TFGDA: Exploring Topology and Feature Alignment in Semi-supervised Graph Domain Adaptation through Robust Clustering

Jun Dan[1] Weiming Liu[1] Chunfeng Xie[2] Hua Yu[3]

Shunjie Dong[4] Yanchao Tan[5,6,7]

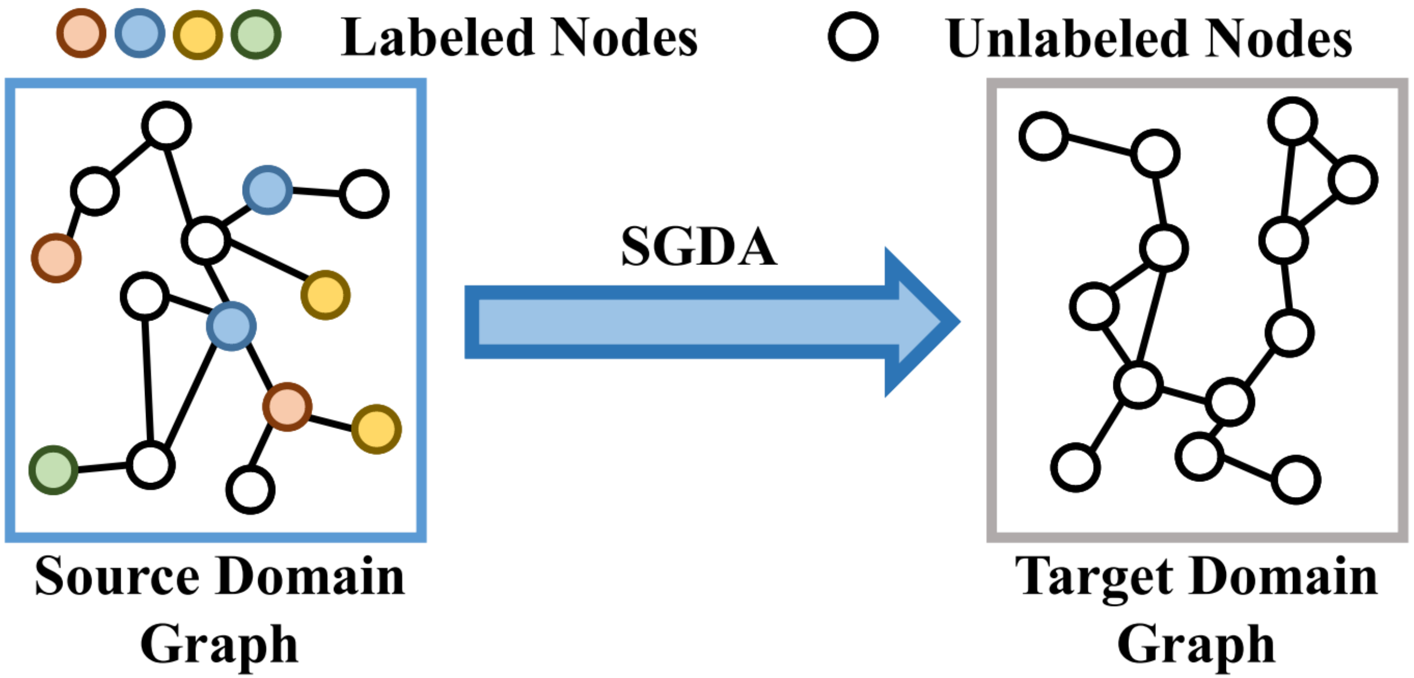
[1]Zhejiang University [2]Queen Mary University of London [3]Dalian University of Technology

[4]Shanghai Jiao Tong University [5]Fuzhou University

[6]Research Center of Big Data Intelligence, Ministry of Education

[7]Fujian Key Laboratory of Network Computing and Intelligent Information Processing

Background Problem



Semi-supervised Graph Domain Adaptation (**SGDA**), as a subfield of graph transfer learning, seeks to precisely annotate unlabeled target graph nodes by leveraging transferable features acquired from the limited labeled source nodes.

Challenges

Existing Semi-Supervised Graph Domain Adaptation (SGDA) models mainly face three challenges:

CH1: Current GCNs-based methods typically mine the graph structure information in the deep feature space, which is a suboptimal way because the graph structure information may be lost or destroyed after passing through the GCNs-based feature extractors.

CH2: Existing GDA methods often employ adversarial learning to reduce domain discrepancy. However, adversarial training is an unstable process that may destroy the category information hidden in node features.

CH3: Given the limited number of labeled nodes in the source domain graph, the model is prone to overfitting when only relying on the source domain classification loss for optimization.

Methodology

CH1: Mining the graph structure information of graphs in the feature space is a suboptimal operation because the node features extracted by the GCNs-based feature extractor have already lost some structure information.

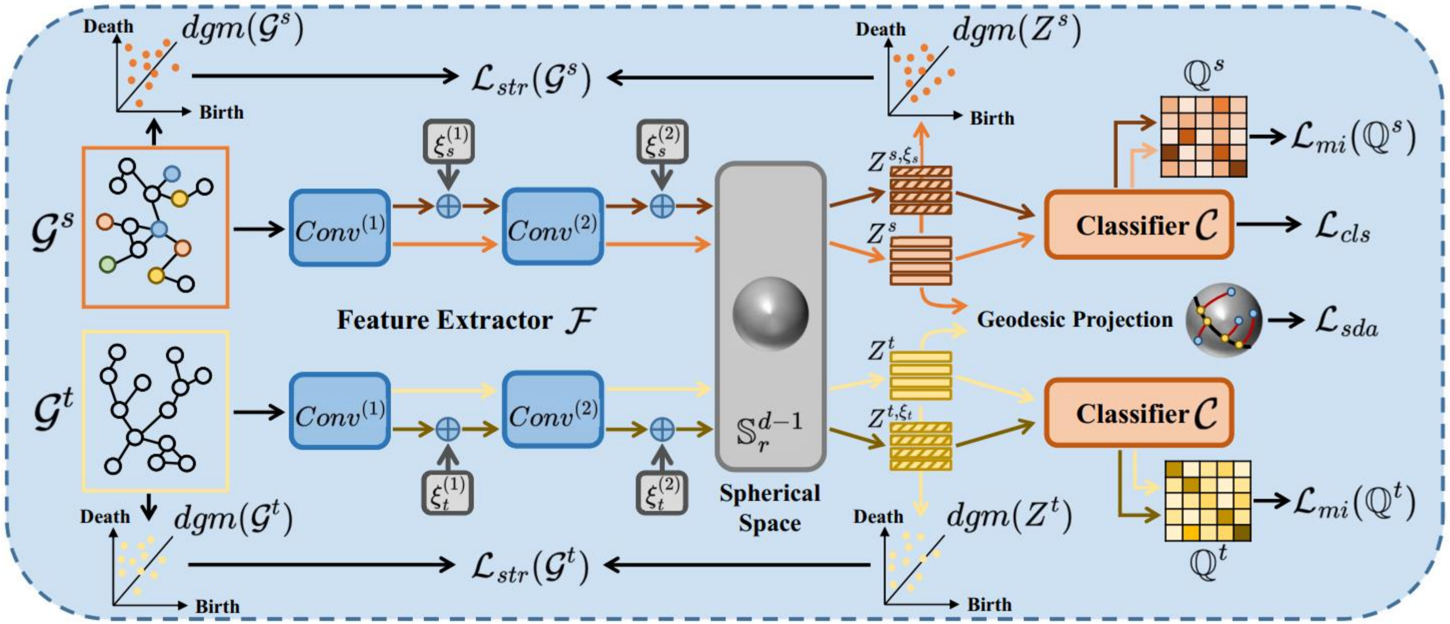
Step 1: Sample multiple subgraphs

$$\left\{\hat{\mathcal{G}}_1, \hat{\mathcal{G}}_2, \cdots, \hat{\mathcal{G}}_a\right\}$$

Step 2: Compute the topological features of subgraphs separately in the input space and the hidden layer space.

$$dgm(\hat{\mathcal{G}}_i) \qquad dgm(\hat{Z}_i)$$

Subgraph Topological Structure Alignment



Step 3: Align the topological structures of the input and latent spaces.

$$\mathcal{L}_{str}^{sub}(\hat{\mathcal{G}}_i) = \sum_{(\alpha,\beta)\in\gamma^*} \|\alpha - \beta\|_2^2 .$$

Step 4: Aggregate the topological discrepancy of multiple subgraphs and generalize across two domains.

$$\mathcal{L}_{str}(\mathcal{G}) = \sum_i \mathcal{L}_{str}^{sub}(\hat{\mathcal{G}}_i) \quad \mathcal{L}_{stsa} = \mathcal{L}_{str}(\mathcal{G}^s) + \mathcal{L}_{str}(\mathcal{G}^t).$$

Methodology

CH2: Using adversarial learning to align feature distributions across domains is an unstable process that may destory the class information hidden in node features.

Step 1: Map node features to spherical space

$$\mathbb{S}_r^{d-1} = \left\{z_i \in \mathbb{R}^d : \|z_i\|_2 = r\right\}$$

Step 2: Use geodesic projection to map spherical features onto multiple circles

$$I^U(z) = U^\top \underset{g \in \mathrm{span}(UU^\top)\cap \mathbb{S}_r^{d-1}}{\arg\min} \mathcal{M}_{\mathbb{S}_r^{d-1}}(z, g)$$

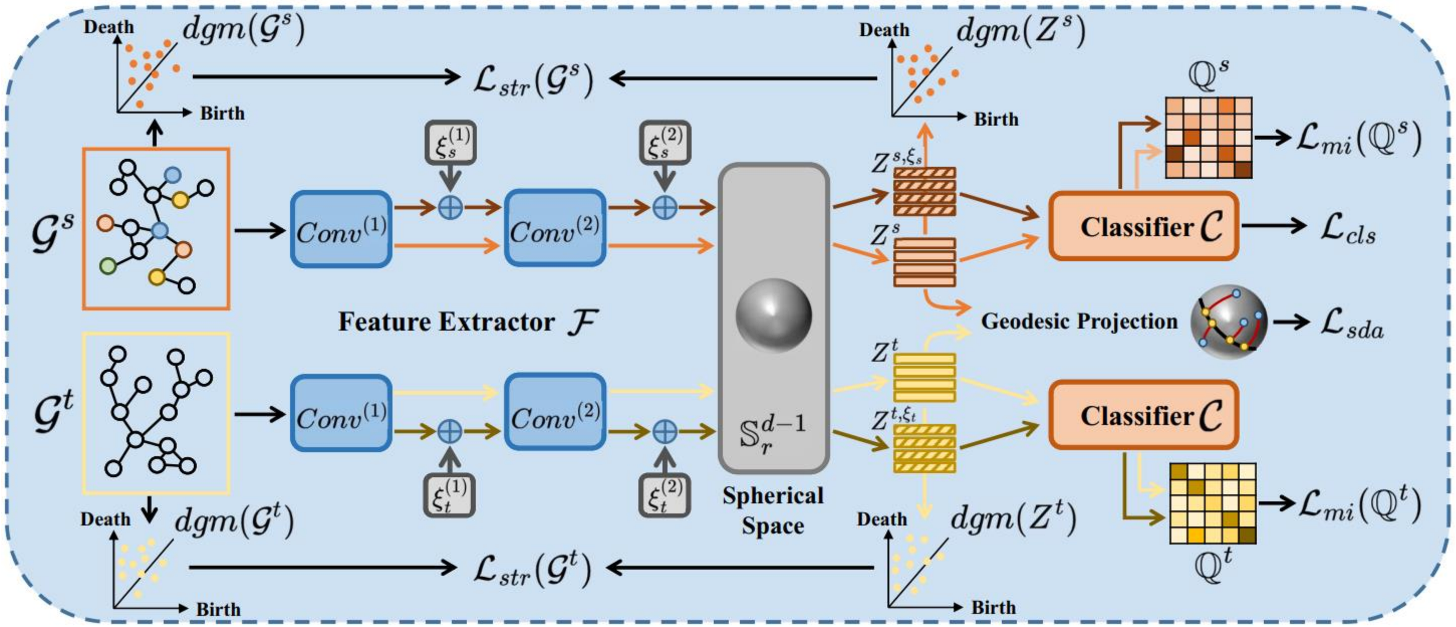
Step 3: utilize the spherical sliced-Wasserstein (SSW) distance to measure domain discrepancy

$$SSW_q^q(u, v) = \int_{\mathbb{V}_{d,2}} \mathcal{W}_q^q(I^U_{\#}\mu, I^U_{\#}\nu)\mathrm{d}\sigma(\mathrm{U})$$

Step 4: Minimize the domain discrepancy during training

$$\mathcal{L}_{sda} = SSW_p^p(\tilde{\mu}, \tilde{\nu}) \approx \frac{1}{b}\sum_{m=1}^{b} \mathcal{W}_q^q(\tilde{\mu}, \tilde{\nu})$$

Sphere-guided Domain Alignment



Methodology

CH3: The source domain graph has too few labeled nodes, making the model prone to overfitting issues.

Step 1: introduces shift parameters to perturb source and target domains node features

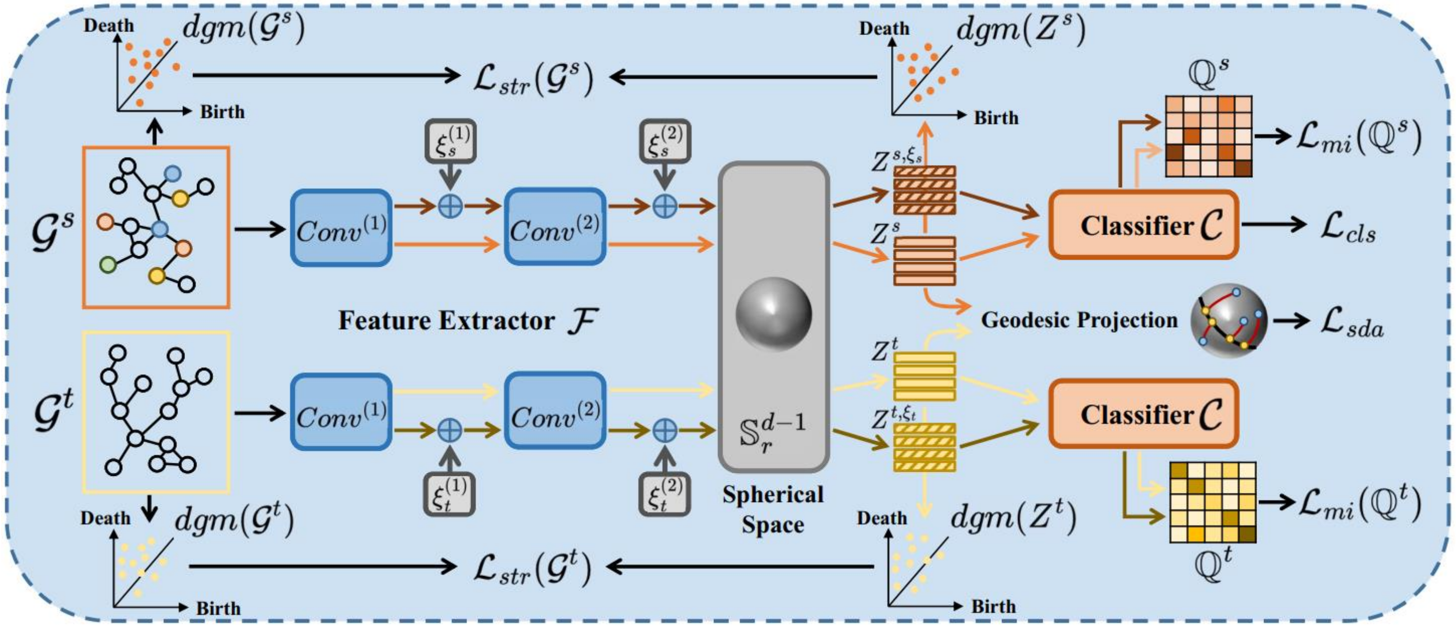
$$H^{s/t,\xi_{s/t},(l)} = \begin{cases} Conv^{(l)}(P^{s/t}, X^{s/t}) + \xi^{(l)}_{s/t}, & l = 1 \\ Conv^{(l)}(P^{s/t}, H^{s/t,(l-1)}) + \xi^{(l)}_{s/t}, & 1 < l \leq L \end{cases}$$

Step 2: Maximize the mutual information between the predictions of the original node features and the perturbed node features

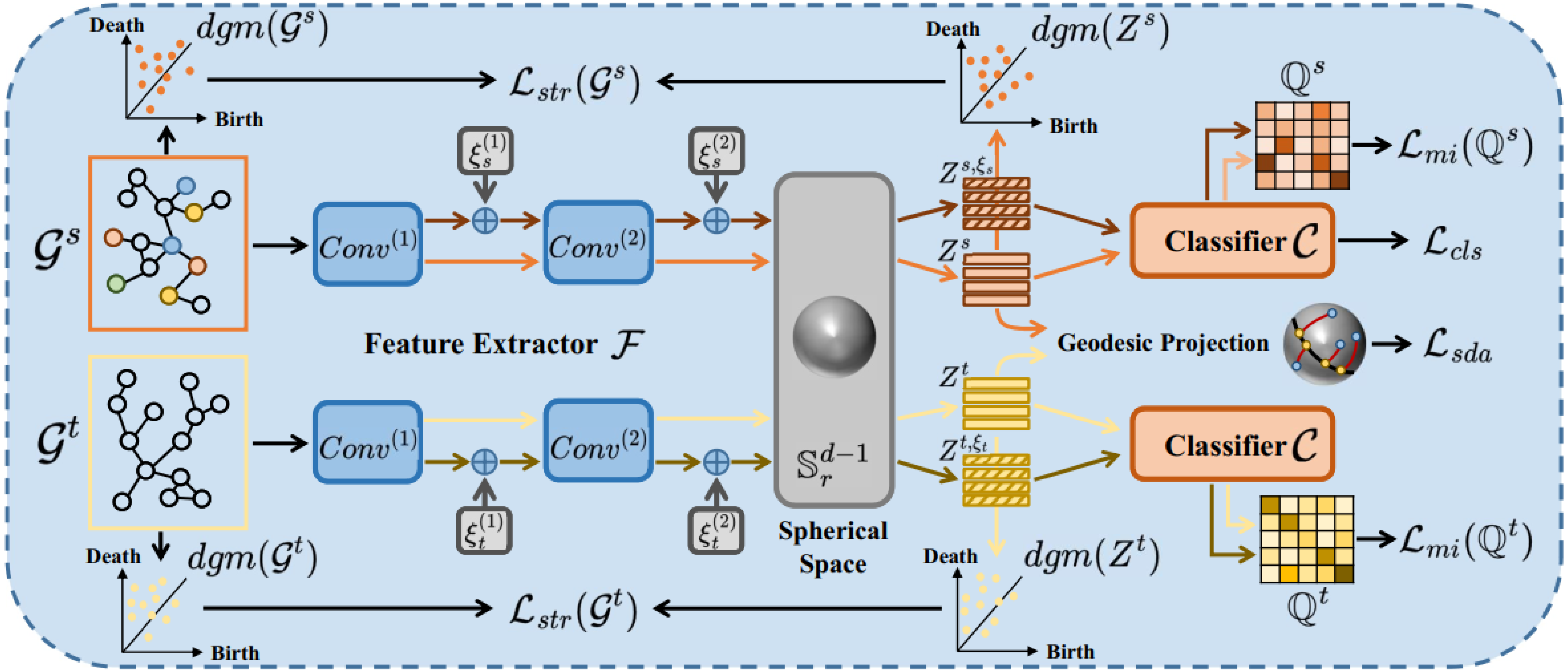
$$\mathbb{Q} = \frac{1}{|\mathcal{V}|} \sum_{i=1}^{|\mathcal{V}|} \Psi(z_i) \cdot \Psi(z_i^{\xi})^{\top}$$

$$\mathcal{L}_{mi}(\mathbb{Q}) = \sum_{k=1}^{K} \sum_{k^{\xi}=1}^{K} \mathbb{Q}_{kk^{\xi}} \cdot \ln \frac{\mathbb{Q}_{kk^{\xi}}}{\mathbb{Q}_k \cdot \mathbb{Q}_{k^{\xi}}}, s.t., ||\xi^{(l)}||_F \leq \varpi, \forall \xi^{(l)} \in \xi.$$

Robustness-guided Node Clustering



Model Optimization



The overall objective of TFGDA:

$$\min_{\mathcal{F},\mathcal{C},\xi^s,\xi^t} \mathcal{L}_{cls} + \eta\mathcal{L}_{sda} + \varepsilon\mathcal{L}_{stsa} - \lambda\mathcal{L}_{rnc}$$

$$s.t., ||\xi^{s,(l)}||_F \leq \varpi, \forall\xi^{s,(l)} \in \xi^s, ||\xi^{t,(l)}||_F \leq \varpi, \forall\xi^{t,(l)} \in \xi^t.$$

Experiments

Datasets：

We run experiments on three real-world graphs.

Six typical transfer tasks are considered in our experiments: **A→C, A→D, C→A, C→D, D→A** and **D→C**.

Table 1: The statistics of three real-world graphs. Note that '#' means 'the number of'. 'Attr.' refer to 'Attributes'. 'Avg.' represents 'Average'.

Graph	#Nodes	#Edges	#Attr.	Avg. Degree	Label Proportion (%)
ACMv9 (**A**)	9,360	15,602	5,571	1.667	20.5/29.6/22.5/8.6/18.8
Citationv1 (**C**)	8,935	15,113	5,379	1.691	25.3/26.0/22.5/7.7/18.5
DBLPv7 (**D**)	5,484	8,130	4,412	1.482	21.7/33.0/23.8/6.0/15.5

Experiments

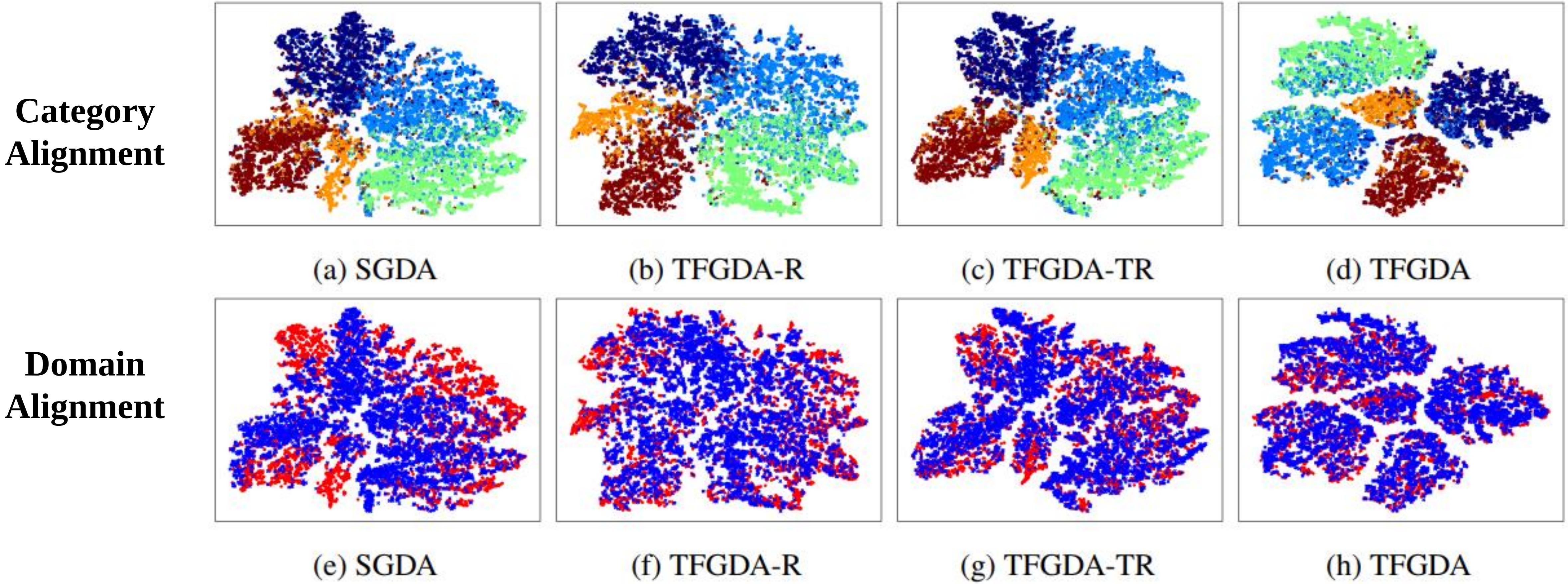
Table 1: Transfer performance (%) on six transfer tasks with a source graph label rate of 5% for semi-supervised graph domain adaptation.

Methods	A→C		A→D		C→A		C→D		D→A		D→C	
	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1
MLP [2]	41.3$_{\pm1.15}$	35.8$_{\pm0.72}$	42.8$_{\pm0.88}$	36.3$_{\pm0.77}$	39.4$_{\pm0.57}$	33.7$_{\pm0.58}$	43.7$_{\pm0.69}$	36.7$_{\pm0.55}$	37.3$_{\pm0.32}$	30.8$_{\pm0.37}$	39.4$_{\pm0.99}$	32.8$_{\pm0.99}$
GCN [28]	54.4$_{\pm1.52}$	52.0$_{\pm1.62}$	56.9$_{\pm2.33}$	53.4$_{\pm2.81}$	54.1$_{\pm1.40}$	52.3$_{\pm1.98}$	58.9$_{\pm0.99}$	54.5$_{\pm1.55}$	50.1$_{\pm2.14}$	48.0$_{\pm3.28}$	56.0$_{\pm1.24}$	51.9$_{\pm1.49}$
GSAGE [29]	49.3$_{\pm2.18}$	46.4$_{\pm2.06}$	51.8$_{\pm1.35}$	47.4$_{\pm1.62}$	46.8$_{\pm2.56}$	45.0$_{\pm2.78}$	51.7$_{\pm1.95}$	48.1$_{\pm1.97}$	41.7$_{\pm2.17}$	37.4$_{\pm4.59}$	45.4$_{\pm2.11}$	39.3$_{\pm3.45}$
GAT [30]	55.1$_{\pm3.22}$	50.8$_{\pm1.45}$	55.3$_{\pm2.52}$	51.8$_{\pm2.60}$	50.0$_{\pm1.20}$	45.6$_{\pm2.36}$	55.4$_{\pm2.73}$	49.2$_{\pm2.59}$	44.8$_{\pm2.74}$	38.3$_{\pm4.84}$	50.4$_{\pm3.35}$	42.0$_{\pm4.46}$
GIN [80]	64.6$_{\pm2.47}$	56.0$_{\pm2.73}$	60.0$_{\pm2.09}$	51.3$_{\pm3.99}$	57.1$_{\pm1.19}$	54.4$_{\pm2.57}$	62.0$_{\pm1.05}$	56.8$_{\pm1.40}$	51.9$_{\pm2.00}$	45.4$_{\pm2.16}$	60.2$_{\pm3.05}$	53.0$_{\pm2.10}$
DANN [81]	44.3$_{\pm2.03}$	39.3$_{\pm1.86}$	44.0$_{\pm1.42}$	38.7$_{\pm1.47}$	41.8$_{\pm1.95}$	37.6$_{\pm1.24}$	45.5$_{\pm0.71}$	39.6$_{\pm1.55}$	37.8$_{\pm3.66}$	33.2$_{\pm2.23}$	41.7$_{\pm2.32}$	35.6$_{\pm2.55}$
CDAN [11]	44.6$_{\pm1.30}$	38.6$_{\pm1.07}$	45.5$_{\pm0.85}$	38.0$_{\pm0.86}$	42.4$_{\pm0.64}$	36.2$_{\pm1.17}$	46.7$_{\pm1.17}$	39.2$_{\pm0.96}$	39.0$_{\pm1.08}$	32.3$_{\pm1.09}$	41.7$_{\pm1.55}$	34.8$_{\pm1.56}$
DANN$_{GCN}$ [2]	63.0$_{\pm6.75}$	59.6$_{\pm6.02}$	62.2$_{\pm1.90}$	57.7$_{\pm3.16}$	56.7$_{\pm0.38}$	55.2$_{\pm1.03}$	65.3$_{\pm2.04}$	59.0$_{\pm2.39}$	52.3$_{\pm2.59}$	48.6$_{\pm4.52}$	58.1$_{\pm2.78}$	52.4$_{\pm3.81}$
CDAN$_{GCN}$ [2]	70.3$_{\pm0.84}$	66.5$_{\pm0.66}$	65.0$_{\pm1.00}$	61.3$_{\pm0.96}$	56.3$_{\pm1.78}$	53.6$_{\pm2.70}$	65.2$_{\pm2.19}$	58.8$_{\pm2.38}$	53.0$_{\pm1.34}$	48.7$_{\pm3.51}$	59.0$_{\pm1.52}$	53.3$_{\pm1.99}$
UDA-GCN [70]	72.4$_{\pm2.75}$	65.2$_{\pm6.51}$	68.0$_{\pm6.38}$	64.3$_{\pm7.12}$	62.9$_{\pm0.33}$	62.2$_{\pm1.44}$	71.4$_{\pm2.56}$	67.5$_{\pm2.25}$	55.8$_{\pm3.50}$	52.4$_{\pm2.68}$	65.2$_{\pm4.41}$	60.7$_{\pm6.84}$
AdaGCN [1]	70.8$_{\pm0.95}$	68.5$_{\pm0.73}$	68.2$_{\pm3.84}$	64.2$_{\pm3.91}$	61.5$_{\pm2.20}$	60.4$_{\pm3.15}$	69.1$_{\pm1.96}$	65.8$_{\pm2.87}$	56.1$_{\pm1.75}$	53.8$_{\pm2.95}$	64.1$_{\pm0.91}$	62.8$_{\pm1.56}$
CoCo [25]	72.7$_{\pm1.36}$	66.8$_{\pm1.15}$	68.3$_{\pm2.31}$	64.1$_{\pm2.68}$	62.7$_{\pm0.95}$	61.5$_{\pm1.18}$	71.6$_{\pm1.76}$	67.3$_{\pm1.93}$	56.7$_{\pm1.47}$	54.1$_{\pm1.29}$	66.0$_{\pm0.88}$	64.4$_{\pm1.13}$
StruRW [21]	72.9$_{\pm1.21}$	67.1$_{\pm1.07}$	68.5$_{\pm0.94}$	64.4$_{\pm1.03}$	63.6$_{\pm1.05}$	61.9$_{\pm1.19}$	71.8$_{\pm2.06}$	67.6$_{\pm2.45}$	57.0$_{\pm1.72}$	54.2$_{\pm1.38}$	65.7$_{\pm0.96}$	63.1$_{\pm1.25}$
SGDA [2]	75.6$_{\pm0.57}$	71.4$_{\pm0.82}$	69.2$_{\pm0.73}$	64.7$_{\pm2.36}$	66.3$_{\pm0.68}$	62.3$_{\pm0.96}$	72.9$_{\pm1.26}$	68.9$_{\pm1.83}$	60.6$_{\pm0.86}$	56.0$_{\pm0.90}$	73.2$_{\pm0.59}$	69.3$_{\pm1.01}$
TFGDA-S	55.8$_{\pm1.76}$	53.6$_{\pm1.84}$	54.2$_{\pm2.11}$	44.9$_{\pm2.04}$	58.2$_{\pm1.52}$	48.9$_{\pm1.94}$	57.0$_{\pm1.03}$	46.3$_{\pm1.60}$	49.8$_{\pm2.33}$	41.0$_{\pm3.46}$	55.9$_{\pm1.41}$	45.2$_{\pm1.72}$
TFGDA-T	72.6$_{\pm0.55}$	66.3$_{\pm0.83}$	65.9$_{\pm0.85}$	62.4$_{\pm1.39}$	64.3$_{\pm0.88}$	61.8$_{\pm0.76}$	65.6$_{\pm0.97}$	54.7$_{\pm1.24}$	56.2$_{\pm0.78}$	53.4$_{\pm0.87}$	68.5$_{\pm0.44}$	67.4$_{\pm0.92}$
TFGDA-D	75.8$_{\pm0.38}$	70.7$_{\pm0.69}$	71.2$_{\pm0.61}$	67.5$_{\pm1.15}$	68.5$_{\pm0.57}$	63.7$_{\pm1.02}$	72.2$_{\pm0.88}$	68.1$_{\pm1.13}$	63.1$_{\pm0.74}$	58.5$_{\pm0.79}$	73.4$_{\pm0.51}$	71.1$_{\pm0.88}$
TFGDA-R	74.4$_{\pm0.46}$	70.1$_{\pm0.77}$	68.8$_{\pm0.54}$	64.7$_{\pm0.98}$	65.6$_{\pm0.49}$	62.4$_{\pm0.85}$	69.7$_{\pm0.94}$	63.3$_{\pm1.22}$	62.7$_{\pm0.60}$	56.8$_{\pm0.83}$	72.1$_{\pm0.43}$	69.6$_{\pm0.85}$
TFGDA-TD	78.9$_{\pm0.59}$	76.9$_{\pm0.62}$	72.9$_{\pm0.86}$	70.8$_{\pm1.34}$	70.1$_{\pm0.65}$	68.9$_{\pm0.89}$	73.7$_{\pm1.14}$	71.1$_{\pm1.39}$	64.8$_{\pm0.68}$	62.6$_{\pm0.76}$	75.2$_{\pm0.62}$	72.4$_{\pm0.93}$
TFGDA-TR	78.4$_{\pm0.64}$	75.8$_{\pm0.48}$	72.3$_{\pm0.60}$	68.3$_{\pm1.13}$	70.5$_{\pm0.58}$	69.6$_{\pm0.95}$	73.4$_{\pm0.93}$	70.8$_{\pm1.52}$	64.2$_{\pm0.71}$	61.8$_{\pm0.81}$	76.3$_{\pm0.54}$	72.6$_{\pm0.83}$
TFGDA-DR	79.2$_{\pm0.41}$	77.4$_{\pm0.50}$	73.2$_{\pm0.49}$	71.6$_{\pm0.94}$	72.0$_{\pm0.53}$	71.4$_{\pm0.92}$	74.5$_{\pm1.10}$	71.5$_{\pm1.47}$	65.3$_{\pm0.63}$	63.0$_{\pm0.70}$	77.1$_{\pm0.49}$	72.9$_{\pm0.87}$
TFGDA	**81.0**$_{\pm0.34}$	**78.9**$_{\pm0.46}$	**75.3**$_{\pm0.51}$	**73.2**$_{\pm0.89}$	**73.6**$_{\pm0.61}$	**72.3**$_{\pm0.94}$	**76.0**$_{\pm1.02}$	**72.6**$_{\pm1.35}$	**66.9**$_{\pm0.59}$	**64.3**$_{\pm0.72}$	**78.9**$_{\pm0.47}$	**74.4**$_{\pm0.91}$

Results and Discusssion：

We can find that TFGDA greatly outperforms all competitors and becomes a state-of-the-art method.

Experiments



Visualization of Node Features：
Compared with other competitors, TFGDA achieves exactly 5 clusters with clean decision boundaries, indicating that our model can capture more fine-grained transferable features as well as align more complex distributions.