Supplementary Material of GIG

Shaohua Teng[a], Zefeng Zheng[a], Naiqi Wu[b], Luyao Teng[c,*], Wei Zhang[a], Yan Hou[a], Lunke Fei[a]

[a]*School of Computer Science and Technology, Guangdong University of Technology, Guangzhou 510006, China*
[b]*Institute of Systems Engineering, Macau University of Science and Technology, Macao 999078, China*
[c]*Faculty of Information Technology, Monash University, 20 Exhibition Walk Clayton, VIC 3800, Australia*

Introduction

This text gives the additional information about GIG. Appendix A and B give the optimization of GIG, and the proofs of the observations. Appendix C and D give the details of involved datasets and comparison methods. The experiments setting is reported in Appendix E, while the experiment results on Multi-PIE datasets are listed in Appendix F. The optimization process of ablation experiment methods is given in Appendix G, and the visualization analysis of GIG is reported in Appendix H.

Appendix A: Optimization

The objective function of GIG is

$$\begin{aligned}\min\ & \alpha tr(W_1^T X_s L_s X_s^T W_1) + \alpha\|W_1^T X_s - \ddot{Y}_s\|_F^2 + \gamma\|W_1 - V\|_F^2 \\ & + \beta\|W_2^T X_s - V^T X_s\|_F^2 + tr(W_2^T X M X^T W_2) + \lambda\|W_2\|_F^2 \\ & + \beta\|W_2^T X_t - W_3^T X_t\|_F^2 + \gamma\|W_3 - U\|_F^2 \\ & + \alpha tr(U^T X_t L_t X_t^T U)\end{aligned} \tag{1}$$

*This work is supported in part by the Key-Area Research and Development Program of Guangdong Province under Grant 2020B010166006, and by the National Natural Science Foundation of China under Grant 61972102.

**The source code of this paper is available at `https://github.com/zzf495/GIG`.

*Corresponding author
Email address: `luna.teng@mona sh.edu` (Luyao Teng)

According to Eq.(1), there are five variables, i.e., W_1, W_2, W_3, V and U that need to be optimized. We update each of them alternatively while keeping the others fixed.

Update $\mathbf{W_1}$: Fix other variables, then W_1 is updated by the following formula:

$$\begin{aligned} L_1 =& \alpha tr(W_1^T X_s L_s X_s^T W_1) + \alpha \|W_1^T X_s - \ddot{Y}_s\|_F^2 \\ &+ \gamma \|W_1 - V\|_F^2 \end{aligned} \tag{2}$$

In the first iteration, because the second and third terms are close to zero, W_1 can be solved approximately as a generalized eigen-decomposition problem by:

$$(X_s L_s X_s^T) W_1 = X_s X_s^T W_1 \xi \tag{3}$$

After that, let $\frac{\partial L_1}{\partial W_1} = 0$, the solution of W_1 is computed by:

$$W_1 = (\alpha(X_s(L_s + I_s)X_s^T) + \gamma I)^{-1}(\alpha X_s \ddot{Y}_s + \gamma V) \tag{4}$$

Update $\mathbf{V}$: Fix other variables, then V is updated by the following formula:

$$L_2 = \beta \|W_2^T X_s - V^T X_s\|_F^2 + \gamma \|W_1 - V\|_F^2 \tag{5}$$

Let $\frac{\partial L_2}{\partial V} = 0$, the solution of V is computed by:

$$V = (\beta X_t X_t^T + \gamma I)^{-1}(\beta X_t X_t^T W_2 + \gamma W_1) \tag{6}$$

Update $\mathbf{W_2}$: Fix other variables, then W_2 is updated as:

$$\begin{aligned} L_3 =& tr(W_2^T X M X^T W_2) + \beta \|W_2^T X_s - V^T X_s\|_F^2 \\ &+ \beta \|W_2^T X_t - W_3^T X_t\|_F^2 + \lambda \|W_2\|_F^2 \end{aligned} \tag{7}$$

Let $\frac{\partial L_3}{\partial W_2} = 0$, the solution of W_2 is computed by:

$$W_2 = (X M X^T + \beta G_1 + \lambda I)^{-1}(\beta G_2) \tag{8}$$

where $G_1 = X_s X_s^T + X_t X_t^T$, and $G_2 = X_s X_s^T V + X_t X_t^T W_3$.

Update $\mathbf{U}$: Fix other variables, then U is updated as:

$$L_4 = \alpha tr(U^T X_t L_t X_t^T U) + \gamma \|W_3 - U\|_F^2 \tag{9}$$

Let $\frac{\partial L_4}{\partial U} = 0$, the solution of U is solved by:

$$U = (\alpha X_t L_t X_t^T + \gamma I)^{-1}(\gamma W_3) \tag{10}$$

Update $\mathbf{W_3}$: Fix other variables, then W_3 is updated as:

$$W_3 = (\beta X_t X_t^T + \gamma I)^{-1}(\beta X_t X_t W_2 + \gamma U) \tag{11}$$

Kernelization: For nonlinear problems, denote kernel matrix $K = \phi(X)^T\phi(X) \in \mathbb{R}^{n\times n}$. Let $X = K$, $X_s = K(:, 1 : n_s)$, and $X_t = K(:, n_s + 1 : n)$, then the solutions are the same as Eqs.(4), (5), (8), (10), and (11).

Appendix B. Proof

Lemma 1. *Given a sample set X, a label regression matrix W_1, and another projected matrix W_2. Classification knowledge can be transfer from W_1 to W_2 if the following formula holds.*

$$\min \|W_1^T X - W_2^T X\| + \ell(W_1, X, Y) \tag{12}$$

Proof. Because classification information is embedded into W_1 by minimizing $\ell(W_1, X, Y)$, W can be used to classify. According to subspace pursuit, classification knowledge can be transfer from W_1 to W_2. ■

Lemma 2. *The Guidance subspace $W_1^T X_s$ can transfer classification knowledge to the Imitation subspace $W_2^T X_s$.*

Proof. Substituting Eqs.(4) and (6) into Eq.(8) in the $\mathcal{T}$-th iteration, W_2 can be computed by the following formula:

$$W_1^{\mathcal{T}} = (\alpha(X_s(L_s + I_s)X_s^T) + \gamma I)^{-1}(\alpha X_s \ddot{Y}_s + \gamma V^{\mathcal{T}-1}) \tag{13}$$

$$V^{\mathcal{T}} = (\beta X_t X_t^T + \gamma I)^{-1}(\beta X_t X_t^T W_2^{\mathcal{T}-1} + \gamma W_1^{\mathcal{T}}) \tag{14}$$

$$W_2^{\mathcal{T}} = (XMX^T + \beta G_1 + \lambda I)^{-1}(\beta X_s X_s^T V^{\mathcal{T}} + X_t X_t^T W_3^{\mathcal{T}-1}) \tag{15}$$

$W_1^{\mathcal{T}}$ is embedded with distilled label information and is transferred to $W_2^{\mathcal{T}}$ through $V^{\mathcal{T}}$. According to Lemma 1, $W_2^{\mathcal{T}}$ can achieve the classification knowledge. ■

Lemma 3. *The Imitation subspace $W_2^T X_t$ can transfer classification knowledge to the Generalization subspace $W_3^T X_t$.*

Proof. Substituting Eqs.(8) and (10) into Eq.(11) in the $\mathcal{T}$-th iteration, W_3 can be computed by the following formula:

$$U^{\mathcal{T}} = (\alpha X_t L_t X_t^T + \gamma I)^{-1}(\gamma W_3^{\mathcal{T}-1}) \tag{16}$$

$$W_3^{\mathcal{T}-1} = (\beta X_t X_t^T + \gamma I)^{-1}(\beta X_t X_t W_2^{\mathcal{T}} + \gamma U^{\mathcal{T}}) \tag{17}$$

According to Lemmas 1 and 2, the Generalization subspace $W_3^T X_t$ can acquire the classification ability. ■

Theorem 1. *Given the Guidance subspace matrix W_1, the Imitation subspace matrix W_2, the Generalization subspace matrix W_3, the source domain X_s, and the target domain X_t, GIG can transfer knowledge from the Guidance subspace $W_1^T X_s$ to the Generalization subspace $W_3^T X_t$ through the Imitation subspace $W_2^T X$.*

Proof. According to Lemmas 1 and 2, discriminative classification knowledge is transferred from W_1 to W_2. By Lemma 3, this knowledge is further transferred from W_2 to W_3. Hence, the knowledge can be transferred from the Guidance subspace $W_1^T X_s$ to the Generalization subspace $W_3^T X_t$ through Imitation subspace $W_2^T X$. ■

Corollary 1. *Given the Guidance subspace matrix W_1, intermediate subspace matrix V_1, V_2, ..., and the Generalization subspace matrix W_3, the source domain X_s, and the target domain X_t, Knowledge can be transferred from the Guidance subspace $W_1^T X_s$ to the Generalization subspace $W_3^T X_t$ through multiple intermediate subspace matrices V_1, V_2, and so on.*

Proof. According to Lemma 1, knowledge can be transferred from $W_1^T X_s$ to $V_1^T X_s$, and from $V_1^T X_s$ to $V_2^T X_s$, and so on. According to Theorem 1, it is

obvious that classification knowledge can finally be transferred from $W_1^T X_s$ to $W_3^T X_t$ through V_1, V_2, etc. ■

Corollary 2. *Under GIG framework, DLR is helpful to learn discriminative representation.*

Proof. Since projection matrices are updated alternately, chain transfer rule exists:

$$W_1^{\mathcal{T}} \rightarrow V^{\mathcal{T}+1} \rightarrow W_2^{\mathcal{T}+1} \rightarrow U^{\mathcal{T}+1} \rightarrow W_3^{\mathcal{T}+1} \tag{18}$$

Because DLR distills discriminative label information during iteration, rich label information is maintained in $W_1^{\mathcal{T}}$. By the chain transfer rule, $V^{\mathcal{T}+1}$ can learn discriminative representation based on $W_1^{\mathcal{T}}$ and $W_2^{\mathcal{T}}$. According to Theorem 1, $W_2^{\mathcal{T}+1}$ can transfer knowledge to $W_3^{\mathcal{T}+1}$. Hence, DLR is helpful to learn discriminative representation. ■

Appendix C: Involved Datasets

The experiments are conducted on a series of popular recognition datasets, including object recognition, facial recognition, digit classification, and document analysis. The details of the involved datasets are as follows.

Office-Home [1] contains 65 kinds of different objects with 30,475 original samples for training. In this paper, we follow [2] and use ResNet50 models to extract features. Four domains, i.e., Art (Ar), Clipart (Cl), Product (Pr), and Real-World (Re) with 2,048 features are obtained. The sizes of these subdomains are 2,427, 4,365, 4,439, and 4,357 with 65 categories, respectively. In the evaluation experiments, 12 cross-domain recognition tasks are conducted, e.g., Ar $\rightarrow$ Cl, Ar $\rightarrow$ Pr, ..., Re $\rightarrow$ Pr.

Multi-PIE [3] is a benchmark dataset of face recognition and includes 41,368 facial images of 68 different categories totally. In the transfer learning, five facial domains are commonly used in evaluation, including PIE5: left pose with 3,332 images, PIE7: upward pose with 1,629 images, PIE9: downward pose with 1,632 images, PIE27: front pose with 3,329 images, and PIE29: right pose with 1,632 images. The resolution of each image is 32 $\times$ 32 pixels. In the

experiments, 20 cross-domain tasks are employed, e.g., PIE5 $\rightarrow$ PIE7, PIE5 $\rightarrow$ PIE9, ..., and PIE29 $\rightarrow$ PIE27.

Office and Caltech256 [4, 5] is a popular benchmark dataset used for transfer learning. It is formed of four sub-domains, including Caltech256 (C), Amazon (A), Webcam (W), and DSLR (D), where the distributions of domains are totally different from each other. Even for the same category, their corresponding features are also dissimilar. This dataset includes 1,410 and 1,123 images of objects in 10 common categories from real-world. Moreover, it has two different datasets that are processed differently: one is formed of 800 features, called SURF [4], while the other is composed of 4,096 features, called Decaf_6 [5]. In the experiments, both datasets are involved by pairwise selection of sub-domains. Twelve pairs of cross-domain recognition tasks of two datasets are conducted in the experiments, namely C $\rightarrow$ A, C $\rightarrow$ W, C $\rightarrow$ D, A $\rightarrow$ C, ..., D $\rightarrow$ W. Besides, we also adopt 12 multi-domains transfer tasks, e.g., AC $\rightarrow$ W, AC $\rightarrow$ D, ..., DW $\rightarrow$ C to verify the effectiveness of the proposal.

USPS-MNIST [6] is a digit recognition subset with 3,800 samples randomly selected from USPS and MNIST. In fact, USPS has 7,291 training images and 2,007 test images, while MNIST has 60,000 training images and 10,000 test images. USPS-MNIST is a sub-dataset of USPS and MNIST, which is composed of 256 features and shares 10 numeral classes from zero to nine. In this paper, two cross-domain recognition tasks are organized, i.e., M $\rightarrow$ U, and U $\rightarrow$ M.

MSRC and VOC2007 [6] is an object recognition dataset. The original MSRC comprises of 4,323 images with 18 classes, while the original VOC2007 consists of 5,011 images with 20 classes. Following the experiment setting in [6], 2,799 samples are extracted from MSRC and VOC2007. MSRC (M) contains 1,269 samples with 240 features, while VOC2007 (V) contains 1,530 samples with 240 features. There are six categories, including bird, bicycle, car, cow, plane, and sheep. In the experiments, two cross-domain recognition tasks are conducted, i.e., M $\rightarrow$ V, and V $\rightarrow$ M.

COIL20 [7] is a 3D object recognition benchmark with 1,440 images of 20 categories. In fact, these images are captured by rotating five degrees at a time,

so there are 72 images with 1,024 features in each category. There are two subdomains in COIL20, i.e., COIL1 and COIL2, where COIL1 contains 1,440 images in $[0°, 85°] \cup [180°, 265°]$, and COIL2 contains 1,440 images in $[90°, 175°] \cup [270°, 365°]$. In evaluation, two cross-domain tasks are employed, e.g., C1 $\rightarrow$ C2, and C2 $\rightarrow$ C1.

Reuters-21578 is a text categorization dataset. It contains Reuters news articles from 1987 with 10,369 documents in origin. In the transfer learning, the purpose is to distinguish documents from the two top newsgroup categories: Rec and Talk. Following [8, 9], we adopt the same preprocessed way in the experiments. Three subdomains, including Orgs (Org), People (Peo), and Places (Pla), are selected in evaluation. These subdomains consist of 1,237, 1,208, and 1,016 samples with 4,771 features respectively. We perform six crossover experiments for comparison, e.g., Org $\rightarrow$ Peo, Peo $\rightarrow$ Org, ..., Pla $\rightarrow$ Peo.

Appendix D: Involved Methods

We compare GIG with several state-of-the-art traditional methods and one deep learning method. The details are as follows.

Geodesic Flow Kernel (GFK, CVPR'12) [4], which captures the geometric structure by integrating an infinite number of subspaces on the Geodesic Flow Kernel. Since GFK can automatically determine the optimal dimensionality, no hyper-parameters need to be adjusted. GFK is a continuous learning method.

Weakly-Supervised Cross-Domain Dictionary Learning (WSCDDL, IJCV'14) [10], which adopts Dictionary Learning and Sparsity Constraints to address domain adaptation problems by relaxation learning. WSCDDL aims to extract discriminative atoms from a reconstructed dictionary of source domain, and utilizes such extracted dictionary to classify target domain.

Low-rank Transfer Subspace Learning (LTSL, IJCV'14) [11], which employs a low-rank constraint on the source domain to reconstruct the target domain by one-step projection. Specifically, LTSL adopts kernel norm to ensure

the sparse properties of the reconstructed target domain. By LTSL, knowledge is transferred when source and target domains are aligned.

Scatter Component Analysis (SCA, PAMI'17) [12], which minimizes the scatter between the source and target domains. This unified framework adopts a one-step projection approach to minimize the within-class scatter and domain scatter, while maximizing the between-class scatter and total scatter.

Joint Geometrical and Statistical Alignment (JGSA, CVPR'17) [13], which learns two projections, one for the source domain and the other for the target domain, to obtain a joint discriminative representation. JGSA is a two-step projection method and it trains a classifier on the common discriminative subspace by two specific projection matrices.

Latent Elastic-Net Transfer Learning (LET, TIP'19) [14], which adopts two-step projection methods to tackle the cross-domain problems. As for LET, in the first step, it learns a common latent subspace. Based on the learned common latent subspace, LET further utilizes Soft Label Based Regression to regress target samples through source samples.

Projective Double Reconstructions (PDR, TIP'20) [15], which adopts Dictionary Learning and Soft Label Based Regression to learn a target reconstruction presentation. It introduces two relaxation matrices to relax the objective function, resulting in significant performance improvement.

Class-specific Reconstruction Transfer Learning (CRTL, TIP'20) [16], which aims to learn a class-specific reconstruction representation based on a deep learning method. Projected Hilbert-Schmidt Independency Criterion (pHSIC) is proposed by CRTL to exploit the latent relation between features and labels.

Appendix E. Experiment Setting

As for GIG, we have seven hyper-parameters that should be tuned, i.e., tradeoff parameter μ, α, β, γ, λ temperature parameter T_e, and K-nearest neighbor k. For simplicity, we fix $k = 1$ for USPS-MNIST and MSRC-VOOC2007,

Table 1: The best parameters of datasets

Dataset	δ_s	δ_t	α	β	γ
Office-Home	5	5	10^{-3}	10	10^{-5}
Multi-PIE	5	5	10^{5}	10^{-2}	10
Office+Caltech256(SURF)	20/8	8	10^{2}	10^{2}	10
Office+Caltech256(Multi vs Single)	20/8	8	10^{2}	10^{2}	10
Office+Caltech256(Decaf_6)	20/8	8	10^{3}	10	10
USPS-MNIST	5	5	10^{5}	10^{-5}	10^{-5}
MSRC-VOC2007	10	10	10^{5}	10^{-5}	10^{-5}
COIL20	5	5	10^{2}	10^{2}	10
Reuters-21578	5	5	10^{5}	10^{2}	10^{2}

and $k = 7$ for other datasets. We fix the tradeoff parameter $\mu = 0.1$. Further, α, β, and γ are grid-searched in 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^{2}, 10^{3}, 10^{4}, 10^{5}, while λ is fixed as 10^{3}. The temperature T_e is fixed as 10^{2}, and reduced to 1 during iteration. The optimal parameters of GIG are listed in Table 1

Appendix F: The Results of Multi-PIE

Table 2 shows that GIG achieves the second-best performance compared with the seven state-of-the-art cross-domain recognition methods, with 0.78% less than PDR. Besides, GIG achieves 93.94% classification accuracy, and 19 second-best performances on 20 facial recognition tasks. The results verify the efficiency of GIG.

Appendix G: Optimization of the Ablation Experiment Methods

For GIG, we have:

$$W_1 = (\alpha(X_s(L_s + I_s)X_s^T) + \gamma I)^{-1}(\alpha X_s \ddot{Y}_s + \gamma V) \tag{19}$$

Table 2: Classification accuracies (%) on MULTI-PIE. The best results are highlighted in bold black letters, with the second-placed competitor underlined

Source	Target	GFK	WSCDLL	LTSL	SCA	JGSA	PDR	CRTL	GIG
PIE5	PIE7	62.62±1.76	91.93±1.38	84.51±2.45	75.60±1.02	81.67±1.50	**94.20±0.59**	88.96±0.75	93.13±0.77
PIE5	PIE9	63.71±1.56	91.87±0.74	82.33±2.57	68.67±1.13	78.56±1.45	**93.40±0.53**	87.20±0.69	93.15±0.67
PIE5	PIE27	63.40±1.15	91.56±0.80	85.64±2.39	75.32±1.40	81.96±1.65	**96.56±0.45**	90.99±0.74	95.65±0.53
PIE5	PIE29	65.73±1.12	88.16±0.64	73.46±2.90	74.68±1.02	76.09±1.18	**92.18±0.69**	86.67±0.81	91.16±0.76
PIE7	PIE5	63.25±1.01	92.95±0.61	82.20±2.60	77.60±1.56	80.76±1.54	**96.76±0.55**	90.90±0.56	95.70±0.71
PIE7	PIE9	67.21±1.55	91.98±0.70	81.18±2.44	83.22±1.11	82.21±1.93	**94.02±0.56**	89.73±0.62	92.46±0.64
PIE7	PIE27	67.65±1.27	91.87±0.65	86.41±2.64	79.60±1.34	80.36±1.77	**95.61±0.59**	89.63±0.36	95.55±0.47
PIE7	PIE29	65.77±1.54	88.93±0.35	72.69±2.76	74.45±0.86	70.25±2.16	**92.34±0.66**	86.92±0.55	91.11±0.80
PIE9	PIE5	63.42±1.27	93.89±0.84	74.38±2.52	69.97±1.14	76.85±1.49	**96.45±0.71**	91.19±0.76	95.84±0.41
PIE9	PIE7	64.88±1.57	90.93±0.33	77.48±2.18	81.55±1.49	77.63±1.83	**93.17±0.75**	88.70±0.43	93.05±0.89
PIE9	PIE27	69.37±0.90	92.17±0.43	80.63±2.22	78.96±0.97	77.60±1.10	**96.56±0.48**	91.12±0.52	96.04±0.41
PIE9	PIE29	66.27±1.88	88.80±0.29	66.47±3.31	73.21±1.63	72.24±2.06	**92.70±0.60**	86.54±0.41	91.59±0.81
PIE27	PIE5	64.89±1.42	92.78±0.60	86.18±2.76	82.62±1.54	86.50±1.63	**96.92±0.30**	92.65±0.45	96.03±0.48
PIE27	PIE7	68.59±1.38	91.69±0.63	82.12±2.34	79.54±0.88	79.69±1.58	94.25±0.47	89.48±0.71	**94.57±0.30**
PIE27	PIE9	70.67±1.08	92.00±0.70	80.60±2.23	81.25±0.82	82.29±1.07	**95.42±0.96**	90.91±0.40	94.21±0.53
PIE27	PIE29	67.25±1.07	89.79±0.27	79.82±2.22	80.80±0.87	82.60±1.13	**93.52±0.53**	86.68±0.96	92.34±0.67
PIE29	PIE5	62.08±1.08	89.98±0.94	72.51±2.62	75.54±0.76	79.98±1.54	**96.45±0.58**	89.34±0.67	95.32±0.55
PIE29	PIE7	62.32±1.48	90.49±0.54	74.86±2.64	76.93±1.28	77.76±1.31	**93.95±0.63**	89.16±0.52	93.20±0.83
PIE29	PIE9	64.21±1.28	91.18±0.60	72.14±2.90	74.35±2.04	76.12±1.36	**94.25±0.62**	87.76±0.66	92.90±0.55
PIE29	PIE27	61.44±1.15	92.39±0.68	77.75±2.06	80.12±0.84	81.25±1.15	95.65±0.44	91.58±0.43	**95.72±0.43**
AVERAGE		65.24±1.33	91.27±0.64	78.67±2.54	77.20±1.19	79.12±1.52	**94.72±0.58**	89.31±0.60	93.94±0.61

$$W_2 = (XMX^T + \beta G_1 + \lambda I)^{-1}(\beta G_2) \tag{20}$$

$$W_3 = (\beta X_t X_t^T + \gamma I)^{-1}(\beta X_t X_t W_2 + \gamma U) \tag{21}$$

The optimizations of the variants are as follows.

GIG_{RDLR}, which removes the DLR term of GIG to verify its effectiveness. The objective function of GIG becomes

$$\begin{aligned} \min\ & \alpha tr(W_1^T X_s L_s X_s^T W_1) + \gamma \|W_1 - V\|_F^2 \\ & + \beta \|W_2^T X_s - V^T X_s\|_F^2 + tr(W_2^T XMX^T W_2) + \lambda \|W_2\|_F^2 \\ & + \beta \|W_2^T X_t - W_3^T X_t\|_F^2 + \gamma \|W_3 - U\|_F^2 \\ & + \alpha tr(U^T X_t L_t X_t^T U) \end{aligned} \tag{22}$$

Here, W_2 and W_3 are obtained by Eqs.(20) and (21) while W_1 is obtained by

$$W_1 = (\alpha X_s L_s X_s^T + \gamma I)^{-1}(\gamma V) \tag{23}$$

GIG_{RDLR} + LR, which replaces DLR with LR. The objective function becomes:

$$\begin{aligned}\min\ & \alpha tr(W_1^T X_s L_s X_s^T W_1) + \alpha\|W_1^T X_s - \hat{Y}_s\|_F^2 + \gamma\|W_1 - V\|_F^2 \\ & + \beta\|W_2^T X_s - V^T X_s\|_F^2 + tr(W_2^T X M X^T W_2) + \lambda\|W_2\|_F^2 \\ & + \beta\|W_2^T X_t - W_3^T X_t\|_F^2 + \gamma\|W_3 - U\|_F^2 \\ & + \alpha tr(U^T X_t L_t X_t^T U)\end{aligned} \tag{24}$$

Here, W_2 and W_3 are obtained by Eqs.(20) and (21) while W_1 is obtained by

$$W_1 = (\alpha(X_s(L_s + I_s)X_s^T) + \gamma I)^{-1}(\alpha X_s \hat{Y}_s + \gamma V) \tag{25}$$

GIG_{RW}, which removes the Guidance subspace learning to verify its effectiveness, then the overall solution becomes:

$$\begin{aligned}\min\ & \alpha tr(W_2^T X_s L_s X_s^T W_2) + \alpha\|W_2^T X_s - \ddot{Y}_s\|_F^2 \\ & + tr(W_2^T X M X^T W_2) + \lambda\|W_2\|_F^2 \\ & + \beta\|W_2^T X_t - W_3^T X_t\|_F^2 + \gamma\|W_3 - U\|_F^2 \\ & + \alpha tr(U^T X_t L_t X_t^T U)\end{aligned} \tag{26}$$

Here, W_3 is obtained by Eq.(21) while W_2 is obtained by

$$W_2 = G_1^{-1}(\beta X_t X_t^T W_3 + \gamma X_s \ddot{Y}_s) \tag{27}$$

where $G_1 = \alpha X_s(L_s + I_s)X_s^T + \beta X_t X_t^T + XMX^T + \lambda I$.

GIG_{RA} , which removes the Generalization subspace learning, and the solution becomes to solve P and train classifier in the Imitation subspace. The objective function is re-expressed as:

$$\begin{aligned}\min\ & \alpha tr(W_1^T X_s L_s X_s^T W_1) + \alpha\|W_1^T X_s - \ddot{Y}_s\|_F^2 + \gamma\|W_1 - V\|_F^2 \\ & + \beta\|W_2^T X_s - V^T X_s\|_F^2 + tr(W_2^T X M X^T W_2) \\ & + \alpha tr(W_2^T X_t L_t X_t^T W_2)\end{aligned} \tag{28}$$

Here, the solution of W_2 is

$$W_2 = G_2^{-1}(\beta X_s X_s^T V) \tag{29}$$

where $G_2 = \alpha X_t L_t X_t^T + \beta X_s X_s^T + XMX^T + \lambda I$.

GIG_{RAW}, which removes the Guidance subspace $W_1^T X_s$ and the Generalization subspace $W_3^T X_t$. the objective function becomes to one-step projection learning as:

$$\begin{aligned} \min\, & tr(W_2^T XMX^T W_2) + \alpha tr(W_2^T X_s L_s X_s^T W_2) \\ & + \alpha \|W_2^T X_s - \ddot{Y}_s\|_F^2 + \alpha tr(W_2^T X_t L_t X_t^T W_2) \end{aligned} \tag{30}$$

Here, the solution of W_2 is

$$W_2 = G_3^{-1}(\alpha X_s \ddot{Y}_s) \tag{31}$$

where $G_3 = XMX^T + \alpha(X_s(L_s + I_s)X_s^T + X_t L_t X_t^T) + \lambda I$.

Appendix H: Visualization analysis

To further verify the effectiveness of GIG, we visualize the source and target domain samples by t-SNE [17] on tasks V $\rightarrow$ M. The visualizations of the original domain are shown in Figs. 1 (a) and (b). The source domain is composed of 120 samples (20 samples each class) by randomly selecting from M and V. The unlabeled target domain is composed of the rest samples of M. We adopt GIG and report the final results as shown in Figs. 1 (c) to (e). From Fig. 1 (c), DLR works well in the source domain, since all of the source samples can be correctly tagged in the source domain by supervised spectral clustering. As shown in Fig. 1 (d), by Imitation subspace learning, classification knowledge is transferred from the Guidance subspace to the Imitation subspace, and the discrepancies of domains are reduced. At last, domain-invariant features are achieved in the Imitation subspace. Through transitive learning, the Generalization subspace further learns geometric structure in the target domain, and generalizes the classification knowledge in the Imitation subspace, referred to Fig. 1 (e). By learning the three subspaces step by step, discriminative knowledge-transferable features are achieved in the Generalization subspace, thus significant performance is achieved. By above observations, we can confirm

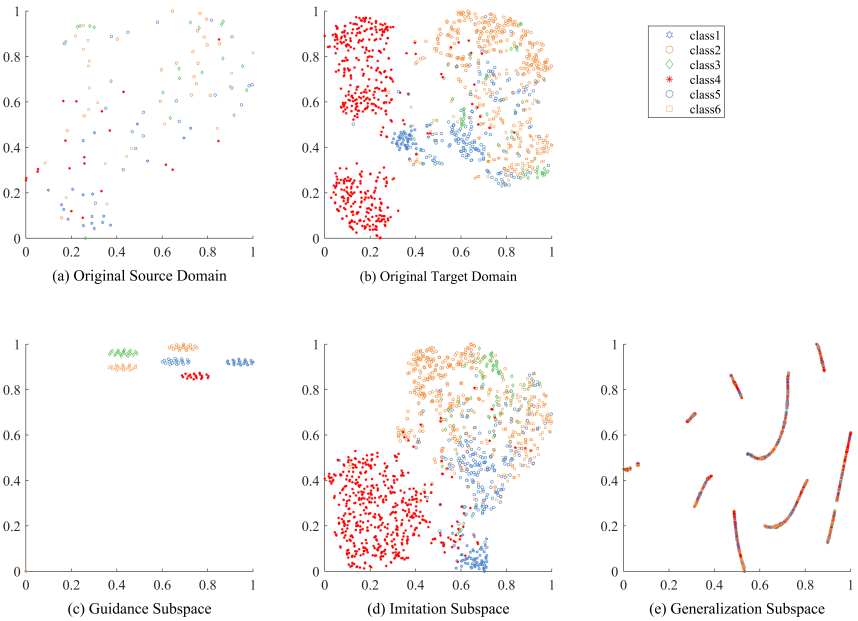

Figure 1: Visualization of GIG on tasks V to M of MSRC-VOC2007 datasets. (a) original source domain; (b) original target domain; (c) Guidance subspace (projected source subspace); (d) Imitation subspace (shared subspace); (e) Generalization subspace (projected target subspace).

that GIG can achieve remarkable performance through both transitive learning and relaxed learning.

References

[1] H. Venkateswara, J. Eusebio, S. Chakraborty, S. Panchanathan, Deep hashing network for unsupervised domain adaptation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 5018–5027.

[2] J. Liang, R. He, Z. Sun, T. Tan, Aggregating randomized clustering-

promoting invariant projections for domain adaptation, IEEE transactions on pattern analysis and machine intelligence 41 (5) (2018) 1027–1042.

[3] L. Zhang, W. Zuo, D. Zhang, Lsdt: Latent sparse domain transfer learning for visual adaptation, IEEE Transactions on Image Processing 25 (3) (2016) 1177–1191.

[4] B. Gong, Y. Shi, F. Sha, K. Grauman, Geodesic flow kernel for unsupervised domain adaptation, in: 2012 IEEE conference on computer vision and pattern recognition, IEEE, 2012, pp. 2066–2073.

[5] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, T. Darrell, Decaf: A deep convolutional activation feature for generic visual recognition, in: International conference on machine learning, PMLR, 2014, pp. 647–655.

[6] M. Long, J. Wang, G. Ding, J. Sun, P. S. Yu, Transfer feature learning with joint distribution adaptation, in: Proceedings of the IEEE international conference on computer vision, 2013, pp. 2200–2207.

[7] S. A. Nene, S. K. Nayar, H. Murase, et al., Columbia object image library (coil-100).

[8] W. Dai, G.-R. Xue, Q. Yang, Y. Yu, Co-clustering based classification for out-of-domain documents, in: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, 2007, pp. 210–219.

[9] J. Gao, W. Fan, J. Jiang, J. Han, Knowledge transfer via multiple model local structure mapping, in: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, 2008, pp. 283–291.

[10] F. Zhu, L. Shao, Weakly-supervised cross-domain dictionary learning for visual recognition, International Journal of Computer Vision 109 (1-2) (2014) 42–59.

[11] M. Shao, D. Kit, Y. Fu, Generalized transfer subspace learning through low-rank constraint, International Journal of Computer Vision 109 (1-2) (2014) 74–93.

[12] M. Ghifary, D. Balduzzi, W. B. Kleijn, M. Zhang, Scatter component analysis: A unified framework for domain adaptation and domain generalization, IEEE transactions on pattern analysis and machine intelligence 39 (7) (2016) 1414–1430.

[13] J. Zhang, W. Li, P. Ogunbona, Joint geometrical and statistical alignment for visual domain adaptation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1859–1867.

[14] N. Han, J. Wu, X. Fang, S. Xie, S. Zhan, K. Xie, X. Li, Latent elastic-net transfer learning, IEEE Transactions on Image Processing 29 (2019) 2820–2833.

[15] N. Han, J. Wu, X. Fang, S. Teng, G. Zhou, S. Xie, X. Li, Projective double reconstructions based dictionary learning algorithm for cross-domain recognition, IEEE Transactions on Image Processing 29 (2020) 9220–9233.

[16] S. Wang, L. Zhang, W. Zuo, B. Zhang, Class-specific reconstruction transfer learning for visual recognition across domains, IEEE Transactions on Image Processing 29 (2019) 2424–2438.

[17] L. Van der Maaten, G. Hinton, Visualizing data using t-sne., Journal of machine learning research 9 (11).