Appearance and Structure Aware Robust Deep Visual Graph Matching: Attack, Defense and Beyond Supplementary Material

Anonymous CVPR submission

Paper ID 7620

In this Supplementary Material, we first present additional results to verify the robustness of our model (Appendix A). Then, we verify the applicability of our locality attack and AAR on another popular deep GM baseline (Appendix B), followed by our further analysis of the advantage of our locality attack as a data augmentation technique (Appendix C). We also delve into analyzing the vulnerability of deep GM when attacks target on different parts of feature space (Appendix D). Finally, we give time complexity analysis (Appendix E), more visual results about our adversarial attack and defense (Appendix F) and the implementation details (Appendix G).

Table 1. Robust accuracy (%) of (non-)robust models. W-B means white-box attack with B-B as black-box attack. For B-B, MI-FGSM is transfer-based with NGMv2 as the surrogate model.

Attackers \ Defenders	PCA	CIE	NGMv2		
	Baseline	Baseline	Baseline	Pixel AT	ASAR-GM
C&W-20 (W-B)	3.72	3.27	2.31	70.72	**70.79**
PGD-20 (W-B)	9.8	9.77	24.37	70.5	**70.58**
PGD-50 (W-B)	8.76	9.18	23.94	70.5	**70.58**
combo-50 (W-B)	7.99	8.16	21.46	54.09	**69.6**
MI-FGSM (B-B)	35.89	27.25	-	71.33	**73.23**

A. More results about White-box and Black-box Attacks

To verify that our robust model does not suffer from incomplete evaluations, we follow the series of sanity checks introduced in [2] and present more results about white-box and black-box attacks, which are in line with the conclusions in our paper. For white-box attacks, we perform PGD attack with multiple steps, e.g., PGD-20, PGD-50 and also the targeted C&W attack [3]. Note that except the number of attack steps, the white-box setting is the same as reported in our paper. For black-box attacks, we choose another powerful attack: the transfer-based MI-FGSM attack [4] to evaluate robustness. Table 1 shows that under stronger PGD attack, e.g., PGD-50 and combo-50 that represents PGD-50 combo attack, our model consistently exhibits the best robustness, whose accuracy converges with the increasing of attack steps. Moreover, results under MI-FGSM attack also verify that our robust model achieves true robustness without causing obfuscated gradients [1], since MI-FGSM does not exploit gradient information of our model.

B. Applicability of Locality Attack and AAR

To verify the applicability, we select another popular deep GM baseline, PCA-GM [5]). Table 2 shows our AAR achieves both better clean accuracy and robustness over Pixel AT. Moreover, our locality attack boosts the accuracy of PCA-GM on clean examples, being also consistent with our conclusion.

Table 2. Clean and Robust accuracy(%) of (non-)robust PCA-GM

Defenders \ Attackers	Clean	pixel	locality	combo
		PGD-20	PGD-20	PGD-20
PCA-GM(baseline)	64.78	14.04	31.47	12.56
Pixel AT$_{FGSM}$	54.45	54.15	36.73	36.48
Locality AT(ours)	**66.17**	46.18	**51.1**	38.13
AAR(ours)	60.85	**56.77**	46.10	**44.20**

C. Generalization Analysis of ASAR-GM

In our paper, we claim that our defense mechanism can act as a data augmentation technique to achieve better clean accuracy on the test dataset, since our perturbations on keypoint locality induce various graph structures. To verify the advantage of our locality attack, we train baseline model with random locality noise. As introduced in Table 5 in our paper, we devise three types of locality attack: "location" (only perturb keypoint locations with the original graph structure), "structure" (only reconstruct graph structure with location of keypoints unchanged), and "both" (perturb keypoint location and graph structure together). We generate random noise for the three kinds of locality attack respectively and train the baseline model. The clean accuracy of the three models is 70.91% (random noise

Figure 1. Visualization of the matching results of the baseline NGMv2 and our robust model under our adversarial attack. Our model exhibits superior robustness on Pascal VOC dataset. One image pair is randomly sampled and visualized for each of the 10 classes.

Table 3. Test accuracy (%) of (non-)robust NGMv2 for vulnerability analysis. Robustness is tested in different cases where only one/both of node and edge similarity matrices is/are perturbed.

Defenders \ Attackers	Attack Scale	pixel attack	locality attack	combo attack
		FGSM	FGSM	FGSM
NGMv2	clean	80.4	-	-
	only nodes	56.1	70.9	55.43
	only edges	40.39	69.49	37.4
	both	36.97	64.47	33.51
ASAR-GM$_{\text{(config 1)}}$	clean	81.14	-	-
	only nodes	78.73	77.12	75.48
	only edges	73.84	75.52	69.61
	both	73.50	73.43	67.42

on "location"), 79.02% (random noise on "structure"), and 80.43% (random noise on "both"), which reveals that random noise on locality during training even harms model generalization ability while our adversarial perturbation on locality helps our model generalize better on the clean test dataset.

D. Vulnerability Analysis of Deep GM

As introduced in Sec. 3.1 in the main paper, deep graph matching establishes node-to-node correspondences between two graphs through learning node-to-node and edge-to-edge affinity, i.e., learning node similarity matrix and edge similarity matrix. As shown in Table 3, we further analyze the vulnerability of deep graph matching at node and edge level by attacking the (node) edge similarity matrix while the (edge) node similarity matrix remains unchanged, i.e., one similarity matrix gets inferenced via adversarial inputs while the other one is obtained via clean inputs. For NGMv2 and its robust model, we find that the learned edge affinity is more vulnerable than node affinity when being attacked. It is probably because that attacks on the edge affinity have a greater impact on graph neural network (GNN), which has been utilized to learn the affinity during the whole pipeline of deep graph matching.

E. Time Complexity Analysis

Our defense (config 1 and config 2) only need **one extra** gradient backward propagation per mini-batch to finish two parts: i) Calculating our appearance aware regularization loss. ii) Crafting our adversarial examples. Compared to the common PGD-n AT setting with loss backward **n more times**, our defense achieves better clean accuracy and robustness with a lower time computation cost.

F. More Visual Results on Attack and Defense

The Pascal VOC dataset contains 20 classes in total and in our paper, we select 10 of them to visualize. We also visualize the other 10 classes in Fig. 1, which demonstrates the effectiveness of our attack and defense mechanism on the whole dataset.

G. Implementation Details

Attack. We generate our adversarial examples within the bounded ℓ_∞-norm ball. For pixel attack, we set the perturbation budget $\epsilon_{\mathbf{c}}$ as $8/255$ while for locality attack, $\epsilon_{\mathbf{z}}$ is set as 8 (the image size is 256×256). For each iteration during PGD attack in Eq. 3 in our paper, we keep the step size α as $2/255$ for pixel attack and 2 for locality attack.

Defense. For our appearance aware regularizer (AAR) in Eq. 10a, we set β as 1.5 across all training examples. In the maximization step of Eq. 10b, for config 1 & 2, we apply single step FGSM attack as our attack to generate our adversarial examples while for config 3 we perform stronger PGD-2 attack for better robustness. All of the other hyperparameters of our robust model are in line with the official setting of NGMv2.

References

[1] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing

defenses to adversarial examples. In *Int. Conf. Mach. Learn.*, pages 274–283, 2018. 1

[2] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019. 1

[3] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *ieee symposium on security and privacy (sp)*, pages 39–57. IEEE, 2017. 1

[4] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Comput. Vis. Pattern Recog.*, pages 9185–9193, 2018. 1

[5] Runzhong Wang, Junchi Yan, and Xiaokang Yang. Learning combinatorial embedding networks for deep graph matching. In *Int. Conf. Comput. Vis.*, pages 3056–3065, 2019. 1