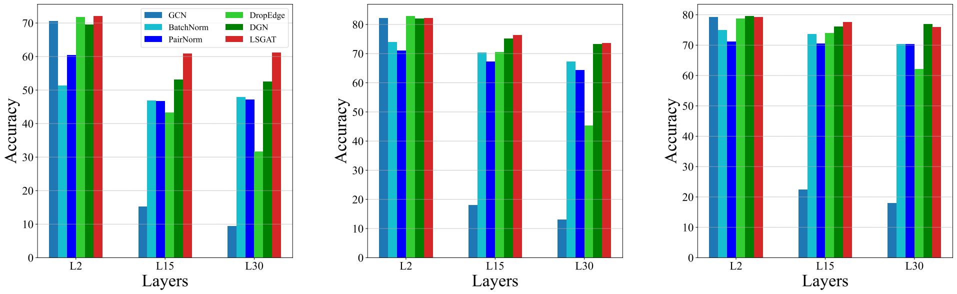
A Further Experiments Details

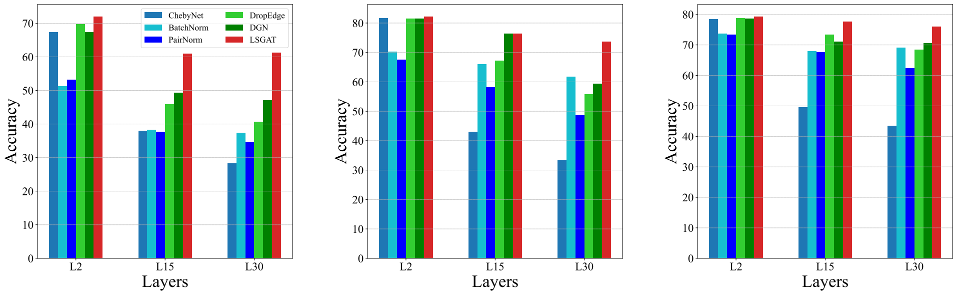
A.1 Comparison based on GCN and ChebyNet

Here, we analysis the comparison between LSGAT and other deep GNNs methods, which tried to relieve the oversmoothing problem based on GCN and ChebyNet two models. As shown in Figure 5 and Figure 6, LSGAT consistently exhibits superior performance. Especially the results with mainly compared layers: fifteen layers and thirty layers, among seventy-two cases based on six algorithms, our LSGAT outperforms in seventy-one cases. The comparison further validates the effectiveness of the LSGAT.



(a) Citeseer (b) Cora (c) Pubmed

Figure 5: The comparison of test accuracies between LSGAT and other general Deep Graph Neural Network methods, which are equipped with Graph Convolutional Neural Network (GCN).



(a) Citeseer (b) Cora (c) Pubmed

Figure 6: The comparison of test accuracies between LSGAT and other general Deep Graph Neural Network methods, which are equipped with ChebyNet.

A.2 Combination of LSGAT and GAT-Lip, and Hyper-parameter Study

Combining with GAT-Lip. To address the gradient explosion issue happened in deep GAT, Dasoulas et al. (2021) proposed GAT-Lips, which could be applied to GAT and our LSGAT. From the results, we can see that both algorithms contribute to the improvement of performance, and the performance is further significantly improved on the original GAT and LSGAT.

Hyper-parameter study. Detailed results have been put in Table 6.

A.3 Comparison with GAT-Based Algorithms

The full results with standard deviations of Section 5.2 has been put in Table 5. Node representation visualization for GAT and LSGAT on Cora dataset in 30 layers could be found in Figure 7 and Figure 8. This can help us understand the specific classification situation.

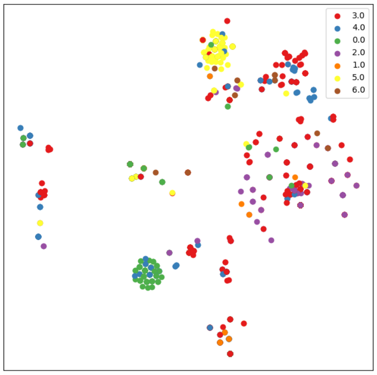


Figure 7: Node representation visualization for GAT on Cora dataset in 30 layers (node colors represent classes).

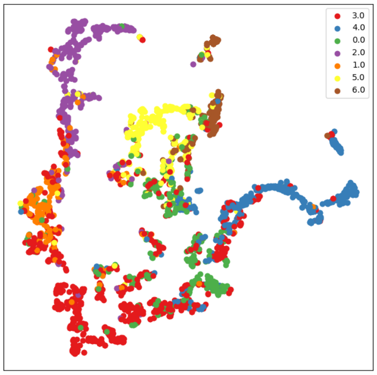


Figure 8: Node representation visualization for LSGAT on Cora dataset in 30 layers (node colors represent classes).

B Experimental Setting

B.1 Datasets Statistics

Datasets. Joining the practice of previous work, we evaluate GNN models by performing the node classification tasks on six real-world datasets: Cora, Citeseer, Pubmed (Sen et al., 2008), CoauthorPhysics, CoauthorCS (Shchur et al., 2018), and Ogbn-Arxiv (Hu et al., 2020). Detailed information of the four benchmark datasets is listed as follows and summarized in Table 4.

- The Cora dataset consists of 2,708 scientific publications classified into one of seven classes, and 5,429 links. Each publication is described by a 0/1-valued word vector indicating the absence/presence of the corresponding word from the dictionary. The dictionary consists of 1,433 unique words.
- The Citeseer dataset consists of 3,312 scientific publications classified into one of six classes, and 4,732 links. Each publication is described by a 0/1-valued word vector indicating the absence/presence of the corresponding word from the dictionary. The dictionary consists of 3,703 unique words.
- The Pubmed Diabetes dataset consists of 19,717 scientific publications from PubMed database pertaining to diabetes classified into one of three classes. The citation network consists of 44,338 links. Each publication is described by a TF-IDF weighted word vector from a dictionary comprised of 500 unique words.

Table 4: Details of the Datasets

Dataset	Nodes	Edges	Features	Classes	Train/Val/Test	Ave. Degree
Cora	2,708	5,429	1,433	7	0.05/0.18/0.37	3.88
Citeseer	3,327	4,732	3,703	6	0.04/0.15/0.30	2.84
Pubmed	19,717	44,338	500	3	0.003/0.03/0.05	4.50
Physics	34,493	247,962	8,415	5	0.003/0.004/0.99	14.38
CoauthorCS	18,333	81,894	6,805	15	0.03/0.10/0.87	8.93
Ogbn-Arxiv	169,343	1,166,243	128	40	0.54/0.18/0.28	13.77

- For CoauthorCS and CoauthorPhysics, they are the coauthorship graph datasets from the scientific fields of computer science and physics, respectively. The nodes represent the authors, and the links indicate whether the two corresponding nodes co-authored papers. Node features represent paper keywords for each author's papers. The node classification task is to predict the most active fields of study for the corresponding author
- The Ogbn-arxiv dataset is a citation network with directed edges, where each node corresponds to an arXiv paper and the edges denote citations from one paper to another. The dataset contains node attributes, that are averaged word embeddings of the titles and the abstracts of dimensionality 128. The label of each node is the subject area of the paper and can take 40 values.

B.2 Parameter Settings

All the codes are implemented in Python 3 and Pytorch 1.11.0, and running on one NVIDIA Quadro RTX 2080Ti GPUs and one NVIDIA Quadro RTX A6000 GPUs serve with CUDA 11.6. All models are trained with a maximum of 1000 epochs using the Adam optimizer (Kingma & Ba, 2014) and early stopping. Weights in GATs models are initialized with Glorot algorithm (Glorot & Bengio, 2010).

Experiments setting of Section 5.2 are strictly followed the previous work Dasoulas et al. (2021), and we directly use the best performance of related parts shown in the work.

GAT-Lip Settings. This experiment is a node classification task, where we evaluate the performance of GNN models with respect to increasing model depth. We used again the Adam optimizer Kingma & Ba (2014) with a weight decay $L = 5 * 10^{-4}$ and the initial learning rate was set in $\{0.1, 0.01, 0.005, 0.001\}$.

- **Model Selection.** We performed for all models and datasets cross-validation with predefined train/validation/test splits and reported the best achieved validation accuracy.
- **Model Depth.** In order to examine the model behavior under the depth increase, for each architecture we used models consisting of l GNN layers, where $l \in \{2, 5, 10, 15, 20, 25, 30\}$. We run each experiment 5 times and we keep the configuration with the best average accuracy.
- **GAT Hyper-parameter tuning.** For each model depth and GNN model, we performed grid-search for hyper-parameter tuning. The hyper-parameters of GAT were tuned are the following: The dimensionality of the hidden units was set in $\{8, 16, 64, 128\}$. The number of attention heads was selected between $\{1, 2, 4, 8\}$ and we experimented over two standard aggregators of the attention heads: a) concatenation and b) averaging of the attention heads. The dropout of the attention weights was set in $\{0, 0.2, 0.5\}$.

Experiments settings of Section 5.3 and Section A.1 are strictly followed the work Jin et al. (2022). Specifically, for BatchNorm, PairNorm and DGN, we reuse the performance reported in Zhou et al. (2020a) for GCN and GAT. For ChebyNet, we use their best configuration to run the experiments. For DropEdge, we tune the sampling percent from $\{0.1, 0.3, 0.5, 0.7\}$, weight decay from $\{0, 5e - 4\}$ dropout rate from $\{0, 0.6\}$ and fix the learning rate to be 0.005. For DeCorr, we reuse the best results shown in Jin et al. (2022). For MAGNA, we use their basic configuration to run the experiments, however, for a fair comparison, we remove the resnet (He et al., 2016a) and LayerNorm used its original work, we also limited the maximum hidden number to 128, which is used in other works under attention mechanism. For our work, We set the number of hidden units as $\{8, 16, 32, 64, 128\}$. The number of attention heads was selected between $\{1, 2, 4, 8\}$. We tune the hyperparameters for all datasets from the following sets: $\{0, 0.1, ..., 0.6\}$ (dropout rate), $5 \times 10^{\{-3,-4\}}$ (learning rate), $5 \times 10^{\{-3,-4,-5\}}$ (L2 regularization).

For experiments in Section 5.4, we directly use the best performance and settings of related works reported in GCNII (Chen et al., 2020).

GCNII Settings. 0.1 (α_l for initial residual), 5×10^{-4} (L2 regularization), and other hyperparameters are tuned by grid search. The experiments are randomly repeated for ten times, and the average accuracy and the standard deviation are reported.

Table 5: Summary of classification accuracy (%) results among deep GAT based methods.

		Layers						
Dataset	Method	2	5	10	15	20	25	30
Cora	GAT	**82.2 ± 1.1**	78.9 ± 1.0	57.8 ± 0.6	35.5 ± 1.1	32.2 ± 1.1	30.0 ± 1.2	**23.9 ± 0.6**
	GAT-Lip	**82.2 ± 0.6**	83.3 ± 2.2	80.7 ± 1.1	78.8 ± 1.2	76.6 ± 0.6	71.6 ± 1.1	**68.8 ± 2.0**
	LSGAT	**82.2 ± 1.0**	79.1 ± 1.3	77.5 ± 0.6	76.4 ± 1.1	76.2 ± 0.8	74.8 ± 1.5	**73.5 ± 0.8**
Citeseer	GAT	**66.8 ± 0.4**	65.0 ± 1.1	62.9 ± 0.7	61.2 ± 2.8	60.9 ± 1.1	59.9 ± 2.1	**56.1 ± 3.1**
	GAT-Lip	**67.1 ± 0.8**	65.9 ± 1.6	62.6 ± 1.8	62.1 ± 1.7	60.1 ± 1.5	60.9 ± 2.5	**59.4 ± 3.8**
	LSGAT	**67.6 ± 0.8**	65.5 ± 0.8	63.8 ± 0.8	60.9 ± 2.5	62.4 ± 1.3	62.5 ± 0.4	**61.2 ± 1.4**
Pubmed	GAT	**76.3 ± 1.9**	78.1 ± 1.1	64.5 ± 1.0	57.4 ± 0.5	51.5 ± 1.4	48.8 ± 1.4	**29.5 ± 1.1**
	GAT-Lip	**77.6 ± 1.2**	80.9 ± 0.7	75.4 ± 0.8	72.4 ± 0.5	73.2 ± 0.8	67.7 ± 1.1	**65.0 ± 1.5**
	LSGAT	**76.5 ± 0.5**	77.0 ± 0.8	76.9 ± 1.2	77.6 ± 0.8	77.4 ± 0.7	76.2 ± 1.5	**72.8 ± 1.4**
Physics	GAT	**93.2 ± 0.6**	91.0 ± 0.1	88.3 ± 0.1	77.0 ± 10	50.0 ± 16	15.3 ± 0.3	**13.6 ± 0.2**
	GAT-Lip	**93.7 ± 0.2**	91.6 ± 1.2	90.4 ± 0.8	84.2 ± 4.3	72.6 ± 8.7	71.7 ± 7.6	**63.9 ± 1.3**
	LSGAT	**93.2 ± 0.4**	92.1 ± 0.6	91.7 ± 0.3	91.5 ± 0.4	91.2 ± 1.0	91.0 ± 0.4	**87.0 ± 2.7**
Ogbn-arxiv	GAT	**72.2 ± 2.4**	72.5 ± 2.0	67.8 ± 2.6	59.5 ± 0.7	53.9 ± 0.8	52.9 ± 0.3	**31.4 ± 2.1**
	GAT-Lip	**72.0 ± 2.0**	72.3 ± 4.4	72.4 ± 2.4	69.7 ± 1.7	67.3 ± 2.1	66.8 ± 2.5	**62.2 ± 1.8**
	LSGAT	**72.2 ± 2.2**	72.7 ± 3.5	71.8 ± 2.7	70.5 ± 0.4	67.9 ± 2.1	67.3 ± 3.1	**64.4 ± 1.2**

Table 6: Test accuracies (%) of LSGAT based on different β w/wo GAT-Lip

					Layers			
Dataset	Method	2	5	10	15	20	25	30
Cora	LSGAT 0.2	80.5 ± 1.1	79.1 ± 1.3	77.5 ± 0.6	76.1 ± 1.6	75.8 ± 1.8	74.8 ± 1.5	73.5 ± 0.8
	LSGAT + Lip 0.2	80.5 ± 0.8	79.1 ± 1.3	77.1 ± 0.6	76.0 ± 1.3	75.8 ± 1.5	74.8 ± 1.5	69.3 ± 6.2
	LSGAT 0.4	80.5 ± 0.4	78.9 ± 1.0	77.2 ± 0.5	76.3 ± 0.5	75.5 ± 0.4	74.6 ± 2.2	69.2 ± 1.6
	LSGAT + Lip 0.4	80.1 ± 0.6	78.9 ± 0.5	77.6 ± 1.0	76.4 ± 1.4	75.3 ± 1.8	74.2 ± 0.8	72.4 ± 2.7
	LSGAT 0.6	80.6 ± 1.3	78.7 ± 1.0	77.2 ± 1.6	76.4 ± 1.1	76.2 ± 0.8	74.5 ± 2.3	71.2 ± 3.0
	LSGAT + Lip 0.6	80.1 ± 0.3	78.7 ± 1.2	77.3 ± 1.0	76.8 ± 0.7	75.3 ± 1.0	74.0 ± 1.5	70.2 ± 3.1
	LSGAT 0.8	80.4 ± 1.0	78.6 ± 0.7	77.2 ± 0.7	76.2 ± 0.7	75.9 ± 0.6	74.3 ± 2.5	71.4 ± 2.1
	LSGAT + Lip 0.8	80.2 ± 0.4	78.8 ± 0.8	77.1 ± 0.8	77.0 ± 1.9	75.2 ± 1.1	75.9 ± 1.3	73.0 ± 3.4
Pubmed	LSGAT 0.2	76.1 ± 1.5	76.1 ± 0.7	76.3 ± 0.5	77.3 ± 1.1	76.9 ± 0.9	75.9 ± 1.9	72.8 ± 1.4
	LSGAT + Lip 0.2	76.2 ± 0.8	76.0 ± 1.1	76.4 ± 1.6	77.0 ± 0.6	76.9 ± 1.0	76.2 ± 1.6	71.9 ± 4.4
	LSGAT 0.4	76.3 ± 0.5	76.9 ± 1.1	75.9 ± 1.5	76.8 ± 1.4	77.4 ± 0.7	75.7 ± 2.0	73.2 ± 8.0
	LSGAT + Lip 0.4	76.5 ± 0.8	76.7 ± 1.6	76.8 ± 1.2	76.8 ± 1.5	78.0 ± 0.8	76.4 ± 1.0	75.7 ± 3.0
	LSGAT 0.6	76.2 ± 0.7	76.7 ± 0.8	76.6 ± 1.1	77.3 ± 0.3	77.0 ± 1.3	75.7 ± 0.8	72.1 ± 2.4
	LSGAT + Lip 0.6	76.3 ± 1.8	76.5 ± 0.8	76.8 ± 2.9	76.9 ± 1.0	77.5 ± 1.1	76.1 ± 0.8	72.0 ± 2.6
	LSGAT 0.8	76.5 ± 0.5	77.0 ± 0.8	76.9 ± 1.2	77.6 ± 0.8	76.7 ± 1.0	76.2 ± 1.5	71.3 ± 3.4
	LSGAT + Lip 0.8	76.4 ± 1.2	76.7 ± 0.5	77.0 ± 0.6	77.0 ± 1.0	77.7 ± 0.5	76.7 ± 0.9	71.7 ± 2.2