Automatic Face Naming by Learning Discriminative Affinity Matrices from Weakly Labeled Images

Shijie Xiao, Dong Xu, *Senior Member, IEEE*, Jianxin Wu, *Member, IEEE,*

***Abstract*—Given a collection of images, where each image contains several faces and is associated with a few names in the corresponding caption, the goal of face naming is to infer the correct name for each face. In this work, we propose two new methods to effectively solve this problem by learning two discriminative affinity matrices from these weakly labeled images. We first propose a new method called regularized low-rank representation (rLRR), by effectively utilizing weakly supervised information to learn a low-rank reconstruction coefficient matrix while exploring multiple subspace structures of the data. Specifically, by introducing a specially designed regularizer to the low-rank representation (LRR) method, we penalize the corresponding reconstruction coefficients related to the situations where a face is reconstructed by using face images from other subjects or by using itself. With the inferred reconstruction coefficient matrix, a discriminative affinity matrix can be obtained. Moreover, we also develop a new distance metric learning method called Ambiguously-supervised Structural Metric Learning (ASML) by using weakly supervised information to seek a discriminative distance metric. Hence, another discriminative affinity matrix can be obtained by using the similarity matrix (*i.e.*, the kernel matrix) based on the Mahalanobis distances of the data. Observing that these two affinity matrices contain complementary information, we further combine them to obtain a fused affinity matrix, based on which we develop a new iterative scheme to infer the name of each face. Comprehensive experiments demonstrate the effectiveness of our approach.**

***Index Terms*—Caption-based Face Naming, Low-rank Representation, Distance Metric Learning, Affinity Matrix.**

I. Introduction

In social networking websites (e.g., Facebook), photo sharing websites (e.g., Flickr) and news websites (e.g., BBC), an image which contains multiple faces can be associated with a caption specifying who is in the picture. For instance, multiple faces may appear in a news photo with a caption which briefly describes the news. Moreover, in TV serials, movies and news videos, the faces may also appear in a video clip with scripts. In literature, a few methods were developed for the face naming problem (see Section II for more details).

In this paper, we focus on automatically annotating faces in images based on the ambiguous supervision from the associated captions. Fig. 1 gives an illustration of the face-naming problem. Some pre-processing steps need to be conducted before performing face naming. Specifically, faces in the images are automatically detected by using face detectors such as [1], and names in the captions are automatically extracted by using a name entity detector. Here, the list of names appearing in a caption is denoted as the *candidate name set*. Even after successfully performing these pre-processing steps, automatic face naming is still a challenging task. The faces from the same subject may have different appearances because of the variations in poses, illuminations and expressions. Moreover, the candidate name set may be noisy and incomplete, so a name may be mentioned in the caption but the corresponding face may not appear in the image, and the correct name for a face in the image may not appear in the corresponding caption. Each detected face (including falsely detected ones) in an image can only be annotated by using one of the names in the candidate name set or as "null", which indicates the groundtruth name does not appear in the caption.

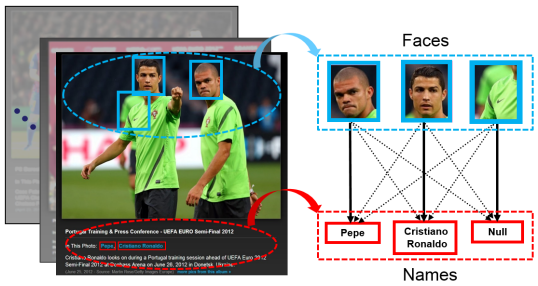


Fig. 1. An illustration of the face-naming task, in which we aim to infer which name matches which face, based on the images and corresponding captions. The solid arrows between faces and names indicate the groundtruth face-name pairs and the dashed ones represent the incorrect face-name pairs, where "null" means the groundtruth name of a face does not appear in the candidate name set.

In this paper, we propose a new scheme for automatic face naming with caption-based supervision. Specifically, we develop two methods to respectively obtain two discriminative affinity matrices by learning from weakly labeled images. The two affinity matrices are further fused to generate one fused affinity matrix, based on which an iterative scheme is developed for automatic face naming.

To obtain the first affinity matrix, we propose a new method called regularized low-rank representation (rLRR) by incorporating weakly supervised information into the low-rank representation (LRR) method, so that the affinity matrix can be obtained from the resultant reconstruction coefficient matrix. To effectively infer the correspondences between the faces based on visual features and the names in the candidate name sets, we exploit the *subspace structures among faces* based on the following assumption: the faces from the same subject/name lie in the same subspace and the subspaces are linearly independent. In [2], Liu et al. showed that such

S. Xiao and D. Xu are with the School of Computer Engineering, Nanyang Technological University, Singapore e-mail: {XIAO0050, DongXu}@ntu.edu.sg

J. Wu is with the National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China e-mail: wujx2001@nju.edu.cn

subspace structures can be effectively recovered by using LRR, when the subspaces are independent and the data sampling rate is sufficient. They also showed that the mined subspace information is encoded in the reconstruction coefficient matrix which is block-diagonal in the ideal case. As an intuitive motivation, we implement LRR on a synthetic dataset and the resultant reconstruction coefficient matrix is shown in Fig. 2(b) in Section V-A (More details can be found in Section V-A and Section V-C). This near block-diagonal matrix validates our assumption on the *subspace structures among faces*. Specifically, the reconstruction coefficients between one face and faces from the same subject are generally larger than others, indicating that the faces from the same subject tend to lie in the same subspace [2]. However, due to the significant variances in poses, illuminations and expressions of in-the-wild faces, the appearances of faces from different subjects may be even more similar when compared with those from the same subject. Consequently, as shown in Fig. 2(b), the faces may also be reconstructed by using faces from other subjects. In this work, we show that the candidate names from the captions can provide important supervision information to better discover the subspace structures.

In Section III-C2, we first propose a method called regularized LRR (rLRR) by introducing a new regularizer that incorporates caption-based weak supervision into the objective of LRR, in which we penalize the reconstruction coefficients when reconstructing the faces using those from different subjects. Based on the inferred reconstruction coefficient matrix, we can compute an affinity matrix which measures the similarity values between every pair of faces. Compared with the one in Fig. 2(b), the reconstruction coefficient matrix from our rLRR exhibits more obvious block-diagonal structure in Fig. 2(c), which indicates that a better reconstruction matrix can be obtained by using the proposed regularizer.

Moreover, we use the similarity matrix (*i.e.*, the kernel matrix) based on the Mahalanobis distances between the faces as another affinity matrix. Specifically, in Section III-D, we develop a new distance metric learning method called Ambiguously-supervised Structural Metric Learning (ASML) to learn a discriminative Mahalanobis distance metric based on weak supervision information. In ASML, we consider the constraints for the label matrix of the faces in each image by using the feasible label set, and we further define the image to assignment (I2A) distance which measures the incompatibility between a label matrix and the faces from each image based on the distance metric. Hence, ASML learns a Mahalanobis distance metric that encourages the I2A distance based on a selected feasible label matrix, which approximates the groundtruth one, to be smaller than the I2A distances based on infeasible label matrices to some extent.

Since rLRR and ASML explore the weak supervision in different ways and they are both effective as shown in our experimental results in Section V, the two corresponding affinity matrices are expected to contain complementary and discriminative information for face naming. Therefore, to further improve the performance, we combine the two affinity matrices to obtain a fused affinity matrix, which is used for face naming. Accordingly, we refer to this method as rLRRml. Based on the fused affinity matrix, we additionally propose a new iterative method, by formulating the face naming problem as an integer programming problem with linear constraints, where the constraints are related to the feasible label set of each image.

Our main contributions are summarized as follows:

- Based on the caption-based weak supervision, we propose a new method rLRR by introducing a new regularizer into LRR and we can calculate the first affinity matrix by using the resultant reconstruction coefficient matrix (see Section III-C).
- We also propose a new distance metric learning approach ASML to learn a discriminative distance metric by effectively coping with the ambiguous labels of faces. The similarity matrix (*i.e.*, the kernel matrix) based on the Mahalanobis distances between all faces is used as the second affinity matrix (see Section III-D).
- With the fused affinity matrix by combining the two affinity matrices from rLRR and ASML, we propose an efficient scheme to infer the names of faces (see Section IV).
- Comprehensive experiments are conducted on one synthetic dataset and two real-world datasets, and the results demonstrate the effectiveness of our approaches (see Section V).

II. Related Work

Recently, there is an increasing research interest in developing automatic techniques for face naming in images [3], [4], [5], [6], [7], [8], [9] as well as in videos [10], [11], [12], [13]. To tag faces in news photos, Berg et al. [3] proposed to cluster the faces in the news images. Ozkan and Duygulu [4] developed a graph-based method by constructing the similarity graph of faces and finding the densest component. In [6], Guillaumin et al. proposed the multiple instance logistic discriminant metric learning (MildML) method. In [7], Luo and Orabona proposed a Structural SVM-like algorithm called Maximum Margin Set (MMS) to solve the face naming problem. Recently, in [9], Zeng et al. proposed the Low-Rank SVM (LR-SVM) approach to deal with this problem, based on the assumption that the feature matrix formed by faces from the same subject is low-rank. In the following, we compare our proposed approaches with several related existing methods.

Our rLRR method is related to LRR [2] and LR-SVM [9]. LRR is an unsupervised approach for exploring multiple subspace structures of data. In contrast to LRR, our rLRR utilizes the weak supervision from image captions and also considers the image-level constraints when solving the weakly supervised face naming problem. Moreover, our rLRR differs from LR-SVM [9] in the following two aspects: 1) To utilize the weak supervision, LR-SVM considers weak supervision information in the partial permutation matrices, while rLRR uses our proposed regularizer to penalize the corresponding reconstruction coefficients. 2) LR-SVM is based on Robust PCA [14] [15] which is not a reconstruction based method, while our rLRR is related to the reconstruction based approach LRR.

Moreover, our ASML is related to the traditional metric learning works, such as Large Margin Nearest Neighbors (LMNN) [16], Frobmetric [17] and Metric Learning to Rank (MLR) [18]. LMNN and Frobmetric are based on accurate supervision without ambiguity (*i.e.*, the triplets of training samples are explicitly given), and they both use the hinge loss in their formulation. In contrast, our ASML is based on the ambiguous supervision, and we use a max margin loss to handle the ambiguity of the structural output, by enforcing the distance based on the best label assignment matrix in the feasible label set to be larger than the distance based on the best label assignment matrix in the infeasible label set by a margin. Although a similar loss that deals with structural output is also used in MLR, it is used to model the ranking orders of training samples and there is no uncertainty regarding supervision information in MLR (*i.e.*, the groundtruth ordering for each query is given).

Our ASML is also related to two recently proposed approaches for the face naming problem by using weak supervision, multiple instance logistic discriminant metric learning (MildML) [6] and maximum margin set (MMS) [7]. MildML follows the multi-instance learning assumption, which assumes that each image should contain a face corresponding to each name in the caption. However, it may not hold for our face naming problem as the captions are not accurate. In contrast, our ASML employs a max margin loss to handle the structural output without using such an assumption. While MMS also uses a max margin loss to handle the structural output, MMS aims to learn the classifiers and it was designed for the classification problem. Our ASML learns a distance metric, which can be readily used to generate an affinity matrix and combined with the affinity matrix from our rLRR method to further improve the face naming performance.

Finally, we compare our face naming problem with multi-instance learning (MIL) [19] and multi-instance multi-label learning (MIML) [20] and the face naming problem in [21]. In the existing MIL and MIML works, a few instances are grouped into bags, in which the bag labels are assumed to be correct. Moreover, the common assumption in MIL is that one positive bag contains at least one positive instance. A straightforward way to apply MIL and MIML methods for solving the face naming problem is to treat each image as a bag, the faces in the image as the instances and the names in the caption as the bag labels. However, the bag labels (based on candidate name sets) may be even incorrect in our problem because the faces corresponding to the mentioned names in the caption may be absent in the image. Besides, one common assumption in face naming is that any two faces in the same image cannot be annotated by using the same name, which indicates that each positive bag contains no more than one positive instance rather than at least one positive instance. Moreover, in [21], each image only contains one face. In contrast, we may have multiple faces in one image which are related to a set of candidate names in our problem.

III. Learning Discriminative Affinity Matrices for Automatic Face Naming

In this section, we propose a new approach for automatic face naming with caption-based supervision. In Section III-A and Section III-B, we formally introduce the problem and definitions, followed by the introduction of our proposed approach. Specifically, we learn two discriminative affinity matrices by effectively utilizing the ambiguous labels, and perform face naming based on the fused affinity matrix. In Section III-C (resp., Section III-D), we introduce our proposed rLRR (resp., ASML) approach to obtain one of the two affinity matrices.

In the remainder of this paper, we use lowercase/uppercase letters in boldface to denote a vector/matrix (*e.g.*, $\mathbf{a}$ denotes a vector and $\mathbf{A}$ denotes a matrix). The corresponding non-bold letter with a subscript denotes the entry in a vector/matrix (*e.g.*, a_i denotes the i-th entry of the vector $\mathbf{a}$, and $A_{i,j}$ denotes an entry at the i-th row and j-th column of the matrix $\mathbf{A}$). The superscript $'$ denotes the transpose of a vector or a matrix. We define $\mathbf{I}_n$ as the $n \times n$ identity matrix, and $\mathbf{0}_n, \mathbf{1}_n \in \mathbb{R}^n$ as the $n \times 1$ column vectors of all zeros and all ones, respectively. For simplicity, we also use $\mathbf{I}$, $\mathbf{0}$ and $\mathbf{1}$ instead of $\mathbf{I}_n$, $\mathbf{0}_n$ and $\mathbf{1}_n$ when the dimensionality is obvious. Moreover, we use $\mathbf{A} \circ \mathbf{B}$ (resp., $\mathbf{a} \circ \mathbf{b}$) to denote the element-wise product between two matrices $\mathbf{A}$ and $\mathbf{B}$ (resp., two vectors $\mathbf{a}$ and $\mathbf{b}$). $\mathrm{tr}(\mathbf{A})$ denotes the trace of $\mathbf{A}$ (*i.e.*, $\mathrm{tr}(\mathbf{A}) = \sum_i A_{i,i}$), and $\langle \mathbf{A}, \mathbf{B} \rangle$ denotes the inner product of two matrices (*i.e.*, $\langle \mathbf{A}, \mathbf{B} \rangle = \mathrm{tr}(\mathbf{A}'\mathbf{B})$). The inequality $\mathbf{a} \leq \mathbf{b}$ means that $a_i \leq b_i \ \forall \ i = 1, \ldots, n$, and $\mathbf{A} \succeq 0$ means that $\mathbf{A}$ is a positive semi-definite (PSD) matrix. $\|\mathbf{A}\|_F = \sqrt{\sum_{i,j} A_{i,j}^2}$ denotes the Frobenious norm of a matrix $\mathbf{A}$. $\|\mathbf{A}\|_\infty$ denotes the largest absolute value of all elements in $\mathbf{A}$.

A. Problem Statement

Given a collection of images, each of which contains several faces and is associated with multiple names, our goal is to annotate each face in these images with these names.

Formally, let us assume we have m images, each of which contains n_i faces and r_i names, $i = 1, \ldots, m$. Let $\mathbf{x} \in \mathbb{R}^d$ denote a face, where d is the feature dimension. Moreover, let $q \in \{1, \ldots, p\}$ denote a name, where p is the total number of names in all the captions. Then, each image can be represented as a pair $(\mathbf{X}^i, \mathcal{N}^i)$, where $\mathbf{X}^i = [\mathbf{x}_1^i, \ldots, \mathbf{x}_{n_i}^i] \in \mathbb{R}^{d \times n_i}$ is the data matrix for faces in the i-th image with each $\mathbf{x}_f^i$ being the f-th face in this image ($f = 1, \ldots, n_i$), and $\mathcal{N}^i = \{q_1^i, \ldots, q_{r_i}^i\}$ is the corresponding set of candidate names with each $q_j^i \in \{1, \ldots, p\}$ being the j-th name ($j = 1, \ldots, r_i$). Moreover, let $\mathbf{X} = [\mathbf{X}^1, \ldots, \mathbf{X}^m] \in \mathbb{R}^{d \times n}$ denote the data matrix of the faces from all m images, where $n = \sum_{i=1}^m n_i$.

By defining a binary label matrix $\mathbf{Y} = [\mathbf{Y}^1, \ldots, \mathbf{Y}^m] \in \{0, 1\}^{(p+1) \times n}$ with each $\mathbf{Y}^i \in \{0, 1\}^{(p+1) \times n_i}$ being the label matrix for each image $\mathbf{X}^i$, then the task is to infer the label matrix $\mathbf{Y}$ based on the candidate name sets $\{\mathcal{N}^i|_{i=1}^m\}$. Considering the situation where the groundtruth name of a face does not appear in the associated candidate name set $\mathcal{N}^i$, we use the $(p+1)$-th name to denote the "null" class, so that the face should be assigned to the $(p+1)$-th name in this situation.

Moreover, the label matrix $\mathbf{Y}^i$ for each image should satisfy the following three image-level constraints [9]:

- **Feasibility**: the faces in the i-th image should be annotated using the names from the set $\tilde{\mathcal{N}}^i = \mathcal{N}^i \cup \{\ (p+1)\ \}$, *i.e.*, $Y^i_{j,f} = 0,\ \forall f = 1, \ldots, n_i$ and $j \notin \tilde{\mathcal{N}}^i$.
- **Non-redundancy**: each face in the i-th image should be annotated by using exactly one name from $\tilde{\mathcal{N}}^i$, *i.e.*, $\sum_j Y^i_{j,f} = 1,\ \forall f = 1, \ldots, n_i$.
- **Uniqueness**: two faces in the same image cannot be annotated with the same name except the $(p+1)$-th name (*i.e.*, the "null" class), *i.e.*, $\sum_{f=1}^{n_i} Y^i_{j,f} \leq 1, \forall j = 1, \ldots, p$.

B. Face Naming using a Discriminative Affinity Matrix

Firstly, based on the image-level constraints, we define the feasible set of $\mathbf{Y}^i$ for the i-th image as follows:

$$\mathcal{Y}^i = \left\{ \mathbf{Y}^i \in \{0,1\}^{(p+1)\times n_i} \left| \begin{array}{l} \mathbf{1}'_{(p+1)}(\mathbf{Y}^i \circ \mathbf{T}^i)\mathbf{1}_{n_i} = 0, \\ \mathbf{1}'_{(p+1)}\mathbf{Y}^i = \mathbf{1}'_{n_i}, \\ \mathbf{Y}^i\mathbf{1}_{n_i} \leq [\mathbf{1}'_p, n_i]' \end{array} \right. \right\}, \quad (1)$$

where $\mathbf{T}^i \in \{0,1\}^{(p+1)\times n_i}$ is a matrix in which the rows related to the indices of the names in $\tilde{\mathcal{N}}^i$ are all zeros and the other rows are all ones.

Accordingly, the feasible set for the label matrix on all images can be represented as

$$\mathcal{Y} = \{\mathbf{Y} = [\mathbf{Y}^1, \ldots, \mathbf{Y}^m] \mid \mathbf{Y}^i \in \mathcal{Y}^i, \forall i = 1, \ldots, m\}.$$

Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be an affinity matrix, which satisfies that $\mathbf{A} = \mathbf{A}'$ and $A_{i,j} \geq 0, \forall i, j$. Each $A_{i,j}$ describes the pairwise affinity/similarity between the i-th face and the j-th face [2]. We aim to learn a proper $\mathbf{A}$, such that $A_{i,j}$ is large if and only if the i-th face and the j-th face share the same groundtruth name. Then, one can solve the face naming problem based on the obtained affinity matrix $\mathbf{A}$. To infer the names of faces, we aim to solve the following problem:

$$\max_{\mathbf{Y}\in\mathcal{Y}} \sum_{c=1}^{p} \frac{\mathbf{y}'_c \mathbf{A} \mathbf{y}_c}{\mathbf{1}'\mathbf{y}_c} \quad s.t.\ \mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{(p+1)}]', \quad (2)$$

where $\mathbf{y}_c \in \{0,1\}^n$ corresponds to the c-th row in $\mathbf{Y}$. The intuitive idea is that we cluster the faces with the same inferred label as one group, and we maximize the sum of the average affinities for each group. The solution of this problem will be introduced in Sec IV. According to (2), a good affinity matrix is crucial in our proposed face naming scheme, because it directly determines the face naming performance.

In this work, we consider two methods to obtain two affinity matrices, respectively. Specifically, to obtain the first affinity matrix, we propose the regularized low rank representation (rLRR) method to learn the low-rank reconstruction coefficient matrix while considering the weak supervision. To obtain the second affinity matrix, we propose the Ambiguously-supervised Structural Metric Learning (ASML) method to learn the discriminative distance metric by effectively using weak supervised information.

C. Learning Discriminative Affinity Matrix with Regularized Low Rank Representation (rLRR)

We firstly give a brief review of LRR, and then present the proposed method which introduces a discriminative regularizer into the objective of LRR.

1) A Brief Review of LRR: LRR [2] was originally proposed to solve the *subspace clustering* problem, which aims to explore the subspace structure in the given data $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$. Based on the assumption that the subspaces are linearly independent, LRR [2] seeks a reconstruction matrix $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_n] \in \mathbb{R}^{n\times n}$, where each $\mathbf{w}_i$ denotes the representation of $\mathbf{x}_i$ using $\mathbf{X}$ (*i.e.*, the data matrix itself) as the "dictionary". Since $\mathbf{X}$ is used as the "dictionary" to reconstruct itself, the optimal solution $\mathbf{W}^*$ by solving (3) encodes the pairwise affinities between the data samples. As discussed in Theorem 3.1 of [2], in the noise-free case, $\mathbf{W}^*$ should be ideally block-diagonal, where $W^*_{i,j} \neq 0$ indicates the i-th sample and the j-th sample are in the same subspace.

Specifically, the optimization problem of LRR is as follows:

$$\min_{\mathbf{W},\mathbf{E}} \ \|\mathbf{W}\|_* + \lambda\|\mathbf{E}\|_{2,1} \quad s.t.\ \mathbf{X} = \mathbf{X}\mathbf{W} + \mathbf{E}, \quad (3)$$

where $\lambda > 0$ is a tradeoff parameter, $\mathbf{E} \in \mathbb{R}^{d\times n}$ is the reconstruction error, the nuclear norm $\|\mathbf{W}\|_*$ (i.e., the sum of all singular values of $\mathbf{W}$) is adopted to replace $rank(\mathbf{W})$ as commonly used in the rank minimization problems, $\|\mathbf{E}\|_{2,1} = \sum_{j=1}^{n} \sqrt{\sum_{i=1}^{d} (E_{i,j})^2}$ is a regularizer to encourage the reconstruction error $\mathbf{E}$ to be column-wise sparse. As mentioned in [2], compared with the sparse representation (SR) method that encourages the sparsity by using the ℓ_1 norm, "LRR is better at handling the global structures and correcting the corruptions in data automatically." Mathematically, the nuclear norm is non-separable w.r.t. the columns, which is different from the ℓ_1 norm. This good property of the nulcear norm is helpful for grasping the global structure and making the model more robust. The toy experiments in [2] (see Section 4.1 in [2]) also clearly demonstrate that LRR outperforms SR (which adopts the ℓ_1 norm). Similarly, in many real-world applications such as face clustering, LRR usually achieves better results than the Sparse Subspace Clustering (SSC) [22] method (see [2], [23] and [24] for more details).

2) LRR with a Discriminative Regularization: In (3), LRR learns the coefficient matrix $\mathbf{W}$ in an unsupervised way. In our face naming problem, although the names from captions are ambiguous and noisy, they still provide us with the weak supervision information which is useful for improving the performance of face naming. For example, if two faces do not share any common name in their related candidate name sets, it is unlikely that they are from the same subject, so we should enforce the corresponding entries in $\mathbf{W}$ to be zeros or close to zeros.

Based on this motivation, we introduce a new regularization term $\|\mathbf{W} \circ \mathbf{H}\|_F^2$ by incorporating the weak supervised information, where $\mathbf{H} \in \{0,1\}^{n\times n}$ is defined based on the candidate name sets $\{\mathcal{N}^i|_{i=1}^m\}$. Specifically, the entry $H_{i,j} = 0$ if the following two conditions are both satisfied: (1) the i-th face and the j-th face share at least one common name in the corresponding candidate name sets; (2) $i \neq j$. Otherwise,

$H_{i,j} = 1$. In this way, we penalize the non-zero entries in $\mathbf{W}$ where the corresponding pair of faces do not share any common names in their candidate name sets, and meanwhile we penalize the entries corresponding to the situations where a face is reconstructed by itself.

As a result, with weak supervision information encoded in $\mathbf{H}$, the resultant coefficient matrix $\mathbf{W}$ is expected to be more discriminative. By introducing the new regularizer $\|\mathbf{W}\circ\mathbf{H}\|_F^2$ into LRR, we arrive at a new optimization problem as follows:

$$\begin{array}{ll}\min\limits_{\mathbf{W},\mathbf{E}} & \|\mathbf{W}\|_* + \lambda\|\mathbf{E}\|_{2,1} + \frac{\gamma}{2}\|\mathbf{W}\circ\mathbf{H}\|_F^2 \\ s.t. & \mathbf{X} = \mathbf{X}\mathbf{W} + \mathbf{E},\end{array} \quad (4)$$

where $\gamma \geq 0$ is a parameter to balance the new regularizer with the other terms. We refer to the above problem as regularized LRR, or rLRR in short. The rLRR problem in (4) can reduce to the LRR problem in (3) by setting the parameter γ to zero. The visual results for the resultant $\mathbf{W}$ from rLRR and the one from LRR can be found in Fig. 2 (see Section V-A).

Once we obtain the optimum solution $\mathbf{W}^*$ after solving (5), the affinity matrix $\mathbf{A}_W$ can be computed as $\mathbf{A}_W = \frac{1}{2}(\mathbf{W}^* + \mathbf{W}^{*\prime})$, similarly as in [2], $\mathbf{A}_W$ is further normalized to be within the range of $[0, 1]$.

3) Optimization: The optimization problem in (4) can be solved similarly as in LRR [2]. Specifically, we introduce an intermediate variable $\mathbf{J}$ to convert the problem in (4) into the following equivalent problem:

$$\begin{array}{ll}\min\limits_{\mathbf{W},\mathbf{E},\mathbf{J}} & \|\mathbf{J}\|_* + \lambda\|\mathbf{E}\|_{2,1} + \frac{\gamma}{2}\|\mathbf{W}\circ\mathbf{H}\|_F^2 \\ s.t. & \mathbf{X} = \mathbf{X}\mathbf{W} + \mathbf{E},\ \ \mathbf{W} = \mathbf{J}.\end{array} \quad (5)$$

By using the Augmented Lagrange Multiplier (ALM) method, we consider the following augmented Lagrangian function:

$$\begin{aligned}\mathcal{L} =& \|\mathbf{J}\|_* + \lambda\|\mathbf{E}\|_{2,1} + \frac{\gamma}{2}\|\mathbf{W}\circ\mathbf{H}\|_F^2 + \langle\mathbf{U}, \mathbf{X} - \mathbf{X}\mathbf{W} - \mathbf{E}\rangle \\ &+ \langle\mathbf{V}, \mathbf{W} - \mathbf{J}\rangle + \frac{\rho}{2}\left(\|\mathbf{X} - \mathbf{X}\mathbf{W} - \mathbf{E}\|_F^2 + \|\mathbf{W} - \mathbf{J}\|_F^2\right),\end{aligned} \quad (6)$$

where $\mathbf{U} \in \mathbb{R}^{d\times n}$ and $\mathbf{V} \in \mathbb{R}^{n\times n}$ are Lagrange multipliers, and ρ is a positive penalty parameter. Following [2], we solve this problem by using inexact ALM [25], which iteratively update the variables, the Lagrange multipliers and the penalty parameter until convergence is achieved. Specifically, we set $\mathbf{W}_0 = \frac{1}{n}(\mathbf{1}_n\mathbf{1}_n' - \mathbf{H})$, $\mathbf{E}_0 = \mathbf{X} - \mathbf{X}\mathbf{W}_0$, $\mathbf{J}_0 = \mathbf{W}_0$, and we set $\mathbf{U}_0, \mathbf{V}_0$ as zero matrices. Then, at the t-th iteration, the following steps are performed until convergence is achieved:

- Fix the others and update $\mathbf{J}_{t+1}$ by

$$\min_{\mathbf{J}_{t+1}} \|\mathbf{J}_{t+1}\|_* + \frac{\rho_t}{2}\left\|\mathbf{J}_{t+1} - \left(\mathbf{W}_t + \frac{\mathbf{V}_t}{\rho_t}\right)\right\|_F^2,$$

  which can be solved in closed form by using the Singular Value Thresholding (SVT) method in [26].

- Fix the others and update $\mathbf{W}_{t+1}$ by

$$\begin{aligned}\min_{\mathbf{W}_{t+1}}\ & \frac{\gamma}{2}\|\mathbf{W}_{t+1}\circ\mathbf{H}\|_F^2 + \langle\mathbf{U}_t, \mathbf{X} - \mathbf{X}\mathbf{W}_{t+1} - \mathbf{E}_t\rangle \\ & + \langle\mathbf{V}_t, \mathbf{W}_{t+1} - \mathbf{J}_{t+1}\rangle + \frac{\rho_t}{2}\|\mathbf{X} - \mathbf{X}\mathbf{W}_{t+1} - \mathbf{E}_t\|_F^2 \\ & + \frac{\rho_t}{2}\|\mathbf{W}_{t+1} - \mathbf{J}_{t+1}\|_F^2.\end{aligned} \quad (7)$$

  Due to the new regularizer $\|\mathbf{W}\circ\mathbf{H}\|_F^2$, this problem cannot be solved as in [2] by using pre-computed SVD. We use the gradient descent method to efficiently solve (7), where the gradient w.r.t. $\mathbf{W}_{t+1}$ is

$$\begin{aligned}&\gamma(\mathbf{H}\circ\mathbf{H})\circ\mathbf{W}_{t+1} + \rho_t(\mathbf{X}'\mathbf{X} + \mathbf{I})\mathbf{W}_{t+1} \\ &+ \mathbf{V}_t - \rho_t\mathbf{J}_{t+1} - \mathbf{X}'(\rho_t(\mathbf{X} - \mathbf{E}_t) + \mathbf{U}_t).\end{aligned}$$

- Fix the others and update $\mathbf{E}_{t+1}$ by

$$\min_{\mathbf{E}_{t+1}} \frac{\lambda}{\rho_t}\|\mathbf{E}_{t+1}\|_{2,1} + \frac{1}{2}\left\|\mathbf{E}_{t+1} - (\mathbf{X} - \mathbf{X}\mathbf{W}_{t+1} + \frac{\mathbf{U}_t}{\rho_t})\right\|_F^2,$$

  which can be solved in closed form based on Lemma 4.1 in [27].

- Update $\mathbf{U}_{t+1}$ and $\mathbf{V}_{t+1}$ by respectively using

$$\begin{aligned}\mathbf{U}_{t+1} &= \mathbf{U}_t + \rho_t(\mathbf{X} - \mathbf{X}\mathbf{W}_{t+1} - \mathbf{E}_{t+1}), \\ \mathbf{V}_{t+1} &= \mathbf{V}_t + \rho_t(\mathbf{W}_{t+1} - \mathbf{J}_{t+1}).\end{aligned}$$

- Update ρ_{t+1} by using

$$\rho_{t+1} = \min(\rho_t(1 + \Delta\rho), \rho_{max}),$$

  where $\Delta\rho$ and ρ_{max} are the constant parameters.

- The iterative algorithm stops if the two convergence conditions are both satisfied:

$$\begin{aligned}\|\mathbf{X} - \mathbf{X}\mathbf{W}_{t+1} - \mathbf{E}_{t+1}\|_\infty &\leq \epsilon, \\ \|\mathbf{W}_{t+1} - \mathbf{J}_{t+1}\|_\infty &\leq \epsilon,\end{aligned}$$

  where ϵ is a small constant parameter.

D. Learning Discriminative Affinity Matrix by Ambiguously-supervised Structural Metric Learning (ASML)

Besides obtaining the affinity matrix from the coefficient matrix $\mathbf{W}^*$ from rLRR (or LRR), we believe the similarity matrix (*i.e.*, the kernel matrix) among the faces is also an appropriate choice for the affinity matrix. Instead of straightforwardly using the Euclidean distances, we seek a discriminative Mahalanobis distance metric $\mathbf{M}$ so that Mahalanobis distances can be calculated based on the learnt metric, and the similarity matrix can be obtained based on the Mahalanobis distances. In the following, we first briefly review the Large Margin Nearest Neighbour (LMNN) method which deals with fully-supervised problems with the groungtruth labels of samples provided, and then introduce our proposed Ambiguously-supervised Structural Metric Learning (ASML) method which extends LMNN for face naming from weakly labeled images.

1) A Brief Review of LMNN: Most existing metric learning methods deal with the supervised learning problems [16] [28] where the groundtruth labels are given. Weinberger and Saul [16] proposed the Large Margin Nearest Neighbour (LMNN) method to learn a distance metric $\mathbf{M}$ which encourages the squared Mahalanobis distances between each data and its target neighbours (*e.g.*, the k nearest neighbours) to be smaller than those between this data and the data from other classes. Let $\{(\mathbf{x}_i, y_i)|_{i=1}^n\}$ be the n labeled samples, where $\mathbf{x}_i \in \mathbb{R}^d$ denotes the i-th sample, with d being the feature dimension, and $y_i \in \{1, \ldots, z\}$ denotes the label of this sample, with

z being the total number of classes. $\eta_{i,j} \in \{0,1\}$ indicates whether $\mathbf{x}_j$ is a target neighbour of $\mathbf{x}_i$, namely, $\eta_{i,j} = 1$ if $\mathbf{x}_j$ is a target neighbour of $\mathbf{x}_i$, and $\eta_{i,j} = 0$ otherwise, $\forall i,j \in \{1,\ldots,n\}$. $\nu_{i,l} \in \{0,1\}$ indicates whether $\mathbf{x}_l$ and $\mathbf{x}_i$ are from different classes, namely, $\nu_{i,l} = 1$ if $y_l \neq y_i$, and $\nu_{i,l} = 0$ otherwise, $\forall i,l \in \{1,\ldots,n\}$. The squared Mahalanobis distance between two samples $\mathbf{x}_i$ and $\mathbf{x}_j$ is defined as

$$d^2_{\mathbf{M}}(\mathbf{x}_i,\mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)'\mathbf{M}(\mathbf{x}_i - \mathbf{x}_j).$$

LMNN minimizes the following optimization problem:

$$\begin{aligned} \min_{\mathbf{M}\succeq 0} \quad & \sum_{(i,j):\eta_{i,j}=1} d^2_{\mathbf{M}}(\mathbf{x}_i,\mathbf{x}_j) + \mu \sum_{(i,j,l)\in\mathcal{S}} \xi_{i,j,l} \qquad (8)\\ s.t. \quad & d^2_{\mathbf{M}}(\mathbf{x}_i,\mathbf{x}_l) - d^2_{\mathbf{M}}(\mathbf{x}_i,\mathbf{x}_j) \geq 1 - \xi_{i,j,l},\ \forall (i,j,l)\in\mathcal{S},\\ & \xi_{i,j,l} \geq 0,\ \forall (i,j,l)\in\mathcal{S}, \end{aligned}$$

where μ is a tradeoff parameter, $\xi_{i,j,l}$ is a slack variable, and $\mathcal{S} = \{(i,j,l)|\eta_{i,j} = 1, \nu_{i,l} = 1, \forall i,j,l \in \{1,\ldots,n\}\}$. So, $d^2_{\mathbf{M}}(\mathbf{x}_i,\mathbf{x}_j)$ is the squared Mahalanobis distance between $\mathbf{x}_i$ and its target neighbour $\mathbf{x}_j$, and $d^2_{\mathbf{M}}(\mathbf{x}_i,\mathbf{x}_l)$ is the squared Mahalanobis distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ that belong to different classes. The difference between $d^2_{\mathbf{M}}(\mathbf{x}_i,\mathbf{x}_l)$ and $d^2_{\mathbf{M}}(\mathbf{x}_i,\mathbf{x}_j)$ is expected to be no less than 1 in the ideal case. The introduction of the slack variable $\xi_{i,j,l}$ can also tolerate the cases when $d^2_{\mathbf{M}}(\mathbf{x}_i,\mathbf{x}_l) - d^2_{\mathbf{M}}(\mathbf{x}_i,\mathbf{x}_j)$ is slightly smaller than 1, which is similar to the one in soft margin SVM for tolerating the classification error. The LMNN problem in (8) can be equivalently rewritten as the following optimization problem:

$$\min_{\mathbf{M}\succeq 0} \sum_{(i,j):\eta_{i,j}=1} d^2_{\mathbf{M}}(\mathbf{x}_i,\mathbf{x}_j) + \mu \sum_{(i,j,l)\in\mathcal{S}} |1 - d^2_{\mathbf{M}}(\mathbf{x}_i,\mathbf{x}_l) + d^2_{\mathbf{M}}(\mathbf{x}_i,\mathbf{x}_j)|_+,$$

with $|\cdot|_+$ being the truncation function, *i.e.*, $|x|_+ = \max(0,x)$.

2) Ambiguously-supervised Structural Metric Learning: In the face naming problem, the groundtruth names of the faces are not available, so LMNN cannot be applied to solve the problem. Fortunately, weak supervision information is available in the captions along with each image, hence we propose a new distance metric learning method called ASML to utilize such weakly supervised information.

Recall that we should consider the image-level constraints when inferring the names of faces in the same image. So we design the losses with respect to each image, by considering the image-level constraints in the feasible label sets $\{\mathcal{Y}^i|_{i=1}^m\}$ defined in (1).

Let us take the i-th image for example. The faces in the i-th image are $\{\mathbf{x}_f^i|_{f=1}^{n_i}\}$. Let $\mathbf{Y}_*^i$ be the groundtruth label matrix for the faces in the i-th image, which is in the feasible label sets $\mathcal{Y}^i$. Let $\bar{\mathbf{Y}}^i$ be an infeasible label matrix for the faces in the i-th image, which is contained in the infeasible label set $\bar{\mathcal{Y}}^i$. Note the infeasible label set $\bar{\mathcal{Y}}^i$ is the set of label matrices which are excluded in $\mathcal{Y}^i$ and meanwhile satisfy the non-redundancy constraint, namely

$$\bar{\mathcal{Y}}^i = \left\{ \bar{\mathbf{Y}}^i \in \{0,1\}^{(p+1)\times n_i} \;\middle|\; \begin{array}{l} \bar{\mathbf{Y}}^i \notin \mathcal{Y}^i, \\ \mathbf{1}'_{(p+1)}\bar{\mathbf{Y}}^i = \mathbf{1}'_{n_i} \end{array} \right\}.$$

Assume that the face $\mathbf{x}_f^i$ is labeled as the name q according to a label matrix, we define face to name (F2N) distance $D_{F2N}(\mathbf{x}_f^i, q, \mathbf{M})$ to measure the disagreement between the face $\mathbf{x}_f^i$ and the name q. Specifically, $D_{F2N}(\mathbf{x}_f^i, q, \mathbf{M})$ is defined as follows:

$$D_{F2N}(\mathbf{x}_f^i, q, \mathbf{M}) = \frac{1}{|\mathcal{X}_q|} \sum_{\tilde{\mathbf{x}}\in\mathcal{X}_q} d^2_{\mathbf{M}}(\mathbf{x}_f^i, \tilde{\mathbf{x}}),$$

where $d^2_{\mathbf{M}}(\mathbf{x}_f^i, \tilde{\mathbf{x}})$ is the squared Mahalanobis distance between $\mathbf{x}_f^i$ and $\tilde{\mathbf{x}}$, $\mathcal{X}_q$ is the set of all the faces from the images with each image associated with the name q, and $|\mathcal{X}_q|$ is the cardinality of $\mathcal{X}_q$. Intuitively, $D_{F2N}(\mathbf{x}, q, \mathbf{M})$ should be small if q is the groundtruth name of the face $\mathbf{x}$, and $D_{F2N}(\mathbf{x}, q, \mathbf{M})$ should be large, otherwise. Recall that in LMNN, we expect $d^2_{\mathbf{M}}(\mathbf{x}_i,\mathbf{x}_j)$ (*i.e.*, the squared Mahalanobis distance between $\mathbf{x}_i$ and its target neighbour $\mathbf{x}_j$) to be somehow smaller than $d^2_{\mathbf{M}}(\mathbf{x}_i,\mathbf{x}_l)$ (*i.e.*, the squared Mahalanobis distance between $\mathbf{x}_i$ and $\mathbf{x}_l$ that belong to different classes). Similarly, we expect that $D_{F2N}(\mathbf{x}_f^i, q, \mathbf{M})$ should be smaller than $D_{F2N}(\mathbf{x}_f^i, \bar{q}, \mathbf{M})$ to some extent, where q is the assigned name of $\mathbf{x}_f^i$ according to the groundtruth label matrix $\mathbf{Y}_*^i$ and $\bar{q}$ is the assigned name of $\mathbf{x}_f^i$ according to an infeasible label matrix $\bar{\mathbf{Y}}^i$. For all the faces in the i-th image and a label matrix $\mathbf{Y}^i$, we define the image to assignment (I2A) distance $D(\mathbf{X}^i, \mathbf{Y}^i, \mathbf{M})$ to be the sum of F2N distances between every face and its assigned names. Mathematically, $D(\mathbf{X}^i, \mathbf{Y}^i, \mathbf{M})$ is defined as

$$D(\mathbf{X}^i, \mathbf{Y}^i, \mathbf{M}) = \sum_{f=1}^{n_i} \sum_{q:Y^i_{q,f}=1} D_{F2N}(\mathbf{x}_f^i, q, \mathbf{M}).$$

In the ideal case, we expect that $D(\mathbf{X}^i, \mathbf{Y}_*^i, \mathbf{M})$ should be smaller than $D(\mathbf{X}^i, \bar{\mathbf{Y}}^i, \mathbf{M})$ by at least $h(\bar{\mathbf{Y}}^i, \mathbf{Y}_*^i)$, where $h(\bar{\mathbf{Y}}^i, \mathbf{Y}_*^i)$ is the number of faces that are assigned with different names based on two label matrices $\bar{\mathbf{Y}}^i$ and $\mathbf{Y}_*^i$. To tolerate the cases where $D(\mathbf{X}^i, \bar{\mathbf{Y}}^i, \mathbf{M}) - D(\mathbf{X}^i, \mathbf{Y}_*^i, \mathbf{M})$ is slightly smaller than $h(\bar{\mathbf{Y}}^i, \mathbf{Y}_*^i)$, we introduce a *non-negative* slack variable ξ_i for the i-th image, and have the following constraint for any $\bar{\mathbf{Y}}^i \in \bar{\mathcal{Y}}^i$:

$$D(\mathbf{X}^i, \bar{\mathbf{Y}}^i, \mathbf{M}) - D(\mathbf{X}^i, \mathbf{Y}_*^i, \mathbf{M}) \geq h(\bar{\mathbf{Y}}^i, \mathbf{Y}_*^i) - \xi_i. \qquad (9)$$

However, the groundtruth label matrix $\mathbf{Y}_*^i$ is unknown, so $h(\bar{\mathbf{Y}}^i, \mathbf{Y}_*^i)$ and $D(\mathbf{X}^i, \mathbf{Y}_*^i, \mathbf{M})$ in (9) are not available. Although $\mathbf{Y}_*^i$ is unknown, it should be a label matrix in the feasible label set $\mathcal{Y}^i$. In this work, we use $\ell(\bar{\mathbf{Y}}^i, \mathcal{Y}^i)$ to approximate $h(\bar{\mathbf{Y}}^i, \mathbf{Y}_*^i)$, where $\ell(\bar{\mathbf{Y}}^i, \mathcal{Y}^i)$ measures the difference between an infeasible label matrix $\bar{\mathbf{Y}}^i$ and the most similar label matrix in $\mathcal{Y}_i$. Similarly as in [7], we define $\ell(\bar{\mathbf{Y}}^i, \mathcal{Y}^i)$ as follows:

$$\ell(\bar{\mathbf{Y}}^i, \mathcal{Y}^i) = \min_{\mathbf{Y}^i\in\mathcal{Y}^i} h(\bar{\mathbf{Y}}^i, \mathbf{Y}^i).$$

On the other hand, since $\mathbf{Y}_*^i$ is in the feasible label set $\mathcal{Y}^i$ and we expect the corresponding I2A distance should be small, we use $\min_{\mathbf{Y}^i\in\mathcal{Y}^i} D(\mathbf{X}^i, \mathbf{Y}^i, \mathbf{M})$ to replace $D(\mathbf{X}^i, \mathbf{Y}_*^i, \mathbf{M})$, where $\min_{\mathbf{Y}^i\in\mathcal{Y}^i} D(\mathbf{X}^i, \mathbf{Y}^i, \mathbf{M})$ is the smallest I2A distance based on the feasible label matrix inside $\mathcal{Y}_i$. In summary, by replacing $h(\bar{\mathbf{Y}}^i, \mathbf{Y}_*^i)$ and $D(\mathbf{X}^i, \mathbf{Y}_*^i, \mathbf{M})$ with $\ell(\bar{\mathbf{Y}}^i, \mathcal{Y}^i)$ and $\min_{\mathbf{Y}^i\in\mathcal{Y}^i} D(\mathbf{X}^i, \mathbf{Y}^i, \mathbf{M})$, respectively, the constraint in (9) becomes the following one for any $\bar{\mathbf{Y}}^i \in \bar{\mathcal{Y}}^i$:

$$\xi_i \geq \ell(\bar{\mathbf{Y}}^i, \mathcal{Y}^i) - D(\mathbf{X}^i, \bar{\mathbf{Y}}^i, \mathbf{M}) + \min_{\mathbf{Y}^i\in\mathcal{Y}^i} D(\mathbf{X}^i, \mathbf{Y}^i, \mathbf{M}), \quad (10)$$

Instead of enforcing ξ_i to be no less than every $\ell(\bar{\mathbf{Y}}^i, \mathcal{Y}^i) - D(\mathbf{X}^i, \bar{\mathbf{Y}}^i, \mathbf{M}) + \min_{\mathbf{Y}^i \in \mathcal{Y}^i} D(\mathbf{X}^i, \mathbf{Y}^i, \mathbf{M})$ (each based on an infeasible label matrix $\bar{\mathbf{Y}}^i$ in $\bar{\mathcal{Y}}^i$) as in (10), we can equivalently enforce ξ_i to be no less than the largest one of them. Note that the term $\min_{\mathbf{Y}^i \in \mathcal{Y}^i} D(\mathbf{X}^i, \mathbf{Y}^i, \mathbf{M})$ is irrelevant to $\bar{\mathbf{Y}}^i$. Accordingly, we rewrite (10) *w.r.t.* the non-negative slack variable ξ_i in the following equivalent form:

$$\xi_i \geq \max_{\bar{\mathbf{Y}}^i \in \bar{\mathcal{Y}}^i} \big[\ell(\bar{\mathbf{Y}}^i, \mathcal{Y}^i) - D(\mathbf{X}^i, \bar{\mathbf{Y}}^i, \mathbf{M})\big] + \min_{\mathbf{Y}^i \in \mathcal{Y}^i} D(\mathbf{X}^i, \mathbf{Y}^i, \mathbf{M}).$$

Hence, we propose a new method called Ambiguously-supervised Structural Metric Learning (ASML) to learn a discriminative Mahalanobis distance metric $\mathbf{M}$, by solving the following problem:

$$\begin{aligned} \min_{\mathbf{M} \succeq 0} \quad & \frac{\sigma}{2}\|\mathbf{M} - \mathbf{I}\|_F^2 + \frac{1}{m}\sum_{i=1}^{m} |\max_{\bar{\mathbf{Y}}^i \in \bar{\mathcal{Y}}^i}[\ell(\bar{\mathbf{Y}}^i, \mathcal{Y}^i) \\ & - D(\mathbf{X}^i, \bar{\mathbf{Y}}^i, \mathbf{M})] + \min_{\mathbf{Y}^i \in \mathcal{Y}^i} D(\mathbf{X}^i, \mathbf{Y}^i, \mathbf{M})|_+ . \end{aligned} \tag{11}$$

where $\sigma > 0$ is a tradeoff parameter, the regularizer $\|\mathbf{M} - \mathbf{I}\|_F^2$ is used to enforce $\mathbf{M}$ to be not too far away from the identity matrix $\mathbf{I}$, and we also rewrite ξ_i as $|\max_{\bar{\mathbf{Y}}^i \in \bar{\mathcal{Y}}^i}[\ell(\bar{\mathbf{Y}}^i, \mathcal{Y}^i) - D(\mathbf{X}^i, \bar{\mathbf{Y}}^i, \mathbf{M})] + \min_{\mathbf{Y}^i \in \mathcal{Y}^i} D(\mathbf{X}^i, \mathbf{Y}^i, \mathbf{M})|_+$, similar as that in LMNN. Note that we have incorporated weak supervision information in the max margin loss in (11). A nice property of such max margin loss is the robustness to label noise.

Optimization: Since $\min_{\mathbf{Y}^i \in \mathcal{Y}^i} D(\mathbf{X}^i, \mathbf{Y}^i, \mathbf{M})$ in (11) is concave, the objective function in (11) is non-convex w.r.t. $\mathbf{M}$. For convenience, we define two convex functions $f_i(\mathbf{M}) = \max_{\bar{\mathbf{Y}}^i \in \bar{\mathcal{Y}}^i}[\ell(\bar{\mathbf{Y}}^i, \mathcal{Y}^i) - D(\mathbf{X}^i, \bar{\mathbf{Y}}^i, \mathbf{M})]$ and $g_i(\mathbf{M}) = -\min_{\mathbf{Y}^i \in \mathcal{Y}^i} D(\mathbf{X}^i, \mathbf{Y}^i, \mathbf{M})$, $\forall i = 1, \ldots, m$. Inspired by the concave-convex procedure (CCCP) method [29], we equivalently rewrite (11) as follows:

$$\min_{\mathbf{M} \succeq 0} \frac{\sigma}{2}\|\mathbf{M} - \mathbf{I}\|_F^2 + \frac{1}{m}\sum_{i=1}^{m} |f_i(\mathbf{M}) - g_i(\mathbf{M})|_+ . \tag{12}$$

We solve the problem in (12) in an iterative fashion. Let us denote $\mathbf{M}$ at the s-th iteration as $\mathbf{M}_{(s)}$. Similarly as in CCCP, at the $(s+1)$-th iteration, we replace the non-convex term $|f_i(\mathbf{M}) - g_i(\mathbf{M})|_+$ with a convex term $|f_i(\mathbf{M}) - \big\langle \mathbf{M}, \tilde{g}_i(\mathbf{M}_{(s)})\big\rangle|_+$, where $\tilde{g}_i(\cdot)$ is the subgradient [7] of $g_i(\cdot)$. Hence, at the $(s+1)$-th iteration, we solve the following relaxed version of the problem in (12):

$$\min_{\mathbf{M} \succeq 0} \frac{\sigma}{2}\|\mathbf{M} - \mathbf{I}\|_F^2 + \frac{1}{m}\sum_{i=1}^{m} \big|f_i(\mathbf{M}) - \big\langle \mathbf{M}, \tilde{g}_i(\mathbf{M}_{(s)})\big\rangle\big|_+ , \tag{13}$$

which is now convex with respect to $\mathbf{M}$. To solve (13), we define $\mathbf{Q} = \mathbf{M} - \mathbf{I}$ and $\mathbf{Q}_{(s)} = \mathbf{M}_{(s)} - \mathbf{I}$, and equivalently rewrite (13) as the following convex optimization problem:

$$\begin{aligned} \min_{\mathbf{Q}, \tilde{\xi}_i} \quad & \frac{\sigma}{2}\|\mathbf{Q}\|_F^2 + \frac{1}{m}\sum_{i=1}^{m} \tilde{\xi}_i \\ s.t. \ & f_i(\mathbf{Q} + \mathbf{I}) - \big\langle \mathbf{Q} + \mathbf{I}, \tilde{g}_i(\mathbf{Q}_{(s)} + \mathbf{I})\big\rangle \leq \tilde{\xi}_i, \ \tilde{\xi}_i \geq 0, \forall i; \\ & \mathbf{Q} + \mathbf{I} \succeq \mathbf{0}. \end{aligned} \tag{14}$$

Although the optimization problem in (14) is convex, it may contain many constraints. To efficiently solve it, we adopt the stochastic subgradient descent method similarly as in Pegasos [30]. Moreover, to handle the positive semi-definite (PSD) constraint on $\mathbf{Q}+\mathbf{I}$ in (14), at each iteration when using the stochastic subgradient descent method, we additionally project the solution onto the PSD cone by thresholding the negative eigenvalues to be zeros, similarly as in [31]. The ASML algorithm is summarized in Algorithm 1.

Algorithm 1 The ASML algorithm

Input: : The training images $\{\mathbf{X}^i|_{i=1}^m\}$, the feasible label sets $\{\mathcal{Y}^i|_{i=1}^m\}$, the parameters σ, N_{iter} and ε.

1: Initialize[1] $\mathbf{M}_{(0)} = \mathbf{I}$.
2: **for** $s = 1 : N_{iter}$ **do**
3: Calcuate $\mathbf{Q}_{(s)}$ as $\mathbf{Q}_{(s)} = \mathbf{M}_{(s)} - \mathbf{I}$.
4: Obtain $\mathbf{Q}_{(s+1)}$ by solving the convex problem in (14) via the stochastic subgradient descent method.
5: Calcuate $\mathbf{M}_{(s+1)}$ as $\mathbf{M}_{(s+1)} = \mathbf{Q}_{(s+1)} + \mathbf{I}$.
6: break if $\|\mathbf{M}_{(s+1)} - \mathbf{M}_{(s)}\|_F \leq \varepsilon$.
7: **end for**

Output: the Mahalanobis distance metric $\mathbf{M}_{(s+1)}$.

IV. Inferring Names of Faces

With the coefficient matrix $\mathbf{W}^*$ learned from rLRR, we can calculate the first affinity matrix as $\mathbf{A}_W = \frac{1}{2}(\mathbf{W}^* + \mathbf{W}^{*\prime})$ and normalize $\mathbf{A}_W$ to the range $[0, 1]$. Furthermore, with the learnt distance metric $\mathbf{M}$ from ASML, we can calculate the second affinity matrix as $\mathbf{A}_K = \mathbf{K}$, where $\mathbf{K}$ is a kernel matrix based on the Mahalanobis distances between the faces. Since the two affinity matrices explore weak supervision information in different ways, they contain complementary information and both of them are beneficial for face naming. For better face naming performance, we combine these two affinity matrices and perform face naming based on the fused affinity matrix. Specifically, we obtain a fused affinity matrix $\mathbf{A}$ as the linear combination of the two affinity matrices, *i.e.*, $\mathbf{A} = (1-\alpha)\mathbf{A}_W + \alpha\mathbf{A}_K$, where α is a parameter in the range [0,1]. Finally, we perform face naming based on $\mathbf{A}$. Since the fused affinity matrix is obtained based on rLRR and ASML, we name our proposed method as "rLRRml". As mentioned in Sec. III-B, given this affinity matrix $\mathbf{A}$, we perform face naming by solving the following optimization problem:

$$\max_{\mathbf{Y} \in \mathcal{Y}} \sum_{c=1}^{p} \frac{\mathbf{y}_c' \mathbf{A} \mathbf{y}_c}{\mathbf{1}' \mathbf{y}_c} , \quad \text{s.t. } \mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_{(p+1)}]'. \tag{15}$$

However, the above problem is an integer programming problem, which is computationally expensive to solve. In this work, we propose an iterative approach to solve a relaxed version of (15). Specifically, at each iteration, we approximate the objective function by using $\frac{\tilde{\mathbf{y}}_c' \mathbf{A} \mathbf{y}_c}{\mathbf{1}' \tilde{\mathbf{y}}_c}$ to replace $\frac{\mathbf{y}_c' \mathbf{A} \mathbf{y}_c}{\mathbf{1}' \mathbf{y}_c}$, where $\tilde{\mathbf{y}}_c$ is the solution of $\mathbf{y}_c$ obtained from the previous iteration. Hence, at each iteration, we only need to solve a linear programming problem as follows,

$$\max_{\mathbf{Y} \in \mathcal{Y}} \sum_{c=1}^{p} \mathbf{b}_c' \mathbf{y}_c , \quad \text{s.t. } \mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_{(p+1)}]', \tag{16}$$

[1]Our experiments show that the results by using this initialization are comparable with those by using random initialization.

where $\mathbf{b}_c = \frac{\mathbf{A}\tilde{\mathbf{y}}_c}{\mathbf{1}'\tilde{\mathbf{y}}_c}$, $\forall c = 1, \ldots, p$. Moreover, the candidate name set $\mathcal{N}^i$ may be incomplete, so some faces in the image $\mathbf{X}^i$ may not have the corresponding groundtruth names in the candidate name set $\mathcal{N}^i$. Therefore, similarly as in [32], we additionally define a vector $\mathbf{b}_{p+1} = \theta\mathbf{1}$ to allow some faces to be assigned to the "null" class, where θ is a predefined parameter. Intuitively, the number of faces assigned to "null" changes when we set θ with different values. In the experiments, to fairly compare the proposed methods and other methods, we report the performances of all methods when each algorithm annotates the same number of faces using real names rather than "null", which can be achieved by tuning the parameter θ (see Sec V-C for more details).

By defining $\mathbf{B} \in \mathbb{R}^{(p+1)\times n}$ as $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_{p+1}]'$, we can reformulate the problem in (16) as follows,

$$\max_{\mathbf{Y}\in\mathcal{Y}} \quad \langle \mathbf{B}, \mathbf{Y} \rangle . \tag{17}$$

Recall that the feasible set for $\mathbf{Y}$ is defined as $\mathcal{Y} = \{\mathbf{Y} = [\mathbf{Y}^1, \ldots, \mathbf{Y}^m] | \mathbf{Y}^i \in \mathcal{Y}^i, \forall i = 1, \ldots, m\}$, which means the constraints on $\mathbf{Y}^i$'s are separable. Let us decompose the matrix $\mathbf{B}$ as $\mathbf{B} = [\mathbf{B}^1, \ldots, \mathbf{B}^m]$ with each $\mathbf{B}^i \in \mathbb{R}^{(p+1)\times n_i}$ corresponding to $\mathbf{Y}^i$, then the objective function in (17) can be expressed as $\langle \mathbf{B}, \mathbf{Y} \rangle = \sum_{i=1}^m \langle \mathbf{B}^i, \mathbf{Y}^i \rangle$, which is also separable w.r.t. $\mathbf{Y}^i$'s. Hence, we optimize (17) by solving m subproblems, with each subproblem related to one image in the following form:

$$\max_{\mathbf{Y}^i\in\mathcal{Y}^i} \quad \langle \mathbf{B}^i, \mathbf{Y}^i \rangle, \tag{18}$$

$\forall i = 1, \ldots, m$. In particular, the i-th problem in (18) can equivalently rewritten as a *minimization* problem with detailed constraints as follows:

$$\begin{aligned} \min_{Y^i_{q,f}\in\{0,1\}} \quad & \sum_{q\in\tilde{\mathcal{N}}^i} \sum_{f=1}^{n_i} -B^i_{q,f} Y^i_{q,f}, \\ \text{s.t.} \quad & \sum_{q\in\tilde{\mathcal{N}}^i} Y^i_{q,f} = 1, \forall f = 1, \ldots, n_i \\ & \sum_{f=1}^{n_i} Y^i_{q,f} \le 1, \forall q \in \mathcal{N}^i, \\ & \sum_{f=1}^{n_i} Y^i_{(p+1),f} \le n_i, \end{aligned} \tag{19}$$

in which we have dropped the elements $\{Y^i_{q,f}|_{q\notin\tilde{\mathcal{N}}^i}\}$, because these elements are zeros according to the feasibility constraint in (1). Similarly as in [32], we solve the problem in (19) by converting it to a minimum cost bipartite graph matching problem, for which the objective is the sum of the costs for assigning faces to names. In this work, we adopt the Hungarian algorithm to efficiently solve it. Specifically, for the i-th image, the cost $c(f, q)$ for assigning a face $\mathbf{x}^i_f$ to a real name q is set to $-B^i_{q,f}$, and the cost $c(f, p+1)$ for assigning a face $\mathbf{x}^i_f$ to the corresponding "null" name is set to $-B^i_{(p+1),f}$.

In summary, to infer the label matrix $\mathbf{Y}$ for all faces, we iteratively solve the linear programming problem in (17), which can be efficiently addressed by solving m subproblems as in (19) with the Hungarian algorithm. Let $\mathbf{Y}(t)$ be the label matrix at the t-th iteration. The initial label matrix $\mathbf{Y}(1)$ is set to the label matrix which assigns each face to all names in the caption associated with the corresponding image that contains this face. The iterative process continues until the convergence condition is satisfied. In practice, this iterative process always converges in about 10 to 15 iterations, so we empirically set $\tilde{N}_{iter}$ as 15. The iterative algorithm for face naming is summarized in Algorithm 2.

Algorithm 2 The face naming algorithm

Input: : The feasible label sets $\{\mathcal{Y}^i|_{i=1}^m\}$, the affinity matrix $\mathbf{A}$, the initial label matrix $\mathbf{Y}(1)$ and the parameters $\tilde{N}_{iter}$, θ.

1: **for** $t = 1 : \tilde{N}_{iter}$ **do**
2: Update $\mathbf{B}$ by using $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_{p+1}]'$, where $\mathbf{b}_c = \frac{\mathbf{A}\tilde{\mathbf{y}}_c}{\mathbf{1}'\tilde{\mathbf{y}}_c}$, $\forall c = 1, \ldots, p$ with $\tilde{\mathbf{y}}_c$ being the c-th column of $\mathbf{Y}(t)$, and $\mathbf{b}_{p+1} = \theta\mathbf{1}$.
3: Update $\mathbf{Y}(t+1)$ by solving m subproblems in (19).
4: break if $\mathbf{Y}(t+1) = \mathbf{Y}(t)$.
5: **end for**

Output: the label matrix $\mathbf{Y}(t+1)$.

V. Experiments

In this section, we compare our proposed methods rLRR, ASML and rLRRml with four state-of-the-art algorithms for face naming, as well as two special cases LRR and LRRml by using a synthetic dataset and two real-world datasets.

A. Introduction of the datasets

One synthetic dataset and two real-world benchmark datasets are used in the experiments. The synthetic dataset is collected from the Faces and Poses dataset in [33]. We first find out the top 10 popular names and then for each name, we randomly sample 50 images where this name appears in the image tags. In total, the synthetic dataset contains 602 faces in 500 images, with a total number of 20 names appearing in the corresponding tags, which include these top 10 popular names and other names associated with these 500 images.

Other than the synthetic dataset, the experiments are also conducted on the following two real-world datasets:

Soccer player dataset. This dataset was used in [9], with the images of soccer players from famous European clubs and names mentioned in the captions. The detected faces are manually annotated by using names from the captions or as "null". Following [9], we retain 170 names that occur at least 20 times in the captions and treat others as the "null" class. The images without containing any of these 170 names are discarded.

Labeled Yahoo! News dataset. This dataset was collected in [34] and further processed in [6]. It contains news images as well as the names in the captions. Following [9] and [7], we retain the 214 names occurred at least 20 times in the captions and treat others as the "null" class. The images that do not contain any of the 214 names are removed.

The detailed information about the synthetic and real-world datasets is shown in Table I, where the "*groundtruth real name*

TABLE I
DETAILS OF THE DATASETS. THE COLUMNS IN TURNS ARE THE TOTAL NUMBER OF IMAGES, FACES AND NAMES, THE AVERAGE NUMBER OF DETECTED FACES PER IMAGE, THE AVERAGE NUMBER OF DETECTED NAMES PER CAPTION AND THE GROUNDTRUTH RATIO, RESPECTIVELY.

Dataset	#images	#faces	#names	#faces/image	#names/caption	groundtruth ratio
Synthetic	500	602	20	1.20	1.07	0.89
Soccer player	8640	17472	170	2.02	1.74	0.51
Labeled Yahoo! News	10128	15868	214	1.57	1.37	0.56

ratio" (or "*groundtruth ratio*" in short) is the percentage of faces whose groundtruth names are real names (rather than "null") among all the faces in the dataset. In the Soccer player dataset, there are more images with multiple faces and multiple names in the captions when compared with the Labeled Yahoo! News dataset, which indicates the Soccer player dataset is more challenging. For the synthetic dataset and the two real-world datasets, we extract the feature vectors to represent the faces in the same way as in [10]. For each face, 13 interest points (facial landmarks) are located. For each interest point, a simple pixel-wised descriptor is formed by using the gray-level intensity values of pixels in the elliptical region based on each interest point, which is further normalized to achieve local photometric invariance [10]. Finally, a 1,937-dimensional descriptor for each face is obtained by concatenating the descriptors from 13 interest points.

B. Baseline methods and two special cases

The following four state-of-the-art methods are used as baselines:

- Maximum Margin Set (**MMS**) learning algorithm [7], which solves the face naming problem by learning SVM classifiers for each name.
- Multiple Instance Logistic Discriminant Metric Learning (**MildML**) [6], which learns a Mahalanobis distance metric such that the bags (images) with common labels (names in captions) are pulled closer, while the bags that do not share any common label are pushed apart.
- Constrained-Gaussian Mixture Model (**cGMM**) [35] [32]. For this Gaussian mixture model based approach, each name is associated with a Gaussian density function in the feature space with the parameters estimated from the data, and each face is assumed to be independently generated from the associated Gaussian function. The overall assignments are chosen to achieve the maximum log-likelihood.
- Low-rank SVM (**LR-SVM**) [9], which simultaneously learns the partial permutation matrices for grouping the faces and minimize the rank of the data matrices from each group. SVM classifiers are also trained for each name to deal with the out-of-sample cases.

More details of these methods can be found in Section II. For detailed analysis of the proposed rLRRml, we also report the results of the following two special cases:

- **LRRml**. rLRRml reduces to LRRml if we do not introduce the proposed regularizer on $\mathbf{W}$. In other words, we set the parameter γ in rLRR to 0 when learning $\mathbf{W}$.
- **LRR**. rLRRml reduces to LRR if we neither consider the affinity matrix $\mathbf{A}_K$ nor pose the proposed regularizer on $\mathbf{W}$. In other words, we set the parameter γ in rLRR to 0 when learning the coefficient matrix $\mathbf{W}$, and we use $\mathbf{A}_W$ as the input affinity matrix $\mathbf{A}$ in Algorithm 2.

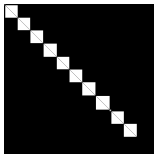
On the synthetic dataset, we empirically set γ to 100 for our rLRR, and we empirically set λ to 0.01 for both LRR and rLRR.[1] On the real-world datasets, for MMS, we tune the parameter C in the range of $\{1, 10, \ldots, 10^4\}$ and report the best results from the optimal C. For MildML, we tune the parameter about the metric rank in the range of $\{2^2, 2^3, \ldots, 2^7\}$ and report the best results. For cGMM, there are no parameters to be set. For the parameters λ and C in LR-SVM, instead of fixing $\lambda = 0.3$ and choosing C in the range of $\{0.1, 1, 10\}$ as in [9], we tune these parameters in larger ranges. Specifically, we tune λ in the range of $\{1, 0.3, 0.1, 0.01\}$ and C in the range of $\{10^{-2}, 10^{-1}, \ldots, 10^2\}$, and report the best results from the optimal λ and C. The parameter α for fusing the two affinity matrices in rLRRml and LRRml is empirically fixed as 0.1 on both real-world datasets, namely we calculate $\mathbf{A}$ as $\mathbf{A} = 0.9\mathbf{A}_W + 0.1\mathbf{A}_K$. On the two real-world datasets, after tuning λ in LRR in the range of $\{1, 0.1, 0.01, 0.001\}$, we observe that LRR achieves the best results when setting λ to 0.01 on both datasets, so we fix the parameter λ for LRR, rLRR, LRRml and rLRRml to 0.01 on both datasets. The parameter γ for rLRR and rLRRml is empirically set to 100, and the tradeoff parameter σ for ASML, LRRml and rLRRml is empirically fixed to 1. For the kernel matrix $\mathbf{K}$ in ASML, LRRml and rLRRml, we use the kernel matrix based on the Mahalanobis distances, namely we have $K_{i,j} = \exp(-\sqrt{\nu} D_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j))$, where $D_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)'\mathbf{M}(\mathbf{x}_i - \mathbf{x}_j)}$ is the Mahalanobis distance between $\mathbf{x}_i$ and $\mathbf{x}_j$, and ν is the bandwidth parameter set as the default value $1/\beta$, with β being the mean of squared Mahalanobis distances between all samples [36].

C. Experimental Results

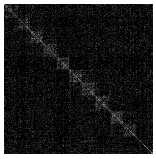
1) Results on the synthetic dataset: Firstly, to validate the effectiveness of our proposed method rLRR for recovering subspace information, we compare the coefficient matrices obtained from LRR and rLRR with the ideal coefficient matrix $\mathbf{W}^*$ according to the groundtruth.

Fig. 2(a) shows the ideal coefficient matrix $\mathbf{W}^*$ according to the groundtruth. For better viewing, the faces are re-ordered by grouping the faces belonging to the same name at contiguous positions. Note the white points indicate that the corresponding

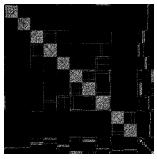
[1] We set the parameters ρ_{max}, $\Delta\rho$ and ϵ to the default values in the code from http://www.columbia.edu/~js4038/software.html. We set the number of iterations to 20 since the result becomes stable after about 20 iterations.



(a) $\mathbf{W}^*$ according to the groundtruth



(b) $\mathbf{W}^*$ from LRR



(c) $\mathbf{W}^*$ from rLRR

Fig. 2. The coefficient matrix $\mathbf{W}^*$ according to the groundtruth and the ones obtained from LRR and rLRR.

faces belong to the same subject (i.e., with the same name), and the bottom-right part corresponds to the faces from the "null" class. The diagonal entries are set to be zeros since we expect self-reconstruction can be avoided.

Fig. 2(b) shows the coefficient matrix $\mathbf{W}^*$ obtained from LRR. While there exists block-wise diagonal structure to some extents, we also observe that 1) The diagonal elements are large, meaning that a face is reconstructed mainly by itself. It should be avoided. 2) Generally, the coefficients between faces from the same subject are not significantly larger than the ones between faces from different subjects.

Fig. 2(c) shows the coefficient matrix $\mathbf{W}^*$ obtained from our rLRR. It has smaller values for the diagonal elements. Generally, the coefficients between faces from the same subject become larger, while the ones between faces from different subjects become smaller. Compared with Fig. 2(b), Fig. 2(c) is more similar to the ideal coefficient matrix in Fig. 2(a), because the reconstruction coefficients exhibit more obvious block-wise diagonal structure.

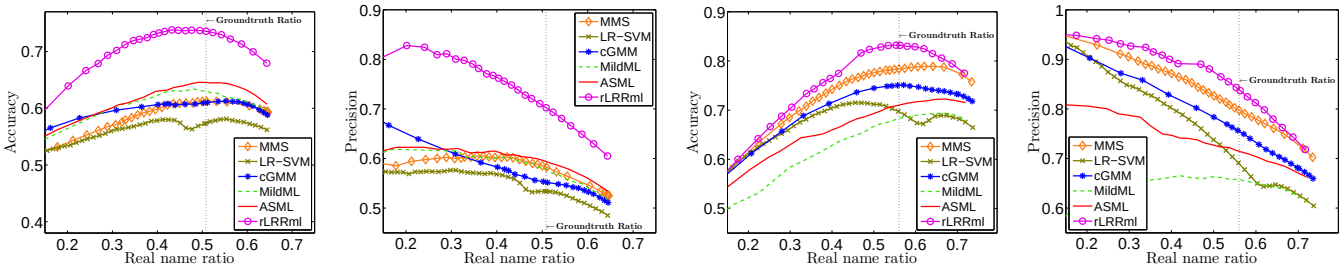
2) Results on the real-world datasets : For performance evaluation, we follow [37] to take the accuracy and precision as two criteria. The accuracy is the percentage of correctly annotated faces (also including the correctly annotated faces whose groundtruth name is the "null" name) over all faces, while the precision is the percentage of correctly annotated faces over the faces which are annotated as real names (*i.e.*, we do not consider the faces annotated as the "null" class by a face naming method). Since all methods aim at inferring names based on the faces in the images with ambiguous captions, we use all the images in each dataset for both learning and testing. To fairly compare all methods, we define the "*real name ratio*" as the percentage of faces which are annotated as real names by using each method over all the faces in the dataset, and we report the performances at the same real name ratio.

In order to achieve the same real name ratio for all methods, we use the minimum cost bipartite graph matching method (introduced in Sec. IV) to infer the names of the faces, and vary the hyper-parameter θ to tune the real name ratio, as suggested in [37]. Specifically, the costs $c(f,q)$ and $c(f,p+1)$ are set as follows. For MildML, we set $c(f,q) = -\sum_{\mathbf{x}\in\mathcal{S}_q} w(\mathbf{x}_f^i, \mathbf{x})$ and $c(f,p+1) = \theta$ as in [6], where $w(\mathbf{x}_f^i, \mathbf{x})$ is the similarity between $\mathbf{x}_f^i$ and $\mathbf{x}$, and $\mathcal{S}_q$ contains all faces assigned to the name q while inferring the names of the faces. For cGMM, we set $c(f,q) = -\ln\mathcal{N}(\mathbf{x}_f^i; \boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q)$, and $c(f,p+1) = -\ln\mathcal{N}(\mathbf{x}_f^i; \boldsymbol{\mu}_{(p+1)}, \boldsymbol{\Sigma}_{(p+1)}) + \theta$, as in [32], where $\boldsymbol{\mu}_q$ and $\boldsymbol{\Sigma}_q$ (resp. $\boldsymbol{\mu}_{(p+1)}$ and $\boldsymbol{\Sigma}_{(p+1)}$) are the mean and covariance of the faces assigned to the q-th class (resp. the "null" class) in cGMM. Similarly, for MMS and LR-SVM, we consider the decision values from the SVM classifiers of the n-th name and the "null" class by setting the cost as $c(f,q) = -dec_q(\mathbf{x}_f^i)$ and $c(f,p+1) = -dec_{null}(\mathbf{x}_f^i) + \theta$, where $dec_q(\mathbf{x}_f^i)$ and $dec_{null}(\mathbf{x}_f^i)$ are the decision values of SVM classifiers from the q-th name and the "null" class, respectively. The accuracies and precisions of different methods on the real-world datasets are shown in Table II, where the real name ratio for each method is set to be close to the groundtruth ratio by using a suitable hyper-parameter θ, as suggested in [37]. For a more comprehensive comparison, we also plot the accuracies and precisions on these two real-world datasets when using different real name ratios for all methods, by varying the value of the parameter θ. In Fig. 3, we compare the performances of our proposed methods ASML and rLRRml with the baseline methods MMS, cGMM, LR-SVM and MildML on these two real-world datasets, respectively. In Fig. 4, we compare the performances of our proposed methods rLRRml, rLRR with the special cases LRRml and LRR on these two real-world datasets, respectively. According to these results, we have the following observations:

TABLE II
PERFORMANCES (AT GROUNDTRUTH RATIOS) OF DIFFERENT METHODS ON TWO REAL-WORLD DATASETS. THE BEST RESULTS ARE IN BOLD.

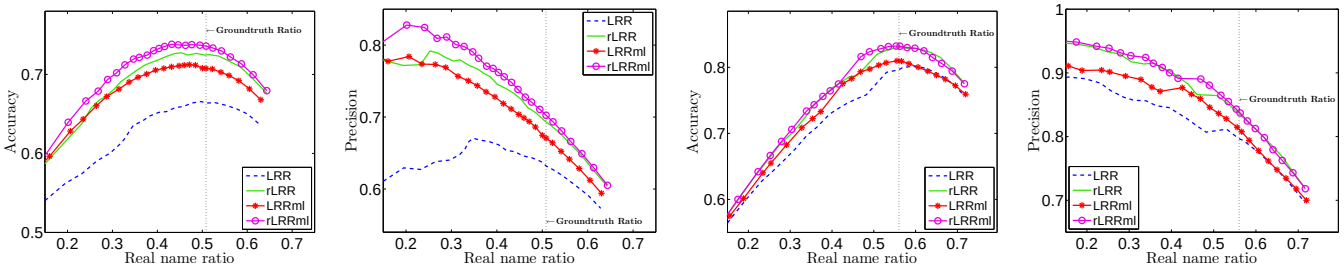
Method	Soccer player		Labeled Yahoo! News	
	Accuracy	Precision	Accuracy	Precision
MMS	0.613	0.583	0.784	0.797
LR-SVM	0.574	0.534	0.697	0.690
cGMM	0.611	0.553	0.750	0.755
MildML	0.630	0.579	0.684	0.658
ASML	0.646	0.594	0.710	0.715
LRR	0.664	0.632	0.797	0.797
rLRR	0.725	0.694	0.830	0.836
LRRml	0.708	0.671	0.810	0.812
rLRRml	**0.736**	**0.703**	**0.832**	**0.839**

- Among the four baseline algorithms MMS, cGMM, LR-SVM and MildML, there is no consistent winner on both datasets in terms of the accuracies and precisions in Table. II. On the Labeled Yahoo! News dataset, MMS achieves the best accuracy and precision among four methods. On the Soccer player dataset, MMS still achieves the best precision, but MildML achieves the best accuracy.
- We also compare ASML with MildML, because both methods use captions-based weak supervision for distance metric learning. According to Table II, ASML outperforms MildML on both datasets in terms of both accuracy and precision. From Fig. 3, we observe that ASML consistently outperforms MildML on the Labeled Yahoo! News dataset, and generally outperforms MildML on the Soccer player dataset. These results indicate that ASML can learn a more discriminative distance metric by better utilizing ambiguous supervision information.
- LRR performs well on both datasets, which indicates that our assumption that faces in a common subspace should belong to the same subject/name is generally satisfied on both real-world datasets. Moreover, rLRR consistently



(a) Accuracy vs. real name ratio on the Soccer player dataset. (b) Precision vs. real name ratio on the Soccer player dataset. (c) Accuracy vs. real name ratio on the Labeled Yahoo! News dataset. (d) Precision vs. real name ratio on the Labeled Yahoo! News dataset.

Fig. 3. The accuracy and precision curves for the proposed methods rLRRml and ASML, as well as the baseline methods MMS, cGMM, LR-SVM and MildML, on the Soccer player dataset and the Labeled Yahoo! News dataset.



(a) Accuracy vs. real name ratio on the Soccer player dataset. (b) Precision vs. real name ratio on the Soccer player dataset. (c) Accuracy vs. real name ratio on the Labeled Yahoo! News dataset. (d) Precision vs. real name ratio on the Labeled Yahoo! News dataset.

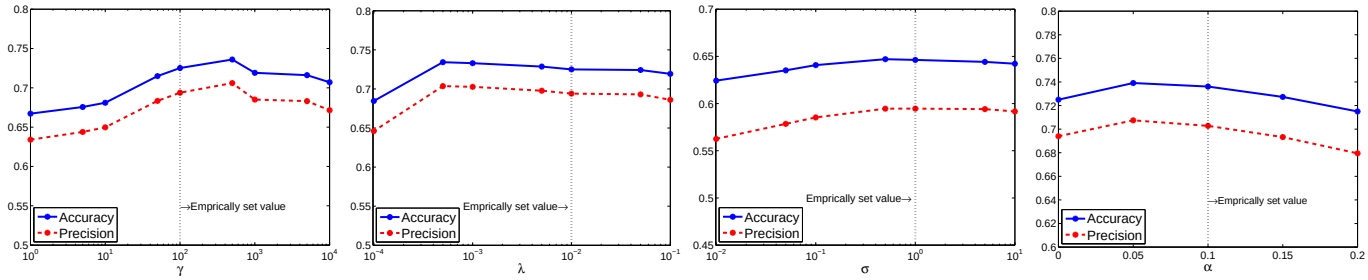
Fig. 4. The accuracy and precision curves for the proposed methods rLRRml and rLRR, as well as the special cases LRRml and LRR, on the Soccer player dataset and the Labeled Yahoo! News dataset.

achieves much better performance when compared with the original LRR algorithm on both datasets (see Table II and Fig. 4), which demonstrates it is beneficial to additionally consider weak supervision information by introducing the new regularizer into LRR while exploring the *subspace structures among faces*.

- According to Table II, rLRRml is better than rLRR, and LRRml also outperforms LRR on both datasets in terms of accuracy and precision. On the Soccer player dataset (see Fig. 4(a) and Fig. 4(b)), rLRRml (resp., LRRml) consistently outperforms rLRR (resp., LRR). On the Labeled Yahoo! News dataset (see Fig. 4(c) and Fig. 4(d)), rLRRml (resp., LRRml) generally outperforms rLRR (resp., LRR). One possible explanation is, these two affinity matrices contain complementary information to some extent because they explore weak supervision information in different ways. Hence, the fused affinity matrix is more discriminative for face naming. Note that the performance of rLRR on the Labeled Yahoo! News dataset is already high, so the improvement of rLRRml over rLRR on this dataset is not as significant as that on the Soccer player dataset.
- Compared with all other algorithms, the proposed rLRRml algorithm achieves the best results in terms of both accuracy and precision on both datasets (see Table II). It can be observed that rLRRml consistently outperforms all other methods on the Soccer player dataset (see Fig. 3(a), Fig. 3(b), Fig. 4(a) and Fig. 4(b)), and rLRRml generally achieves the best performance on the Labeled Yahoo! News dataset (see Fig. 3(c), Fig. 3(d), Fig. 4(c) and Fig. 4(d)). These results demonstrate the effectiveness of our rLRRml for face naming.
- For all methods, the results on the Soccer player dataset are worse than those on the Labeled Yahoo! News dataset. One possible explanation is, the Soccer player dataset is a more challenging dataset because there are more faces in each image, more names in each caption and relatively more faces from the "null" class in the Soccer player dataset (see Table I).

More discussions on $\mathbf{H}$ *in our rLRR:* In our rLRR, we penalize the following two cases by using the specially designed $\mathbf{H}$: 1) a face is reconstructed by the irrelevant faces that do not share any common names with this face according to their candidate name sets. 2) a face is reconstructed by using itself. If we only consider one case when designing $\mathbf{H}$ in our rLRR, the corresponding results will be worse than the current results in Table II. Taking the Soccer player dataset as an example, we re-define $\mathbf{H}$ by only considering the first (resp., second) case, the accuracy and precision of our rLRR method become 0.714 and 0.682 (resp., 0.694 and 0.664), respectively. These results are worse than the results (*i.e.*, the accuracy is 0.725 and the precision is 0.694) of our rLRR in Table II that considers both cases when designing $\mathbf{H}$, which experimentally validates the effectiveness of penalizing both cases.

3) Performance variations of our methods using different parameters: We take the Soccer player dataset as an example to study the performances (i.e., accuracies and precisions) of



(a) The performances of rLRR vs. γ. (b) The performances of rLRR vs. λ. (c) The performances of ASML vs. σ. (d) The performances of rLRRml vs. α.

Fig. 5. The performances (accuracies and precisions) of our methods on the Soccer player dataset when using different parameters. The black dotted line indicates the empirically set value (i.e., the default value) of each parameter.

our methods by using different parameters.

We first study the performances of our rLRR when using different parameters γ and λ and the results are shown in Fig. 5(a) and Fig. 5(b), respectively. Note we vary one parameter and set another parameter as its default value (i.e., $\gamma = 100$ and $\lambda = 0.01$). In (4), γ is the trade-off parameter for balancing the new regularizer $\|\mathbf{W} \circ \mathbf{H}\|_F^2$ (which incorporates weakly supervised information) and other terms. Recall that our rLRR reduces to LRR when γ is set to 0. When setting γ in the range of $(1, 500)$, the performances of rLRR become better as γ increases, and rLRR consistently outperforms LRR, which again shows it is beneficial to utilize weakly supervised information. We also observe that the performances of rLRR are relatively stable when setting γ in the range of $(50, 5000)$. The parameter λ is used in both LRR and our rLRR. We observe that our rLRR is relatively robust to the parameter λ when setting λ in the range of $(5 \times 10^{-4}, 10^{-1})$.

In Fig. 5(c), we show the results of our new metric learning method ASML when using different parameter σ in (11). It can be observed that our ASML is relatively stable to the parameter σ when σ is in the range of $(0.1, 10)$.

Finally, we study the performance variations of our rLRRml when setting the parameter α to different values, as shown in Fig. 5(d). When setting $\alpha = 0$ and $\alpha = 1$, rLRRml reduces to rLRR and ASML, respectively. As shown in Table II, rLRR is better than ASML in both cases in terms of accuracy and precision. So we empirically set α as a smaller value such that the affinity matrix from rLRR contributes more in the fused affinity matrix. When setting α in the range of $(0.05, 0.15)$, we observe that our rLRRml is relatively robust to the parameter α and the results are consistently better than rLRR and ASML, which demonstrates that the two affinity matrices from rLRR and ASML contain complementary information to some extent.

VI. Conclusion

In this paper, we have proposed a new scheme for face naming with caption-based supervision, in which one image that may contain multiple faces is associated with a caption specifying only who is in the image. To effectively utilize the caption-based weak supervision, we propose an LRR based method called rLRR by introducing a new regularizer to utilize such weak supervision information. We also develop a new distance metric learning method ASML using weak supervision information to seek a discriminant Mahalanobis distance metric. Two affinity matrices can be obtained from rLRR and ASML, respectively. Moreover, we further fuse the two affinity matrices and additionally propose an iterative scheme for face naming based on the fused affinity matrix. The experiments conducted on a synthetic dataset clearly demonstrate the effectiveness of the new regularizer in rLRR. In the experiments on two challenging real-world datasets (*i.e.*, the Soccer player dataset and the Labeled Yahoo! News dataset), our rLRR outperforms LRR, and our ASML is better than the existing distance metric learning method MildML. Moreover, our proposed rLRRml outperforms rLRR and ASML as well as several state-of-the-art baseline algorithms.

In order to further improve the face naming performances, we plan to extend our rLRR in the future by additionally incorporating the ℓ_1 norm based regularizer and using other losses when designing new regularizers. We will also study how to automatically determine the optimal parameters for our methods in the future.

References

[1] P. Viola and M. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[2] G. Liu, Z. Lin, and Y. Yu, "Robust subspace segmentation by low-rank representation," in *Proceedings of the 27th International Conference on Machine Learning*, Haifa, Israel, Jun. 2010, pp. 663–670.

[3] T. L. Berg, A. C. Berg, J. Edwards, M. Maire, R. White, Y. W. Teh, E. G. Learned-Miller, and D. A. Forsyth, "Names and faces in the news," in *Proceedings of the 17th IEEE Conference on Computer Vision and Pattern Recognition*, Washington, DC, Jun. 2004, pp. 848–854.

[4] D. Ozkan and P. Duygulu, "A graph based approach for naming faces in news photos," in *Proceedings of the 19th IEEE Conference on Computer Vision and Pattern Recognition*, New York, NY, Jun. 2006, pp. 1477–1482.

[5] P. T. Pham, M.-F. Moens, and T. Tuytelaars, "Cross-media alignment of names and faces," *IEEE Trans. Multimedia*, vol. 12, no. 1, pp. 13–27, Jan. 2010.

[6] M. Guillaumin, J. J. Verbeek, and C. Schmid, "Multiple instance metric learning from automatically labeled bags of faces," in *Proceedings of the 11th European Conference on Computer Vision*, Heraklion, Crete, Sep. 2010, pp. 634–647.

[7] J. Luo and F. Orabona, "Learning from candidate labeling sets," in *Proceedsing of the 23rd Annual Conference on Advances in Neural Information Processing Systems*, Vancouver, Canada, Dec. 2010, pp. 1504–1512.

[8] X. Zhang, L. Zhang, X.-J. Wang, and H.-Y. Shum, "Finding celebrities in billions of web images," *IEEE Trans. Multimedia*, vol. 14, no. 4, pp. 995–1007, Aug. 2012.

[9] Z. Zeng, S. Xiao, K. Jia, T.-H. Chan, S. Gao, D. Xu, and Y. Ma, "Learning by associating ambiguously labeled images," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition*, Portland, OR, Jun. 2013, pp. 708–715.

[10] M. Everingham, J. Sivic, and A. Zisserman, ""hello! my name is... buffy" – automatic naming of characters in TV video." in *Proceedings of the 17th British Machine Vision Conference*, Edinburgh, UK, Sep. 2006, pp. 899–908.

[11] J. Sang and C. Xu, "Robust face-name graph matching for movie character identification," *IEEE Trans. Multimedia*, vol. 14, no. 3, pp. 586–596, Jun. 2012.

[12] Y.-F. Zhang, C. Xu, H. Lu, and Y.-M. Huang, "Character identification in feature-length films using global face-name matching," *IEEE Trans. Multimedia*, vol. 11, no. 7, pp. 1276–1288, Nov. 2009.

[13] M. Tapaswi, M. Bauml, and R. Stiefelhagen, ""Knock! Knock! Who is it?" probabilistic person identification in tv-series," in *Proceedings of the 25th IEEE Conference on Computer Vision and Pattern Recognition*, Providence, RI, Jun. 2012, pp. 2658–2665.

[14] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *Journal of the ACM*, vol. 58, no. 3, pp. 11:1–11:37, 2011.

[15] Y. Deng, Q. Dai, R. Liu, Z. Zhang, and S. Hu, "Low-rank structure learning via nonconvex heuristic recovery," *IEEE Trans. Neural Networks and Learning Systems*, vol. 24, no. 3, pp. 383–396, Mar. 2013.

[16] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *Journal of Machine Learning Research*, vol. 10, pp. 207–244, 2009.

[17] C. Shen, J. Kim, and L. Wang, "A scalable dual approach to semidefinite metric learning," in *Proceedings of the 24th IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, Jun. 2011, pp. 2601–2608.

[18] B. McFee and G. R. G. Lanckriet, "Metric learning to rank," in *Proceedings of the 27th International Conference on Machine Learning*, Haifa, Israel, Jun. 2010, pp. 775–782.

[19] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Proceedsing of the 16th Annual Conference on Neural Information Processing Systems*, Vancouver and Whistler, Canada, Dec. 2003, pp. 65–72.

[20] M.-L. Zhang and Z.-H. Zhou, "M^{3}MIML: A maximum margin method for multi-instance multi-label learning," in *Proceedings of the 8th IEEE International Conference on Data Mining*, Pisa, Italy, Dec. 2008, pp. 688–697.

[21] T. Cour, B. Sapp, C. Jordan, and B. Taskar, "Learning from ambiguously labeled images," in *Proceedings of the 22nd IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, Jun. 2009, pp. 919–926.

[22] E. Elhamifar and R. Vidal, "Sparse subspace clustering," in *Proceedings of the 22nd IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, Jun. 2009, pp. 2790–2797.

[23] C. Lu, J. Feng, Z. Lin, and S. Yan, "Correlation adaptive subspace segmentation by trace Lasso," in *Proceedings of the 12th IEEE International Conference on Computer Vision*, Sydney, Australia, Dec. 2013, pp. 1345–1352.

[24] S. Xiao, M. Tan, and D. Xu, "Weighted block-sparse low rank representation for face clustering in videos," in *Proceedings of the 13th European Conference on Computer Vision*, Zurich, Switzerland, Sep. 2014, pp. 123–138.

[25] Z. Lin, R. Liu, and Z. Su, "Linearized alternating direction method with adaptive penalty for low-rank representation," in *Proceedsing of the 24th Annual Conference on Neural Information Processing Systems*, Granada, Spain, Dec. 2011, pp. 612–620.

[26] J. Cai, C. Emmanuel, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.

[27] J. Yang, W. Yin, Y. Zhang, and Y. Wang, "A fast algorithm for edge-preserving variational multichannel image restoration," *SIAM Journal on Imaging Sciences*, vol. 2, no. 2, pp. 569–592, 2009.

[28] C. Shen, J. Kim, and L. Wang, "Scalable large-margin mahalanobis distance metric learning," *IEEE Trans. Neural Networks*, vol. 21, no. 9, pp. 1524–1530, Sep. 2010.

[29] A. L. Yuille and A. Rangarajan, "The concave-convex procedure," *Neural Computation*, vol. 15, no. 4, pp. 915–936, 2003.

[30] S. Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter, "Pegasos: primal estimated sub-gradient solver for svm," *Mathematical Programming*, vol. 127, no. 1, pp. 3–30, 2011.

[31] K. Q. Weinberger and L. K. Saul, "Fast solvers and efficient implementations for distance metric learning," in *Proceedsing of the 25th IEEE International Conference on Machine Learning*, Helsinki, Finland, Jun. 2008, pp. 1160–1167.

[32] M. Guillaumin, T. Mensink, J. J. Verbeek, and C. Schmid, "Face recognition from caption-based supervision," *International Journal of Computer Vision*, vol. 96, no. 1, pp. 64–82, 2012.

[33] J. Luo, B. Caputo, and V. Ferrari, "Who's doing what: Joint modeling of names and verbs for simultaneous face and pose annotation," in *Proceedsing of the 22rd Annual Conference on Neural Information Processing Systems*, Vancouver, Canada, Dec. 2009, pp. 1168–1176.

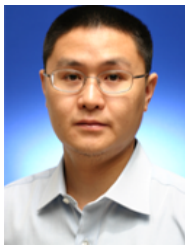
[34] T. L. Berg, E. C. Berg, J. Edwards, and D. Forsyth, "Who's in the picture," in *Proceedsing of the 19th Annual Conference on Neural Information Processing Systems*, Vancouver, Canada, Dec. 2006, pp. 137–144.

[35] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid, "Automatic face naming with caption-based supervision," in *Proceedings of the 21st IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AL, Jun. 2008, pp. 1–8.

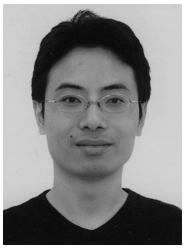
[36] X. Xu, I. W. Tsang, and D. Xu, "Soft margin multiple kernel learning," *IEEE Trans. Neural Networks and Learning Systems*, vol. 24, no. 5, pp. 749–761, May 2013.

[37] V. F. Mert Özcan, Jie Luo and B. Caputo, "A large-scale database of images and captions for automatic face naming," in *Proceedings of the 22nd British Machine Vision Conference*, Dundee, UK, Sep. 2011, pp. 1–11.

Shijie Xiao received the B.E. degree from the Harbin Institute of Technology in China, in 2011. Currently, he is pursuing a Ph.D. degree with the School of Computer Engineering, Nanyang Technological University, Singapore. His current research interests include machine learning and computer vision.



Dong Xu (M'07) received the B.E. and Ph.D. degrees from the University of Science and Technology of China, in 2001 and 2005, respectively. While pursuing his Ph.D., he was with Microsoft Research Asia, Beijing, China, and the Chinese University of Hong Kong, Shatin, Hong Kong, for more than two years. He was a Post-Doctoral Research Scientist with Columbia University, New York, NY, for one year. He is currently an Associate Professor with Nanyang Technological University, Singapore. His current research interests include computer vision, statistical learning, and multimedia content analysis. Dr. Xu was the co-author of a paper that won the Best Student Paper Award in the IEEE International Conference on Computer Vision and Pattern Recognition in 2010.



Jianxin Wu received the B.S. and M.S. degrees in computer science from Nanjing University, and the Ph.D. degree in computer science from the Georgia Institute of Technology. He is currently a Professor with the Department of Computer Science and Technology, Nanjing University. He was an Assistant Professor with Nanyang Technological University, Singapore. His research interests are computer vision and machine learning.