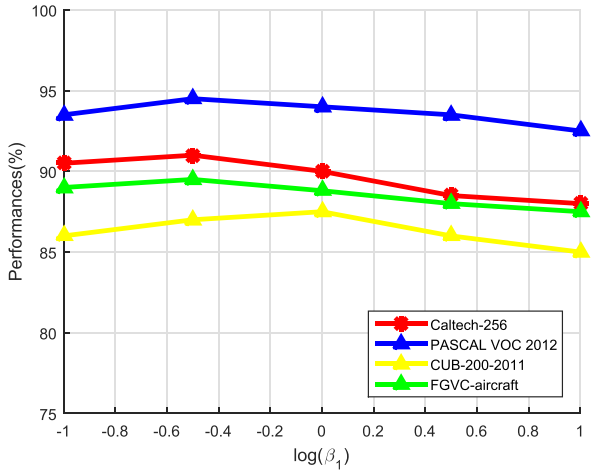


Fig. 2. Influences of $\beta_1$ on the four datasets.

Algorithm 1 describes the steps of the proposed multi-view combination-based image classification method with visual, semantic and view consistency.

## IV. Experiments

To evaluate the effectiveness of the proposed multi-view image classification method using visual, semantic and view consistency (VSVC), we conduct image classification experiments on the Caltech-256 [49], PASCAL VOC 2012 [50], CUB-200-2011 [59], and FGVC-aircraft [60] datasets. Figure 2 shows several example images of the Caltech-256, PASCAL VOC 2012, CUB-200-2011, and FGVC-aircraft datasets.

### *A. Experimental Setup*

We use four widely used deep convolutional neural networks (AlexNet [44], VGG [45], GoogleNet [46] and ResNet [47]) as

TABLE II
CLASSIFICATION PERFORMANCES OF THE PROPOSED VSVC METHOD AND OTHER BASELINE METHODS ON THE CALTECH-256 DATASET

| Methods | 15 images | 30 images | 45 images | 60 images |
|---|---|---|---|---|
| MVLS-LC [2] | 72.53 ± 0.84 | 75.68 ± 0.61 | 77.45 ± 0.72 | 78.82 ± 0.69 |
| MVLS [2] | 77.39 ± 0.62 | 84.23 ± 0.58 | 87.72 ± 0.56 | 89.95 ± 0.63 |
| SWSS-VGG[12] | 69.37 ± 0.46 | 73.56 ± 0.51 | 74.83 ± 0.39 | 76.25 ± 0.47 |
| ObjectBank [31] | 39.00 | – | – | – |
| VGG [45] | – | – | – | 86.20 ± 0.30 |
| KSPM [49] | 23.34 ± 0.42 | 29.51 ± 0.52 | – | – |
| LR-GCC [52] | 39.21 ± 0.48 | 45.87 ± 0.41 | – | – |
| VUCN [53] | 65.70 ± 0.20 | 70.60 ± 0.20 | 72.70 ± 0.40 | 74.20 ± 0.30 |
| NBNN [54] | 30.45 | 38.18 | – | – |
| LLC [55] | 27.74 ± 0.32 | 32.07 ± 0.24 | 35.09 ± 0.44 | 37.79 ± 0.42 |
| FV+$L^2$EMG [69] | 45.00 ± 0.20 | 53.60 ± 0.30 | 58.20 ± 0.30 | 61.80 ± 0.40 |
| SDC [70] | 35.10 | 42.10 | 45.70 | 47.90 |
| MSC [71] | 40.50 ± 0.40 | 48.00 ± 0.20 | 51.90 ± 0.20 | 55.20 ± 0.30 |
| VSVC-LC | 75.18 ± 0.67 | 78.95 ± 0.54 | 79.57 ± 0.61 | 80.69 ± 0.54 |
| VSVC(average) | 80.03 ± 0.59 | 86.16 ± 0.55 | 89.55 ± 0.43 | 90.87 ± 0.46 |
| VSVC | 80.59 ± 0.57 | 86.71 ± 0.52 | 90.07 ± 0.46 | 91.35 ± 0.43 |

four views for image representation and classification. We pre-train these networks on the ImageNet 2012 dataset and fine-tune these networks on the Caltech-256, PASCAL VOC 2012, CUB-200-2011, and FGVC-aircraft datasets. We remove the last fully connected layer and use 4,096 dimensions of the penultimate layer as representations of the corresponding view. Since local feature-based methods are often evaluated on the Caltech-256 dataset, we show the performance of using local features with multi-views on the Caltech-256 dataset (VSVC-LC [2]). We also show the outcome of using the average value of predicted results of multi-views (VSVC(average)). We use 30 DELL EMC PowerEdge C4130 with 200 GPU (Nvidia P80 and P100) which can do 1107T floating-point operations per second. We follow the same experimental setup as in [2] by using the same six views. Local features are densely extracted with 16×16 pixels and 6 pixels of overlap. The codebook size is set to 1,000 as in [2]. We set $\sigma$ to the medium value of all $\|\boldsymbol{x}_n^v - \boldsymbol{x}_m^v\|$. The maximum number of iterations is set to 60 in Algorithm 1. For a fair comparison, we follow the same experimental setup as other researchers and use the same number of training images. The optimal parameters are determined by ten-fold cross validation. We directly compare with the results reported by other baseline methods instead of re-implementing them. The average per-class classification rate is used to quantitatively evaluate the effectiveness of the proposed method.

### B. The Caltech-256 Dataset

There are 256 classes of images in this dataset. The total number of images is 29,780. We randomly select 15/30/45/60 training images per class for training and use the other images for testing, as in [49]. We randomly select images ten times for classification. The mean and standard deviation of the results are used for evaluations. We also show the performance of the proposed VSVC method with local features by combining the same six views as in MVLS-LC [2]. Table 2 presents the performance of the proposed VSVC method and other baseline methods.

We can make four conclusions based on Table 2. First, compared with single-view-based methods [45], [49], the combination of multi-view data is more useful for image classification. Second, the proposed VSVC method improves over other similarity-based methods [52], [54], [55]. This is because we also use the semantic and view consistency information. Third, compared with other multi-view-based methods [2], [12], [54], [69]–[71], the proposed VSVC method achieves a superior performance. We jointly consider visual, semantic and view consistency, while other methods only use visual or semantic correlations. Additionally, VSVC dynamically optimizes over multi-view combination and single-view classifier training. Fourth, VSVC-LC is also able to improve over MVLS-LC [2] using the same views with local features. The performance can be further improved using deep convolutional neural networks.

TABLE III
IMAGE CLASSIFICATION PERFORMANCES OF VSVC METHOD AND OTHER BASELINE METHODS ON THE PASCAL VOC 2012 DATASET

| Methods | mean average precision (mAP) |
|---|---|
| MVLS [2] | 93.1 |
| VGG [45] | 89.0 |
| VGG-16-19-SVM [46] | 89.3 |
| VUCN [53] | 79.0 |
| Chatfield [56] | 83.2 |
| HCP-Alex[57] | 81.8 |
| HCP-VGG [57] | 90.5 |
| HCP-Combined [57] | 93.2 |
| NUS-PSL [72] | 82.2 |
| Zeiler [72] | 79.0 |
| VSVC(average) | 94.3 |
| VSVC | 94.8 |

### C. The PASCAL VOC 2012 Dataset

This dataset has twenty classes (*aeroplane, bicycle, boat, bottle, bus, bird, car, cat, cow, chair, dining table, dog, horse, person, sheep, motorbike, train, potted plant, sofa* and *tv/monitor*). There are a total of 22,531 images. Images are split into a training and validation set and a test set with 11,540 and 10,991 images, respectively. We follow the same experimental setup as in [50]. We use the training set to train classifiers and use the validation set for parameter tuning.

TABLE IV
PER-CLASS MAP COMPARISONS ON THE PASCAL VOC 2012 DATASET

| object class | VUCN[53] | Chatfield[56] | VGG[45] | HCP-VGG[57] | MVLS[2] | VSVC |
|---|---|---|---|---|---|---|
| airplane | 96.0 | 96.8 | 99.0 | 99.1 | 99.4 | 99.6 |
| bicycle | 77.1 | 82.5 | 88.8 | 92.8 | 95.3 | 96.8 |
| bird | 88.4 | 91.5 | 95.9 | 97.4 | 98.7 | 99.3 |
| boat | 85.5 | 88.1 | 93.8 | 94.4 | 96.1 | 97.7 |
| bottle | 55.8 | 62.1 | 73.1 | 79.9 | 82.8 | 88.9 |
| bus | 85.8 | 88.3 | 92.1 | 93.6 | 95.2 | 98.1 |
| car | 78.6 | 81.9 | 85.1 | 89.8 | 92.6 | 96.2 |
| cat | 91.2 | 94.8 | 97.8 | 98.2 | 98.9 | 99.2 |
| chair | 65.0 | 70.3 | 79.5 | 78.2 | 83.4 | 86.1 |
| cow | 74.4 | 80.2 | 91.1 | 94.9 | 96.5 | 98.6 |
| table | 67.7 | 76.2 | 83.3 | 79.8 | 83.6 | 86.4 |
| dog | 87.8 | 92.9 | 97.2 | 97.8 | 98.9 | 99.3 |
| horse | 86.0 | 90.3 | 96.3 | 97.0 | 99.1 | 99.6 |
| motorbike | 85.1 | 89.3 | 94.5 | 93.8 | 96.7 | 98.7 |
| person | 90.9 | 95.2 | 96.9 | 96.4 | 97.2 | 98.9 |
| plant | 52.2 | 57.4 | 63.1 | 74.3 | 79.5 | 81.5 |
| sheep | 83.6 | 83.6 | 93.4 | 94.7 | 95.6 | 96.8 |
| sofa | 61.1 | 66.4 | 75.0 | 71.9 | 77.8 | 81.6 |
| train | 91.8 | 93.5 | 97.1 | 96.7 | 98.4 | 99.1 |
| tv | 76.1 | 81.9 | 87.1 | 88.6 | 92.1 | 93.5 |

The training and validation images are then merged for classifier training with the selected parameters.

Table 3 shows performance comparisons of VSVC with other baseline methods. We can see that the proposed VSVC method improves over these baseline methods. In particular, VSVC outperforms single-view-based methods [45], [56]. Additionally, VSVC improves over [57], which uses detection information for classification. Moreover, the proposed method performs better than other multi-view or combination-based methods [2], [46]. We also show the per-class performance results in Table 4. We can see from Table 4 that VSVC is able to improve over other methods on all the twenty classes with multi-view combination. Furthermore, the improvements on rigid objects are not as large as those on non-rigid objects. The non-rigid objects have larger visual variations than those of rigid objects. The joint consideration of visual, semantic and view consistency helps alleviate this problem.

### *D. The CUB-200-2011 Dataset*

There are 11,788 images that belong to 200 different bird species. The images are pre-divided into training and test sets. Both image labels and bounding boxes are given. In this paper, we only use image labels for classification. Performance comparisons of the proposed method with other baseline methods are presented in Table 5.

We can draw three conclusions from Table 5. First, the proposed VSVC method is able to outperform many the state-of-the-art methods on this dataset, even when bounding box information is used [18], [63], [66]. This result proves the effectiveness of the proposed method. Second, VSVC is able to improve over methods based on VGG [45], [61], [65], [66], AlexNet [18], GoogleNet [63], [64] and ResNet [73], [74]. Third, the proposed method can also achieve performance superior to that of combinational-based methods [64], [65], [74].

TABLE V
PERFORMANCE COMPARISONS OF VSVC AND OTHER BASELINE METHODS ON THE CUB-200-2011 DATASET. BB: BOUNDING BOX

| Methods | BB | Performance (%) |
|---|---|---|
| FC-VGG [45] | no | 70.4 |
| bilinear CNN [61] | no | 84.1 |
| LRBP[62] | no | 84.2 |
| PR-CNN [16] | yes | 73.5 |
| Triplet-A [63] | yes | 80.7 |
| STN [64] | no | 84.1 |
| BoostCNN [65] | no | 86.2 |
| VGG-D [66] | yes | 82.0 |
| RBF [73] | no | 79.0 |
| AGAL [74] | yes | 85.5 |
| VSVC(average) | no | 87.5 |
| VSVC | no | 87.9 |

### *E. The FGVC-Aircraft Dataset*

This dataset has 10,000 images of aircraft. There are 100 classes that are very similar. We test the performance of the proposed VSVC method and show the classification accuracy in Table 6 along with the respective figures for other baseline methods. We can draw similar conclusions as from Table 5. The proposed VSVC method improves over other baseline methods. Additionally, compared with non-rigid bird images, the rigid aircraft often occupy larger portions of images. We are able to improve the classification performance by jointly considering the visual, semantic and view consistency. VSVC also improves over boostCNN, which tries to combine a series of convolutional neural networks for classification. The proposed VSVC method also improves over many manifold regularization-based methods. By jointly imposing visual, semantic and view consistency, we can improve the classification accuracy over that of traditional manifold regularization-based methods. Finally, VSVC(average) performs not as good as VSVC. Average operation can be used to speed up the computation.

TABLE VI

PERFORMANCE COMPARISONS OF VSVC AND OTHER BASELINE METHODS ON THE FGVC-AIRCRAFT DATASET

| Methods | BB | Performance (%) |
|---|---|---|
| bilinear CNN [61] | no | 84.5 |
| BoostCNN [65] | no | 88.5 |
| BoT [69] | yes | 88.4 |
| FV [68] | no | 80.7 |
| VSVC(average) | no | 89.6 |
| VSVC | no | 90.2 |

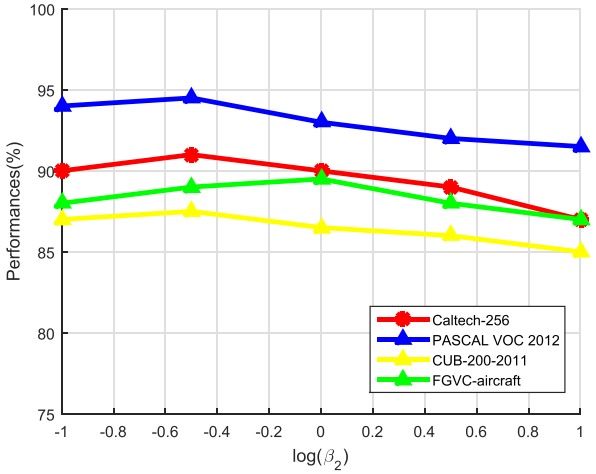


Fig. 3. Influences of $\beta_2$ on the four datasets.

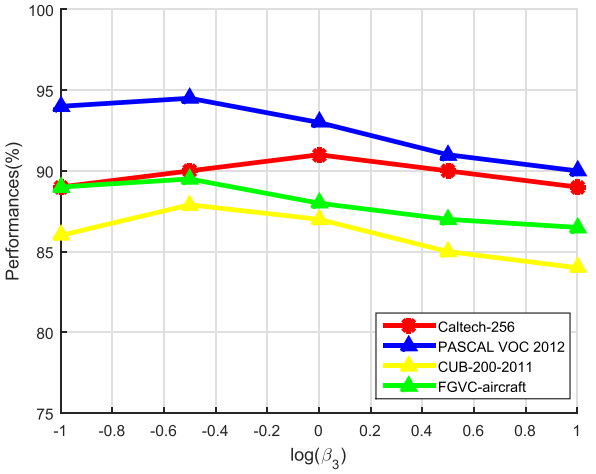


Fig. 4. Influences of $\beta_3$ on the four datasets.

### F. Influence of Parameters

Parameters $\beta_1$, $\beta_2$ and $\beta_3$ control the three aspects of consistency constraints. We plot their influences on the Caltech-256 dataset (60 training images), the PASCAL VOC 2012 dataset, the CUB-200-2011 dataset, and the FGVC-aircraft dataset in Figure 3, Figure 4 and Figure 5, respectively. We observe that the performance is unsatisfactory if $\beta_1$, $\beta_2$ and $\beta_3$ are set to large or small values. This observation shows that imposing too small or very large consistency constraints leads to performance degradation. We can see from Figure 3, Figure 4 and Figure 5 that setting $\beta_1$, $\beta_2$ and $\beta_3$ to 0.1∼1 results in satisfactory performance.

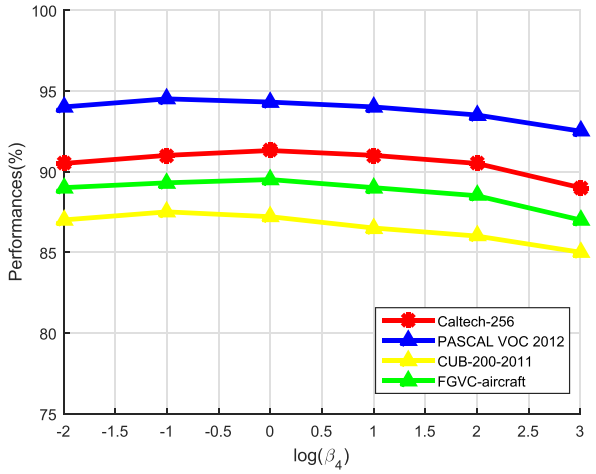


Fig. 5. Influences of $\beta_4$ on the four datasets.

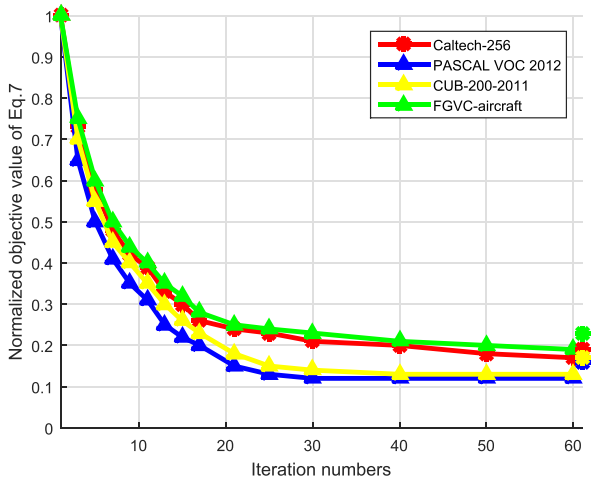


Fig. 6. Normalized objective values of Eq.7 with the number of iterations on the four datasets. The normalized objective value of one trivial solution ($\alpha_n^v = 2y_n/V$, $f^v(x_n^v) = 0.5$) is also given.

Parameter $\beta_4$ controls the influence of the regularization term. We plot its influence on the four datasets in Figure 6. We can see from Figure 6 that the results are not as sensitive to $\beta_4$ as they are to $\beta_1$, $\beta_2$ and $\beta_3$. The performance remains relatively stable as long as $\beta_4$ is not set to very large values.

The proposed method can iteratively minimize the objective value of Eq. 7. To demonstrate the robustness of the proposed method, we randomly initialize the parameters for 100 times and plot the mean of the normalized objective values of Eq.7 of each iteration on the four datasets in Figure 6. For each random initialization process, the normalized objective values of the $M_{it}$ iterations is obtained by normalizing their values with the initial value of Eq.7. The normalized objective value of one trivial solution ($\alpha_n^v = 2y_n/V$, $f^v(x_n^v) = 0.5$) is also given in Figure 6. We can see from Figure 6 that the objective value of Eq. 7 corresponding to the optimized parameters is much smaller than a trivial solution.

### G. Ablation Study

We use visual, semantic and view consistency for image classification. To show their influences, we present the performance resulting from imposing different consistency

TABLE VII

INFLUENCES OF VISUAL, SEMANTIC AND VIEW CONSISTENCY ON THE CALTECH-256 DATASET (60 TRAINING IMAGES), THE PASCAL VOC 2012 DATASET, THE CUB-200-2011 DATASET, AND THE FGVC-AIRCRAFT DATASET. NO VISUAL, NO SEMANTIC AND NO VIEW REPRESENT THE PERFORMANCES OF THE PROPOSED VSVC METHOD BY SETING $\beta_1$, $\beta_2$ AND $\beta_3$ TO 0 RESPECTIVELY

| Datasets | no visual | no semantic | no view | VSVC |
|---|---|---|---|---|
| Caltech-256 (60 images) | 88.7 | 90.2 | 89.5 | 91.3 |
| PASCAL VOC 2012 | 92.4 | 93.1 | 91.6 | 94.8 |
| CUB-200-2011 | 85.3 | 86.4 | 85.5 | 87.9 |
| FGVC-aircraft | 88.2 | 89.1 | 89.5 | 90.2 |

constraints on the four datasets in Table 7. We can see from Table 7 that the three consistency constraints help improve the classification performance. In addition, the relative improvements due to the three consistency constraints vary on different datasets. We believe that this result occurs because of the class variations of different datasets. Visual similarity is less reliable for images with large inter-class variations. This problem can be alleviated by using semantic and view consistency.

## V. CONCLUSION

In this paper, we proposed a multi-view image classification method with visual, semantic and view consistency. We linearly combined multiple views for classification. The combination parameters were determined by exploring the classification loss and visual, semantic and view consistency constraints. First, we ensured that visually similar images were predicted to have similar values. Second, we used semantic consistency to ensure the smoothness of predictions. Third, similar images were also combined with consistent views. We conducted image classification experiments on several public datasets, and the experimental results demonstrated the effectiveness of the proposed method.

## REFERENCES


[1] C. Shi, G. An, R. Zhao, Q. Ruan, and Q. Tian, "Multiview Hessian semisupervised sparse feature selection for multimedia analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 9, pp. 1947–1961, Sep. 2017.

[2] C. Zhang, J. Cheng, and Q. Tian, "Multiview label sharing for visual representations and classifications," *IEEE Trans. Multimedia*, vol. 20, no. 4, pp. 903–913, Apr. 2018.

[3] C. Zhang and W.-S. Zheng, "Semi-supervised multi-view discrete hashing for fast image search," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2604–2617, Jun. 2017.

[4] H. Tao, C. Hou, F. Nie, J. Zhu, and D. Yi, "Scalable multi-view semi-supervised classification via adaptive regression," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4283–4296, Sep. 2017.

[5] Y. Wang, L. Wu, X. Lin, and J. Gao, "Multiview spectral clustering via structured low-rank matrix factorization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 10, pp. 4833–4843, Oct. 2018. doi: 10.1109/TNNLS.2017.2777489.

[6] X. Shen, W. Liu, I. W. Tsang, Q.-S. Sun, and Y.-S. Ong, "Multilabel prediction via cross-view search," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 9, pp. 4324–4338, Sep. 2018. doi: 10.1109/TNNLS.2017.2763967.

[7] J. Wu, S. Pan, X. Zhu, C. Zhang, and P. S. Yu, "Multiple structure-view learning for graph classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 7, pp. 3236–3251, Jul. 2018. doi: 10.1109/TNNLS.2017.2703832.

[8] J. Peng, A. J. Aved, G. Seetharaman, and K. Palaniappan, "Multiview boosting with information propagation for classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 3, pp. 657–669, Mar. 2018. doi: 10.1109/TNNLS.2016.2637881.

[9] C. Zhang, J. Cheng, C. Li, and Q. Tian, "Image-specific classification with local and global discriminations," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 9, pp. 4479–4486, Sep. 2018.

[10] C. Zhang, Q. Huang, and Q. Tian, "Contextual exemplar classifier-based image representation for classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 8, pp. 1691–1699, Aug. 2017.

[11] Y. Lu *et al.*, "Transductive zero-shot learning with a self-training dictionary approach," *IEEE Trans. Cybern.*, vol. 48, no. 10, pp. 2908–2919, Oct. 2018. doi: 10.1109/TCYB.2017.2751741.

[12] C. Zhang, J. Cheng, and Q. Tian, "Structured weak semantic space construction for visual categorization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 8, pp. 3442–3451, Aug. 2018.

[13] W. Li, Z. Xu, D. Xu, D. Dai, and L. Van Gool, "Domain generalization and adaptation using low rank exemplar SVMs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1114–1127, May 2018. doi: 10.1109/TPAMI.2017.2704624.

[14] C. Zhang, J. Cheng, and Q. Tian, "Image-level classification by hierarchical structure learning with visual and semantic similarities," *Inf. Sci.*, vol. 422, pp. 271–281, Jan. 2018.

[15] Z. Hui and A. C. Sankaranarayanan, "Shape and spatially-varying reflectance estimation from virtual exemplars," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 10, pp. 2060–2073, Oct. 2017.

[16] F. Zhu, Z. Yan, J. Bu, and Y. Yu, "Exemplar-based image and video stylization using fully convolutional semantic features," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3542–3555, Jul. 2017.

[17] C. Zhang, J. Liu, Q. Tian, C. Xu, H. Lu, and S. Ma, "Image classification by non-negative sparse coding, low-rank and sparse decomposition," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 1673–1680.

[18] C. Zhang, J. Liu, C. Liang, Z. Xue, J. Pang, and Q. Huang, "Image classification by non-negative sparse coding, correlation constrained low-rank and sparse decomposition," *Comput. Vis. Image Understand.*, vol. 123, pp. 14–22, Jun. 2014.

[19] C. Zhang *et al.*, "Joint image representation and classification in random semantic spaces," *Neurocomputing*, vol. 156, pp. 79–85, May 2015.

[20] X. Cai, F. Nie, W. Cai, and H. Huang, "Heterogeneous image features integration via multi-modal semi-supervised learning model," in *Proc. Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1737–1744.

[21] S. Wang, Z. Ding, and Y. Fu, "Marginalized denoising dictionary learning with locality constraint," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 500–510, Jan. 2018.

[22] C. Zhang, G. Zhu, Q. Huang, and Q. Tian, "Image classification by search with explicitly and implicitly semantic representations," *Inf. Sci.*, vol. 376, pp. 125–135, Jan. 2017.

[23] R. Kwitt, N. Vasconcelos, and N. Rasiwasia, "Scene recognition on the semantic manifold," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 359–372.

[24] C. Zhang, J. Cheng, J. Liu, J. Pang, Q. Huang, and Q. Tian, "Beyond explicit codebook generation: Visual representation using implicitly transferred codebooks," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5777–5788, Dec. 2015.

[25] S. Gao, I. W.-H. Tsang, and L.-T. Chia, "Laplacian sparse coding, hypergraph Laplacian sparse coding, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 92–104, Jan. 2013.

[26] S. Li and Y. Fu, "Learning robust and discriminative subspace with low-rank constraints," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 11, pp. 2160–2173, Nov. 2016.

[27] Z. Ding, M. Shao, and Y. Fu, "Missing modality transfer learning via latent low-rank constraint," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4322–4334, Nov. 2015.

[28] O. Boiman, E. Shechtman, and M. Irani, "In defense of nearest-neighbor based image classification," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[29] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 664–676, Apr. 2017.

[30] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 1778–1785.

[31] L. Li, H. Su, E. P. Xing, and L. Fei-Fei, "Object Bank: A high-level image representation for scene classification & semantic feature sparsification," in *Proc. NIPS*, Vancouver, BC, Canada, 2010, pp. 1378–1386.

[32] C. Zhang, C. Li, D. Lu, J. Cheng, and Q. Tian, "Birds of a feather flock together: Visual representation with scale and class consistency," *Inf. Sci.*, vols. 460–461, pp. 115–127, Sep. 2018.

[33] Z. Xu, S. Huang, Y. Zhang, and D. Tao, "Webly-supervised fine-grained visual categorization via deep domain adaptation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1100–1113, May 2018. doi: 10.1109/TPAMI.2016.2637331.

[34] C. Zhang, G. Zhu, C. Liang, Y. Zhang, Q. Huang, and Q. Tian, "Image class prediction by joint object, context, and background modeling," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 28, no. 2, pp. 428–438, Feb. 2018.

[35] C. Zhang, J. Cheng, and Q. Tian, "Incremental codebook adaptation for visual representation and categorization," *IEEE Trans. Cybern.*, vol. 48, no. 7, pp. 2012–2023, Jul. 2018.

[36] C. Zhang, J. Cheng, and Q. Tian, "Multiview, few-labeled object categorization by predicting labels with view consistency," *IEEE Trans. Cybern.*, vol. 49, no. 11, pp. 3834–3843, Nov. 2019. doi: 10.1109/TCYB.2018.2845912.

[37] J. Tang, Y. Tian, P. Zhang, and X. Liu, "Multiview privileged support vector machines," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 8, pp. 3463–3477, Aug. 2018. doi: 10.1109/TNNLS.2017.2728139.

[38] L. Nie, L. Zhang, Y. Yan, X. Chang, M. Liu, and L. Shaoling, "Multiview physician-specific attributes fusion for health seeking," *IEEE Trans. Cybern.*, vol. 47, no. 11, pp. 3680–3691, Nov. 2017.

[39] J. Li, Y. Wu, J. Zhao, and K. Lu, "Low-rank discriminant embedding for multiview learning," *IEEE Trans. Cybern.*, vol. 47, no. 11, pp. 3516–3529, Nov. 2017.

[40] R. Ma, T. Maugey, and P. Frossard, "Optimized data representation for interactive multiview navigation," *IEEE Trans. Multimedia*, vol. 20, no. 7, pp. 1595–1609, Jul. 2018.

[41] X. Chang, Z. Ma, Y. Yang, Z. Zeng, and A. G. Hauptmann, "Bi-level semantic representation analysis for multimedia event detection," *IEEE Trans. Cybern.*, vol. 47, no. 5, pp. 1180–1197, May 2017.

[42] J. Yu, X. Yang, F. Gao, and D. Tao, "Deep multimodal distance metric learning using click constraints for image ranking," *IEEE Trans. Cybern.*, vol. 47, no. 12, pp. 4014–4024, Dec. 2017.

[43] N. Takahashi, M. Gygli, and L. Van Gool, "AENet: Learning deep audio features for video analysis," *IEEE Trans. Multimedia*, vol. 20, no. 3, pp. 513–524, Mar. 2018.

[44] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

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

[46] C. Szegedy *et al.*, "Going deeper with convolutions," 2014, *arXiv:1409.4842*. [Online]. Available: https://arxiv.org/abs/1409.4842

[47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[48] C. Zhang, J. Liu, C. Liang, Q. Huang, and Q. Tian, "Image classification using Harr-like transformation of local features with coding residuals," *Signal Process.*, vol. 93, no. 8, pp. 2111–2118, Aug. 2013.

[49] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," CalTech, Tech. Rep., 2007.

[50] *The PASCAL Visual Object Classes*. [Online]. Available: http://host.robots.ox.ac.uk/pascal/VOC/

[51] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[52] C. Zhang, C. Liang, L. Li, J. Liu, Q. Huang, and Q. Tian, "Fine-grained image classification via low-rank sparse coding with general and class-specific codebooks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 7, pp. 1550–1559, Jul. 2017.

[53] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," 2013, *arXiv:1311.2901*. [Online]. Available: https://arxiv.org/abs/1311.2901

[54] C. Zhang, J. Cheng, and Q. Tian, "Multiview semantic representation for visual recognition," *IEEE Trans. Cybern.*, to be published. doi: 10.1109/TCYB.2018.2875728.

[55] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3360–3367.

[56] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," 2014, *arXiv:1405.3531*. [Online]. Available: https://arxiv.org/abs/1405.3531

[57] Y. Wei *et al.*, "HCP: A flexible CNN framework for multi-label image classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 9, pp. 1901–1907, Jun. 2015.

[58] J. Donahue *et al.*, "DeCAF: A deep convolutional activation feature for generic visual recognition," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 1–9.

[59] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD birds-200-2011 dataset," Tech. Rep., 2011.

[60] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi, "Fine-grained visual classification of aircraft," 2013, *arXiv:1306.5151*. [Online]. Available: https://arxiv.org/abs/1306.5151

[61] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear convolutional neural networks for fine-grained visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1309–1322, Jun. 2018. doi: 10.1109/TPAMI.2017.2723400.

[62] C. Zhang, S. Wang, Q. Huang, J. Liu, C. Liang, and Q. Tian, "Image classification using spatial pyramid robust sparse coding," *Pattern Recognit. Lett.*, vol. 34, no. 9, pp. 1046–1052, 2013.

[63] Y. Cui, F. Zhou, Y. Lin, and S. Belongie, "Fine-grained categorization and dataset bootstrapping using deep metric learning with humans in the loop," 2015, *arXiv:1512.05227*. [Online]. Available: https://arxiv.org/abs/1512.05227

[64] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc. Neural Inf. Process. Syst.*, 2015, pp. 2017–2025.

[65] M. Moghimi, M. Saberian, J. Yang, L.-J. Li, N. Vasconcelos, and S. Belongie, "Boosted convolutional neural networks," in *Proc. Brit. Mach. Vis. Conf.*, 2016, pp. 24.1–24.13.

[66] J. Krause, H. Jin, J. Yang, and L. Fei-Fei, "Fine-grained recognition without part annotations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 5546–5555.

[67] Y. Wang, J. Choi, V. Morariu, and L. S. Davis, "Mining discriminative triplets of patches for fine-grained classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1163–1172.

[68] P.-H. Gosselin, N. Murray, H. Jégou, and F. Perronnin, "Revisiting the Fisher vector for fine-grained classification," *Pattern Recognit. Lett.*, vol. 49, pp. 92–98, Nov. 2014.

[69] P. Li, Q. Wang, H. Zeng, and L. Zhang, "Local log-Euclidean multivariate Gaussian descriptor and its application to image classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 803–817, Apr. 2017.

[70] K. Sohn, D. Y. Jung, H. Lee, and A. O. Hero, "Efficient learning of sparse, distributed, convolutional feature representations for object recognition," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2643–2650.

[71] L. Bo, X. Ren, and D. Fox, "Multipath sparse coding using hierarchical matching pursuit," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 660–667.

[72] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 391–405.

[73] B. Meyer, B. Harwood, and T. Drummond, "Nearest neighbor radial basis function solvers for deep neural networks," Tech. Rep., 2017.

[74] X. Liu, J. Wang, S. Wen, E. Ding, and Y. Lin, "Localizing by describing: Attribute-guided attention localization for fine-grained recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 4190–4196.

[75] Y. Wang, W. Zhang, L. Wu, X. Lin, and X. Zhao, "Unsupervised metric fusion over multiview data by graph random walk-based cross-view diffusion," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 1, pp. 57–70, Jan. 2017.

[76] S. Eleftheriadis, O. Rudovic, and M. Pantic, "Discriminative shared Gaussian processes for multiview and view-invariant facial expression recognition," *IEEE Trans. Image Process.*, vol. 24, no. 1, pp. 189–204, Jan. 2015.

[77] A. Benton, H. Khayrallah, B. Gujral, D. A. Reisinger, S. Zhang, and R. Arora, "Deep generalized canonical correlation analysis," 2017, *arXiv:1702.02519*. [Online]. Available: https://arxiv.org/abs/1702.02519

[78] W. Wang, R. Arora, K. Livescu, and J. Bilmes, "On deep multi-view representation learning: Objectives and optimization," 2016, *arXiv:1602.01024*. [Online]. Available: https://arxiv.org/abs/1602.01024

[79] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, "Multi-view discriminant analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 188–194, Jan. 2012.

[80] Y. Zhao *et al.*, "Multi-view manifold learning with locality alignment," *Pattern Recognit.*, vol. 78, pp. 154–166, Jun. 2018.

[81] C. Zhang, J. Cheng, and Q. Tian, "Unsupervised and semi-supervised image classification with weak semantic consistency," *IEEE Trans. Multimedia*, to be published. doi: 10.1109/TMM.2019.2903628.
[82] C. Zhang, J. Cheng, and Q. Tian, "Semantically modeling of object and context for categorization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 4, pp. 1013–1024, Apr. 2019. doi: 10.1109/TNNLS.2018.2856096.
[83] X. Lan, S. Zhang, P. C. Yuen, and R. Chellappa, "Learning common and feature-specific patterns: A novel multiple-sparse-representation-based tracker," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 2022–2037, Apr. 2018.
[84] X. Lan, A. J. Ma, P. C. Yuen, and R. Chellappa, "Joint sparse representation and robust feature-level fusion for multi-cue visual tracking," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5826–5841, Dec. 2015.

**Chunjie Zhang** received the B.E. degree from the Nanjing University of Posts and Telecommunications, Jiangsu, China, in 2006, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, China, in 2011. He was an Engineer with the Henan Electric Power Research Institute from 2011 to 2012. He held a postdoctoral position with the School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing, China. He then joined the School of Computer and Control Engineering, University of Chinese Academy of Sciences, as an Assistant Professor. He joined the Institute of Automation, Chinese Academy of Sciences, as an Assistant Professor in 2017. He was then promoted to an Associate Professor in 2017. In 2019, he joined the Institute of Information Science, Beijing Jiaotong University, as a Professor.

He has published over 70 refereed journals and conference papers, such as the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, the IEEE TRANSACTIONS ON CYBERNETICS, the IEEE TRANSACTIONS ON MULTIMEDIA, and the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, CVPR, IJCAI, and ACM MM. His current research interests include image processing, cross-media analysis, machine learning, pattern recognition, and computer vision.

**Jian Cheng** received the B.S. and M.S. degrees from Wuhan University in 1998 and 2001, respectively, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, in 2004. From 2004 to 2006, he held a postdoctoral position with the Nokia Research Center, China. He has been with the National Laboratory of Pattern Recognition since 2006. His current research interests include machine learning methods and their applications for image processing and social network analysis.

**Qi Tian** (F'16) received the B.E. degree in electronic engineering from Tsinghua University in 1992, the M.S. degree in ECE from Drexel University in 1996, and the Ph.D. degree in ECE from the University of Illinois at Urbana–Champaign (UIUC) in 2002.

He was a tenured Associate Professor from 2008 to 2012 and a tenure-track Assistant Professor from 2002 to 2008. During 2008 to 2009, he took one-year Faculty Leave with the Media Computing Group, Microsoft Research Asia (MSRA), as a Lead Researcher. He is currently a Full Professor with the Department of Computer Science, The University of Texas at San Antonio (UTSA). He has coauthored a Best Paper in ACM ICMR 2015, a Best Paper in PCM 2013, a Best Paper in MMM 2013, a Best Paper in ACM ICIMCS 2012, and a Top 10. He has published over 390 refereed journals and conference papers. His research interests include multimedia information retrieval, computer vision, pattern recognition, and bioinformatics.

Dr. Tian research projects are funded by ARO, NSF, DHS, Google, FXPAL, NEC, SALSI, CIAS, Akiira Media Systems, HP, Blippar, and UTSA. He was a recipient of the 2010 Google Faculty Award, the 2010 ACM Service Award, the 2014 Research Achievement Awards from College of Science, UTSA, the 2016 UTSA Innovation Award, and the 2017 UTSA President's Distinguished Award for Research Achievement. He is an Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA (TMM), the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (TCSVT), *ACM Transactions on Multimedia Computing, Communications, and Applications* (TOMM), *Multimedia System Journal* (MMSJ), and in the Editorial Board of the *Journal of Multimedia* (JMM) and the *Journal of Machine Vision and Applications* (MVA). He is the Guest Editor of the IEEE TRANSACTIONS ON MULTIMEDIA and the *Journal of Computer Vision and Image Understanding*.