Certified Robustness on Visual Graph Matching via Searching Optimal Smoothing Range

Anonymous Author(s)*

ABSTRACT

Deep visual graph matching (GM) is a challenging combinatorial task that involves finding a permutation matrix that indicates the correspondence between keypoints from a pair of images. Like many learning systems, empirical studies have shown that visual GM is susceptible to adversarial attacks, with reliability issues in downstream applications. To the best of our knowledge, certifying robustness for deep visual GM remains an open challenge with two main difficulties: how to handle the paired inputs together with the heavily non-linear permutation output space (especially at large scale), and how to balance the trade-off between certified robustness and matching performance.

Inspired by the randomized smoothing (RS) technique, we propose the Certified Robustness based on the Optimal Smoothing Range Search (CR-OSRS) technique to fulfill the robustness guarantee for deep visual GM. First, unlike conventional RS methods that use isotropic Gaussian distributions for smoothing, we build the smoothed model with paired joint Gaussian distributions, which capture the structural information among keypoints, and mitigate the performance degradation caused by smoothing. For the vast space of the permutation output, we devise a similarity-based partitioning method that can lower the computational complexity and certification difficulty. We then derive a stringent robustness guarantee that links the certified space of inputs to their corresponding fixed outputs. Second, we design a global optimization method to search for optimal joint Gaussian distributions and facilitate a larger certified space and better performance. Third, we apply data augmentation and a similarity-based regularizer in training to enhance smoothed model performance. Lastly, for the high-dimensional and multivariable nature of the certified space, we propose two methods (sampling and marginal radii) to evaluate it. Experimental results on public benchmarks show that our method achieves state-of-the-art ℓ_2 certified robustness. Source codes will be made publicly available.

CCS CONCEPTS

• **Security and privacy** → **Logic and verification**.

KEYWORDS

Certified Robustness, Visual Graph Matching, Optimal Smoothing

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

ACM Reference Format:
Anonymous Author(s). 2018. Certified Robustness on Visual Graph Matching via Searching Optimal Smoothing Range. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 21 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

As an essential and popular combinatorial problem, graph matching (GM) has attracted extensive attention over the decades with wide applications in vision [35], text [40], graphics [34], pattern recognition [33], machine learning [47] etc. It refers to establishing correspondences among two [7] or multiple graphs [17], and due to the implicit nature of graphs in many real-world scenarios, visual graph matching is receiving increasing attention by jointly modeling visual perception and the matching procedure by which CNN and GNN are often both used for node/edge feature extraction and matching, respectively [37, 43].

On the other hand, studies on the robustness of machine learning models have attracted intense attention, while the robustness of combinatorial solvers (especially with machine learning) remains a crucial, yet largely unexplored area [13, 26]. Under the deep visual GM paradigm, recent work [29] demonstrates that visual GM models are susceptible to perturbations applied to keypoints and image pixels, which suggests the frailness of both the visual perception CNN model as well as the backend decision model for combinatorial matching, and the authors propose an empirical defense method based on an appearance-aware regularizer. However, there is still a lack of a principled certified defense to provide theoretical robustness guarantees for GM (let alone other combinatorial problems, especially in the practical context of prediction-and-optimization [14]).

Certified robustness and empirical robustness are two distinct concepts in the context of adversarial machine learning. Certified robustness provides a rigorous verification of the model's output invariance under a bounded perturbation set, regardless of the attacks employed. However, empirical robustness lacks such a theoretical guarantee and only evaluates the model's defense capabilities against existing attack methods, which may not generalize to future unseen attacks. Existing certified robustness mechanisms (including randomized smoothing, which we focus on in this study) in the graph domain [3, 16, 31, 49] are confined to the unconstrained node-level or graph-level classification/prediction task with a single graph input and are not readily applicable to address visual GM problems with cross-graph and structured output (permutation matrix).

Certified robustness strives to design solvers whose prediction for any input x is verifiable invariant within some set around the

input [39]. Randomized smoothing (RS) [8, 20] is a promising approach to achieve certified robustness of large-scale neural networks against arbitrary attacks. Given an input x and a base function, RS constructs a smoothed function that is certifiably robust within the region induced by x and the smoothing distribution $\mathcal{D}$ (usually an isotropic Gaussian distribution). RS has been widely applied to certify various models, e.g., image classification [42] and object detection in vision [5], which motivates us to develop RS-based certified robustness for visual GM.

Applying RS to visual GM poses several challenges. **C1: paired inputs.** The input of visual GM consists of paired images and keypoint position matrices, which means that perturbations are also in pairs and mutually constrained in the certified space. This differs from the single input setting of previous certification problems. **C2: dependency of keypoints.** The graph structure derived from Delaunay triangulation of keypoint positions as a whole conveys important structural information. It is an essential intermediate result for the visual GM model, which motivates us to preserve the original graph structure during the smoothing process to maintain the matching performance. **C3: large permutation output space.** The output of visual GM is a $0-1$ permutation matrix, which has an exponential number of theoretical possibilities. For a matching task with n keypoints, the output is an $n \times n$ matrix, and there are $n!$ theoretically possible outputs. This means that we cannot directly apply the existing RS definition, which estimates the occurrence probability for each possible output and would cause a computational explosion. Furthermore, this method is prone to cause certification failure or an exceedingly small certified space without a dominant output (an output with a markedly high probability of occurrence). **C4: performance degradation caused by smoothing.** Smoothing can influence model performance, as corroborated by prior studies. Although data augmentation is a customary method to enhance performance, it is not tailored for visual GM and thus its effect is inadequate if applied directly.

To address these challenges, we propose **C**ertified **R**obustness based on **O**ptimal **S**moothing **R**ange **S**earch (**CR-OSRS**), a novel robustness certification method for visual GM. Specifically, we assume that the two perturbations against paired inputs belong to the joint input space and derive a certification result that adheres to the inter-pair constraints (**C1**). We design a smoothed model by joint Gaussian distributions that captures the correlation of keypoints and employ global optimization to determine the optimal correlation parameters that enhance certified robustness. The rationale of this design is to preserve the difference and avoid confusion among keypoints under perturbations as much as possible (**C2**). Furthermore, we delineate a subspace of the output space by a similarity threshold and characterize the certified robustness as the output that is invariably within the subspace under perturbations. This eliminates the need to count the probability of each possible output and only requires calculating the probability that the output falls into the subspace (**C3**). Additionally, we propose a data augmentation method for visual GM using joint Gaussian noise and an output similarity-based regularizer, which improves both the matching accuracy and certified robustness (**C4**).

The contributions of this paper are as follows:

1) We propose a certification method for visual GM, CR-OSRS, providing the rigorous robustness guarantee by characterizing an ℓ_2 norm certified input space (see Theorem 4.1). This means when the perturbation is within the certified input space, the smoothed model always predicts the output within the output subspace.

2) Specifically, we devise a smoothed model by joint Gaussian distributions and globally optimize the correlation parameters of the distributions, which can capture the connection of keypoints to enhance the anti-disturbance ability of the model (see Sec. 4.2). We also apply data augmentation with joint Gaussian noise and the output similarity-based regularizer during the training phase to further improve the model performance (see Sec. 4.3).

3) We devise two methods, sampling and marginal radii respectively, to measure the certified space for quantitative analysis (see Sec. 4.4). We evaluate our approach on the Pascal VOC dataset [11] with Berkeley annotations [4], the Willow ObjectClass dataset [6] and SPair-71k dataset [27] for six representative GM solvers. The results show that CR-OSRS provides robustness guarantees for visual GM, and the CR-OSRS mechanism performs better than directly applying RS [8] to visual GM, which we denote as RS-GM. Moreover, our designed training methods are also effective (see Sec. 5).

2 RELATED WORKS

We review studies on certified robustness through RS as well as discuss the GM solver and its robustness.

2.1 Certified Robustness by Randomized Smoothing

Randomized smoothing is proposed in Lecuyer et al. [20] as a certified adversarial defense and used to train the pioneering certifiably robust classifier on ImageNet. However, it provides loose guarantees. Cohen et al. [8] shows that Gaussian noise addition provides a tight ℓ_2 certification radius, with subsequent works on new RS-type techniques, e.g. techniques using smoothing distributions at different norms [22, 23, 42], and techniques for different tasks [5, 16, 19, 32]. However, all previous smoothing distributions $\mathcal{D}$ deprive of favorable prior knowledge, which mainly refers to the keypoint positions and graph structure in visual GM. Moreover, all of them only certify a single image or graph but do not consider the combinatorial nature as in visual GM.

2.2 Graph Matching and its Robustness

Approximate GM solvers have evolved from traditional methods without learning [10] to learning-based [41]. In practice, GM methods are often closely related to visual data for matching, whereby the model needs to consider both visual point features and constrained combinatorial matching. Seminal work [45] proposes a deep neural network pipeline for visual GM, in which image features are learned through CNN, with subsequent variants [30, 35], among which a major improvement is to exploit structural information using different techniques, for example GNN, rather than only using appearance for node/edge attributes as done in [45]. Our work, which uses the RS-type technique, treats the GM solver as a black box that guarantees the generality of our method, irrespective of its learning-based or non-learning-based nature.

There are also works on adversarial attacks and defense on (deep) GM. Previous work [44] proposes a robust GM model against perturbations, e.g., distortion, rotation, outliers, and noise. Zhang et al.

[48] propose an adversarial attack model for deep GM networks that employs kernel density estimation to create dense regions where neighboring nodes are indiscernible. The work [28] devises two specific topology attacks in GM: inter-graph dispersion and intra-graph combination attacks, and proposes a resilient defense model. The recent effort Lin et al. [24] integrates the momentum distillation strategy to balance the quadratic contrastive loss and reduce the impact of bi-level noisy correspondence. However, these defense methods are all heuristic and lack robustness certification under unseen attacks.

3 PRELIMINARIES

3.1 Randomized Smoothing

The RS in [8] transforms an arbitrary base classifier f into a smoothed classifier g that is certifiably robust under ℓ_2 norm. For any single input x, the smoothed classifier g returns the most probable prediction of f for the random variable $\mathcal{N}(x;\sigma^2 I)$:

$$g(x) = \arg\max_{c\in\mathcal{Y}} P(f(x+\varepsilon)=c), \tag{1}$$

where $\varepsilon \sim \mathcal{N}\left(0,\sigma^2 I\right)$ is an isotropic Gaussian noise. Then the certified radius within which the output is unchanged for $g(x+\delta) = c_A$ that measures the certified robustness is:

$$\|\delta\|_2 < \frac{\sigma}{2}\left(\Phi^{-1}\left(\underline{p_A}\right) - \Phi^{-1}\left(\overline{p_B}\right)\right), \tag{2}$$

where the most probable class c_A is returned with probability p_A and the "runner-up" class is returned with probability p_B. $\underline{p_A}$ and $\overline{p_B}$ are the lower bound and upper bound of p_A and p_B, respectively, and Φ^{-1} is the inverse of the standard Gaussian cumulative distribution function.

3.2 Visual Graph Matching

We consider the visual GM task f which is a comprehensive setting allowing for both visual appearance and structural perturbation: $(\mathbf{c}^1,\mathbf{c}^2,\mathbf{z}^1,\mathbf{z}^2)\rightarrow\mathbf{X}$, where $(\mathbf{c}^1,\mathbf{c}^2)$ is the image pair with keypoint position pair ($\mathbf{z}^1\in\mathbb{R}^{n_1\times 2},\mathbf{z}^2\in\mathbb{R}^{n_2\times 2}$), n_1 and n_2 are the numbers of keypoints, $\mathbf{X}\in\{0,1\}^{n_1\times n_2}$ represents a 0-1 permutation matrix. Recent deep GM methods tackle images with keypoints as inputs in an end-to-end manner [30, 35, 36, 45] and typically comprise three components: keypoint feature extractor, affinity learning, and final correspondence solver. First, two graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ are constructed by Delaunay triangulation [21]. Then node features are obtained via a feature extractor based on the keypoint positions. Afterward, edge features are constituted based on node features and topology information of $\mathcal{G}_1$ and $\mathcal{G}_2$. Features are obtained by bilinear interpolation on the feature map. Based on these node and edge features, the affinity matrix $\mathbf{K}\in\mathbb{R}^{n_1n_2\times n_1n_2}$ is initialized which is then fed to the affinity learning layer to learn the node-to-node and edge-to-edge correspondence similarity. Finally, the correspondence solver outputs the predicted permutation matrix $\mathbf{X}$ by solving quadratic assignment problem (QAP) [25] which aims to maximize the overall affinity score J:

$$\begin{aligned}&\max_{\mathbf{X}} J(\mathbf{X}) = \mathrm{vec}(\mathbf{X})^\top\mathbf{K}\,\mathrm{vec}(\mathbf{X}),\\ &\text{s.t. } \mathbf{X}\in\{0,1\}^{n_1\times n_2}, \mathbf{X}\mathbf{1}_{n_1}=\mathbf{1}_{n_1}, \mathbf{X}^\top\mathbf{1}_{n_2}\le\mathbf{1}_{n_2},\end{aligned} \tag{3}$$

where $\mathrm{vec}(\mathbf{X})$ denotes the column-wise matrix of $\mathbf{X}$ which is a partial permutation matrix if $n_1 < n_2$.

As discussed above, image pixels affect the extracted node and edge features, while keypoint positions affect the extracted node features by influencing the bilinear interpolation and the graph structure extracted by Delaunay triangulation. However, the keypoint positions are inherently vulnerable due to the randomness of human labeling or keypoint detectors (which are used in the pre-processing step to locate key objects in an image [4]), and the image pixels are extremely sensitive to various noises imperceptible to humans as in other image tasks. Therefore, in this study, we consider the robustness of visual GM under two types of perturbations: perturbations on image pixels and keypoint positions respectively as in Ren et al. [29]. As these two perturbations belong to different spaces and exhibit different effects on GM models, we devise different certification schemes for them. We examine the certified robustness of GM models under perturbations on image pixels with fixed keypoint positions, and under perturbations on keypoint positions with fixed pixel values.

4 METHODOLOGY

This section introduces the methodology of this work that comprises four parts: (1) the definition of a smoothed GM model and the theoretical framework developed for certified robustness analysis (Sec. 4.1); (2) the construction of the joint Gaussian distribution and an optimization method that assists in searching the optimal correlation parameter to optimize the trade-off between certified robustness and model performance (Sec. 4.2); (3) a training method that incorporates data augmentation with joint Gaussian noise and an output similarity-based regularizer that limits the discrepancies among smoothed outputs (Sec. 4.3); (4) two methods (sampling and marginal radii) for quantifying the robustness certification (Sec. 4.4). The pipeline is shown in Fig. 1 with the process detailed in Alg. 1.

4.1 Robustness Guarantee for Visual GM

As discussed in Sec. 3, we certify the robustness under two types of perturbations: keypoint position perturbations and image pixel perturbations respectively. In this subsection, we focus on the certified robustness under keypoint position perturbations, and the certified robustness under image perturbations can be derived similarly.

As stated in Sec. 1, visual GM poses a challenge for certified robustness due to its large permutation output space. Previous research e.g. [8] aims to certify that the output remains unchanged under fixed range perturbations, but this may result in a failed certification or an extremely restricted certified space for visual GM due to the lack of a dominant output – the probability difference between the most probable matrix and the "runner-up" matrix is small. Furthermore, it is computationally intractable to enumerate the probabilities of all possible output matrices. We propose a novel certified robustness definition that guarantees the output always belongs to an output subspace centered on the core output.

We first define a core output $\mathbf{X}_c$. When queried at $(\mathbf{c}^1,\mathbf{c}^2,\mathbf{z}^1,\mathbf{z}^2)$, $\mathbf{X}_c$ is a more likely output of base GM function f when $(\mathbf{z}^1,\mathbf{z}^2)$ is

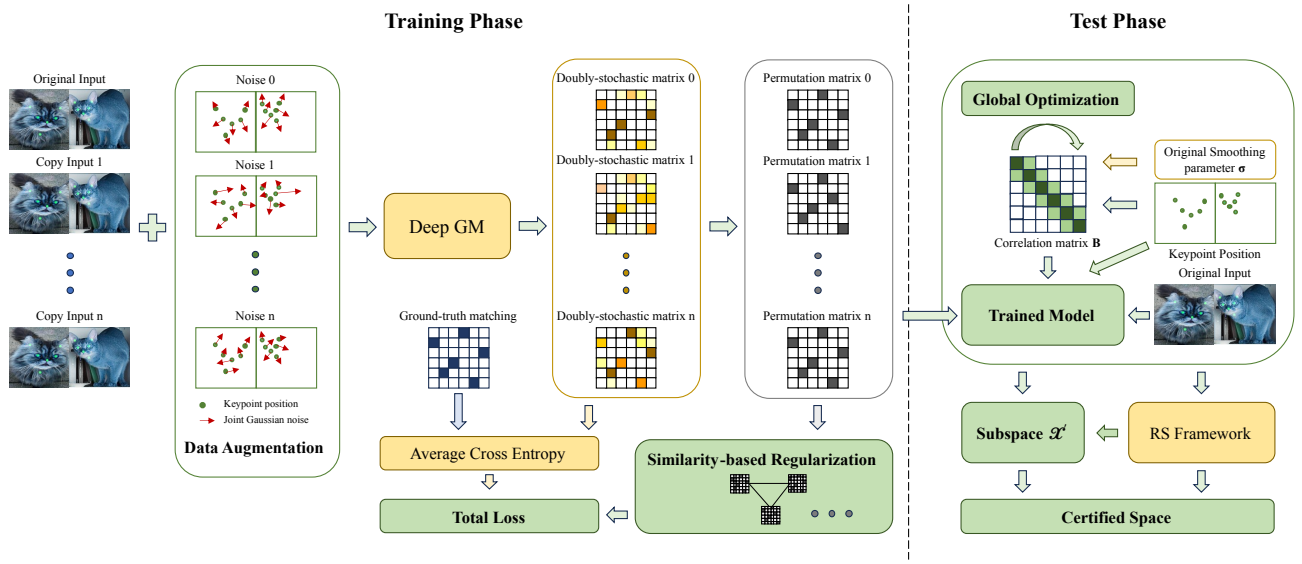


Figure 1: The pipeline of this work consists of two phases: training and testing. In the training phase, we enhance the certified robustness and matching accuracy of the model simultaneously by applying data augmentation and a regularizer as defined in Eq. 11. In the testing stage, we first construct joint Gaussian distributions and employ the global optimization in Eq. 10 to search for the optimal smoothing range. Moreover, we use the optimization results and the trained model to construct a smoothed model, and then compute the output subspace and input certified space following the procedure in Sec. 4.1. It depicts a sample pipeline of robustness certification under keypoint position perturbations.

perturbed by joint Gaussian noise:

$$\mathbf{X}_c = H(S(\sum f\left(\mathbf{c}^1, \mathbf{c}^2, \mathbf{z}^1 + \varepsilon_1, \mathbf{z}^2 + \varepsilon_2\right))), \quad \text{where } \varepsilon_1 \sim \mathcal{N}(0, \Sigma_1), \varepsilon_2 \sim \mathcal{N}(0, \Sigma_2), \tag{4}$$

where the smoothing noise ε_1 and ε_2 follow joint Gaussian distributions with covariance matrices Σ_1 and Σ_2, which represent constraints between keypoints $\mathbf{z}^1$ and $\mathbf{z}^2$ respectively (for solving **C1**). S is the Sinkhorn operator that converts the vertex score matrix into a doubly-stochastic matrix and H is the Hungarian operator that transforms a doubly-stochastic matrix into a permutation matrix. The computation of Eq. 4 takes into account the "majority decision" of RS while only needing to store the sum of matching matrices rather than the statistics of each possible matrix. Note that $\mathbf{X}_c$ is not the output we aim to certify; it is merely the center point of the subspace to be constructed, and thus there is no necessity to consider whether this computation process is provably robust.

Next, we define a subspace $\mathcal{X}'$ of the entire output space $\mathcal{X}$ by a similarity threshold $s \in [0, 1]$. The similarity between the points in $\mathcal{X}'$ and the core output $\mathbf{X}_c$ is no less than s (for solving **C3**).

$$\mathcal{X}' = \left\{ \mathbf{X}_i \middle| \frac{\mathbf{X}_i \cdot \mathbf{X}_c}{\mathbf{X}_c \cdot \mathbf{X}_c} \geq s, \mathbf{X}_i \in \mathcal{X} \right\}, \tag{5}$$

where we employ a dot product $\mathbf{X}_i \cdot \mathbf{X}_c$ to measure the number of identical matching keypoints in $\mathbf{X}_i$ and $\mathbf{X}_c$, because all outputs are 0-1 permutation matrices. Similarly, $\mathbf{X}_c \cdot \mathbf{X}_c$ calculates the total number of keypoints to be matched.

By the above definition, we construct a new base function f_0 based on f. Specifically, we partition the entire output space into two parts by Eq. 5, then assign all points inside $\mathcal{X}'$ with 1 and the rests with 0, and finally convert f to a binary function f_0 as:

$$f_0\left(\mathbf{c}^1, \mathbf{c}^2, \mathbf{z}^1, \mathbf{z}^2\right) = \begin{cases} 1, & \text{if } f(\mathbf{c}^1, \mathbf{c}^2, \mathbf{z}^1, \mathbf{z}^2) \in \mathcal{X}' \\ 0, & \text{otherwise} \end{cases}. \tag{6}$$

Then we build a smoothed function g_0 from f_0. When queried at the input $(\mathbf{c}^1, \mathbf{c}^2, \mathbf{z}^1, \mathbf{z}^2)$ with fixed $(\mathbf{c}^1, \mathbf{c}^2)$, g_0 outputs the binary labels when $(\mathbf{z}^1, \mathbf{z}^2)$ is perturbed by joint Gaussian noise:

$$g_0\left(\mathbf{c}^1, \mathbf{c}^2, \mathbf{z}^1, \mathbf{z}^2\right) = \begin{cases} 1, & \text{if } P(f(\mathbf{c}^1, \mathbf{c}^2, \mathbf{z}^1 + \varepsilon_1, \mathbf{z}^2 + \varepsilon_2) \in \mathcal{X}') \geq 1/2 \\ 0, & \text{otherwise} \end{cases}, \quad \text{where } \varepsilon_1 \sim \mathcal{N}(0, \Sigma_1), \varepsilon_2 \sim \mathcal{N}(0, \Sigma_2). \tag{7}$$

Theorem 4.1 (ℓ_2 **norm certified space for visual GM**). *Let f be a matching function, f_0 and g_0 be defined as in Eq. 6 and Eq. 7, $\varepsilon_1 \sim \mathcal{N}(0, \Sigma_1)$, $\varepsilon_2 \sim \mathcal{N}(0, \Sigma_2)$. Suppose $\underline{p} \in (\frac{1}{2}, 1]$ satisfy:*

$$P(f_0\left(\mathbf{c}^1, \mathbf{c}^2, \mathbf{z}^1 + \varepsilon_1, \mathbf{z}^2 + \varepsilon_2\right) = 1) = P(f(\mathbf{c}^1, \mathbf{c}^2, \mathbf{z}^1 + \varepsilon_1, \mathbf{z}^2 + \varepsilon_2) \in \mathcal{X}') = p \geq \underline{p}. \tag{8}$$

Then we obtain the ℓ_2 norm certified space for the perturbation pair (δ_1, δ_2):

$$\frac{\delta_1^\top \Sigma_1^{-1} \delta_1 + \delta_2^\top \Sigma_2^{-1} \delta_2}{\|\delta_1^\top \Sigma_1^{-1} \mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1} \mathbf{B}_2\|} < \Phi^{-1}\left(\underline{p}\right), \tag{9}$$

which guarantees $g_0\left(\mathbf{c}^1, \mathbf{c}^2, \mathbf{z}^1 + \delta_1, \mathbf{z}^2 + \delta_2\right) = 1$. $\mathbf{B}_1 \in \mathbb{R}^{n_1 \times n_1}$ and $\mathbf{B}_2 \in \mathbb{R}^{n_2 \times n_2}$ are full rank and real symmetric matrics based on the keypoint correlation of keypoint position matrices $\mathbf{z}^1$ and $\mathbf{z}^2$, satisfying $\mathbf{B}_1{}^\top \mathbf{B}_1 = \Sigma_1$ and $\mathbf{B}_2{}^\top \mathbf{B}_2 = \Sigma_2$.

Finally, we formulate a robustness guarantee of g_0 that ensures the similarity between the matching matrix and $\mathbf{X}_c$ being no less than s, that is, the matching matrix always belongs to the subspace $\mathcal{X}'$. We present and illustrate the detailed settings as well as the properties of $\mathbf{B}_1$ and $\mathbf{B}_2$ in Sec. 4.2. The complete proof of Theorem 4.1 is provided in Appendix A.

4.2 Joint Smoothing Distribution

This subsection presents the detailed settings and properties of $\mathbf{B}_1$ and $\mathbf{B}_2$ under keypoint position perturbations. Additionally, we introduce an optimization method to search the optimal smoothing range for enhancing robustness. Besides, refer to Appendix D.2 for the case under pixel perturbations.

As stated in Sec. 3, keypoint positions influence the extracted features through bilinear interpolation and directly determine the graph structure derived by Delaunay triangulation. If the smoothing noise for each keypoint position is completely independent, then the perturbed keypoint set may exhibit partial overlaps or high similarities. This may cause the extracted features to overlap and thus degrade the matching performance. Therefore, our objective is to design a smoothing distribution that can preserve the diversity of keypoints (for solving **C2**).

To construct the correlation matrices $\mathbf{B}_1$ and $\mathbf{B}_2$, we use a correlation parameter b. The diagonals of $\mathbf{B}_1$ and $\mathbf{B}_2$ are original σ as in RS [8], the off-diagonal elements adjacent to the main diagonal are $\sigma \times b$, and the remaining elements are 0. This setting not only maintains the correlation between keypoints but also allows b and σ to be global parameters that can be optimized. We devise an optimization method that aims to maximize the volume of the certified space through the proxy radius, which will be defined in Sec. 4.4. We impose a constraint on b in the optimization method to keep it within a reasonable range, as a large b may enhance the matching performance but diminish the certified space. The optimization problem can be written as:

$$\arg\max_{\sigma,b} \Phi^{-1}\left(\underline{p}\right) \sum_{i=1,2} \left(\sqrt[2m_i]{\prod_j^{m_i} \lambda_{ij}} \right) + \kappa b, \tag{10}$$

where $\kappa \in \mathbb{R}^+$ is a hyperparameter, λ_{ij} is the j-th eigenvalue of Σ_i, and m_i is the eigenvalue number of Σ_i. This optimization idea is inspired by the framework in [1, 9], but deviates considerably from them: their optimization is for individual input test points, while our optimization method is a global optimization for the entire dataset. Consequently, our method circumvents the data independence problem in [1, 9].

4.3 Training with Data Augmentation and an Output Similarity-based Regularizer

As noted in the previous RS methods [8, 20], the smoothing noise influences the model performance. Therefore, to improve both the matching performance and the provable robustness, we adopt two strategies (for solving **C4**). The first one is data augmentation, which is motivated by [8]. We use a joint Gaussian distribution to generate augmented data, which matches the smoothing distribution we use to build the smoothed model. The second one is a regularizer based on the similarity among smoothed outputs. Since RS operates on the principle of "majority decision", minimizing the loss between each smoothed output and the true matching result is not adequate. We also need to ensure that smoothed outputs are as consistent as possible for a given input under multiple smoothing noises. To achieve this, we replicate the same input n times, apply data augmentation to the n data points, calculate their respective outputs, and then use the regularizer that penalizes the discrepancy among the n smoothed outputs. The total loss function can be written as follows:

$$\mathcal{L} = \frac{1}{n} \sum_i^n \mathcal{L}_{GM}\left(\mathbf{X}_i, \mathbf{X}_{gt}\right) + \beta \sum_{i,j}^n \left(1 - \frac{\mathbf{X}_i \cdot \mathbf{X}_j}{\mathbf{X}_{gt} \cdot \mathbf{X}_{gt}}\right), \tag{11}$$

where $\beta \in \mathbb{R}^+$ is a hyperparameter, $\mathbf{X}_{gt}$ is the true matching for input $(\mathbf{c}^1, \mathbf{c}^2, \mathbf{z}^1, \mathbf{z}^2)$, $\mathbf{X}_i$ and $\mathbf{X}_j$ are the smoothed outputs for $f(\mathbf{c}^1, \mathbf{c}^2, \mathbf{z}^1 + \varepsilon_1, \mathbf{z}^2 + \varepsilon_2)$ when $(\varepsilon_1, \varepsilon_2)$ are sampled by the i-th and j-th times, respectively. $\mathcal{L}_{GM}$ is the original loss function in GM methods, such as binary cross-entropy [36] and pixel offset regression [45]. In Eq. 11, the first term in our objective is the average matching loss, which leverages data augmentation from the joint Gaussian distribution to boost the matching accuracy under perturbations. The second term is a regularizer that enforces a similarity constraint among smoothed outputs, which will contribute to increasing p in Eq. 8 and improving the provable robustness.

Algorithm 1 Certified robustness for deep visual GM (CR-OSRS).

Input: $(\mathbf{c}^1, \mathbf{c}^2, \mathbf{z}^1, \mathbf{z}^2)$; base function f; original σ; sample times k_0; similarity threshold s; number of copies n; regularization hyperparameter β.
Output: Core output $\mathbf{X}_c$; evaluation results.

1: Use the data augmentation and regularizer in Sec. 4.3 to train a visual GM model, and then obtain function f_0.
2: Obtain $\mathbf{B}_1, \mathbf{B}_2, \Sigma_1, \Sigma_2$ described in Sec. 4.2 for perturbing keypoint position pair or image pair using an optimization method, and then obtain function g_0.
3: Sample k_0 number of input samples. For example, when perturbing keypoint position pair, we obtain the series: $\{(\mathbf{z}_1^{1'}, \mathbf{z}_1^{2'}), \ldots, (\mathbf{z}_{k_0}^{1'}, \mathbf{z}_{k_0}^{2'})\}$, where $\mathbf{z}_i^{1'} \sim \mathcal{N}\left(\mathbf{z}^1, \Sigma_1\right)$ and $\mathbf{z}_i^{2'} \sim \mathcal{N}\left(\mathbf{z}^2, \Sigma_2\right)$.
4: Predict the core output $\mathbf{X}_c$ and obtain the subspace $\mathcal{X}'$ using the first sampling series.
5: Sample $k = 10k_0$ number of input samples. For example, when perturbing the keypoint position pair, we obtain the series: $\{(\mathbf{z}_1^1, \mathbf{z}_1^2), \ldots, (\mathbf{z}_k^1, \mathbf{z}_k^2)\}$, where $\mathbf{z}_i^1 \sim \mathcal{N}\left(\mathbf{z}^1, \Sigma_1\right)$ and $\mathbf{z}_i^2 \sim \mathcal{N}\left(\mathbf{z}^2, \Sigma_2\right)$.
6: Calculate one-sided confidence lower bound $\underline{p}$ in Eq. 8 using the second sampling series.
7: **if** $\underline{p} < \frac{1}{2}$ **then**
8: This input cannot be robustly certified.
9: **else**
10: Obtain the sampling evaluation result and marginal radii evaluation result as in Sec. 4.4.
11: **end if**
12: **return** $\mathbf{X}_c$, evaluation results.

4.4 Quantify Certification

In Sec. 4.1, we derive Eq. 9 to characterize the certified space with paired perturbations, which is, however, challenging to quantify and compare. Moreover, previous studies have not tackled the problem of certification with multiple perturbations. To address this issue, we propose two quantity metrics to measure the certified robustness: sampling and marginal radii.

4.4.1 Sampling. According to Eq. 9, the certified robustness of g_0 increases when the certified space becomes larger, which means that more pairs of (δ_1, δ_2) satisfy the certified space. However, it is impractical to determine how many and which pairs of (δ_1, δ_2) satisfy Eq. 9, so we propose using a sampling approach to approximate the certified robustness. Specifically, we sample the perturbation pairs (δ_1, δ_2) and verify if they satisfy Eq. 9. The approximate certified robustness of g_0 is given by the probability of perturbation samples that satisfy Eq. 9.

4.4.2 Marginal Radii. Moreover, by fixing one of δ_1 and δ_2, we simplify the joint space in Eq. 9 to a marginal space, which facilitates robustness evaluation. Specifically, we set one of δ_1 and δ_2 to be a zero matrix and derive a simple expression for Eq. 9. As an example, we consider the case of setting δ_2 to a zero matrix as follows:

$$\|\delta_1^\top \mathbf{B}^{-1}\| < \left(\Phi^{-1}\left(\underline{p}\right)\right). \tag{12}$$

Lemma 4.2 (Eigenvalue Comparison). *For a real symmetric matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, with λ_{max} and λ_{min} as its maximum and minimum of eigenvalues, then $\forall \mathbf{X} \in \mathbb{R}^n, \lambda_{min}\mathbf{X}^\top\mathbf{X} \le \mathbf{X}^\top\mathbf{A}\mathbf{X} \le \lambda_{max}\mathbf{X}^\top\mathbf{X}$.*

Using Lemma 4.2, we have $\frac{1}{\lambda_{1\max}}\delta_1^\top\delta_1 \le \delta_1^\top\Sigma_1^{-1}\delta_1 \le \frac{1}{\lambda_{1\min}}\delta_1^\top\delta_1$ and derive minimum and maximum ℓ_2 norm radii from Eq. 12:

$$\|\delta_1\|_{\text{lower}} = \sqrt{\lambda_{1\min}}\left(\Phi^{-1}\left(\underline{p}\right)\right), \tag{13}$$

$$\|\delta_1\|_{\text{upper}} = \sqrt{\lambda_{1\max}}\left(\Phi^{-1}\left(\underline{p}\right)\right), \tag{14}$$

where $\lambda_{1\min}$ and $\lambda_{1\max}$ are the minimum and maximum eigenvalue of Σ_1. Inspired by [9], we can also use the ellipsoidal volume to measure the certified space. The volume of the ellipsoid is given by: $\mathcal{V}(\mathcal{R}) = r^m\sqrt{\pi^m}/\Gamma(m/2+1)\prod_{i=1}^m \xi_i$ [18], which we use to obtain a proxy ℓ_2 norm radius from Eq. 12:

$$\|\delta_1\|_{\text{volume}} = \left(\Phi^{-1}\left(\underline{p}\right)\right)\left(\sqrt{\pi}/\sqrt[m]{\Gamma(m/2+1)}\right)\sqrt[2m]{\prod_i^m \lambda_{1i}}, \tag{15}$$

where λ_{1i} is the i-th eigenvalue of Σ_1, and m is the number of eigenvalues. In summary, the certified space of Eq. 12 can be regarded as a hyperellipsoid with three radii: $\|\delta_1\|_{\text{lower}}$ as the minor axis, $\|\delta_1\|_{\text{upper}}$ as the major axis, and $\|\delta_1\|_{\text{volume}}$ as a proxy radius of a hypersphere whose volume is proportional to the volume of this hyperellipsoid. Eq. 13, Eq. 14 and Eq. 15 are all quantifiable forms: Eq. 13 is the lower bound radius that guarantees robustness against the worst-case adversaries, Eq. 14 is the upper bound radius that indicates the maximum potential to resist adversaries, and Eq. 15 is the closest assessment to the certified space. Similarly, by setting δ_1 as a zero matrix, we obtain the three radii of δ_2 ($\|\delta_2\|_{\text{lower}}$, $\|\delta_2\|_{\text{upper}}$, and $\|\delta_2\|_{\text{volume}}$) in the same manner. We can use these three radii of δ_1 and δ_2 to evaluate the probable robustness thoroughly.

5 EXPERIMENTS

This section presents the experimental settings, including datasets, GM solvers, parameter settings, and evaluation criteria. It then evaluates the robustness certification and matching performance of CR-OSRS and RS-GM for six common GM solvers using sampling evaluation and marginal radii evaluation as described in Sec. 4.4. Moreover, it performs ablation studies to illustrate the impact of different hyperparameters on the outcomes.

5.1 Experiments Settings

In this section, we apply CR-OSRS and RS-GM to transform base solvers into smoothed ones with certified robustness for comparison and analysis. Note that the original RS is not directly applicable for obtaining robustness certification of functions with paired inputs and structured outputs. For comparison, we propose to use RS-GM, a variant of RS that follows Theorem 4.1, with the only modification being the replacement of the smoothing distribution with an isotropic Gaussian distribution.

Following the GM literature [36], we evaluate our method on the Pascal VOC dataset [11] with Berkeley annotations [4], the Willow ObjectClass dataset [6] and SPair-71k dataset [27] for six GM solvers, which are: GMN [45], PCA-GM [35], CIE-H [43], NGMv2 [36], ASAR [29], COMMON [24]. Unless otherwise specified, we use the same data processing and hyperparameter settings as in Wang et al. [36]. All the experiments are conducted on CPU (Intel(R) Core(TM) i7-7820X CPU @ 3.60GHz) and GPU (GTX 2080 Ti GPU).

5.2 Robustness Certification Evaluation

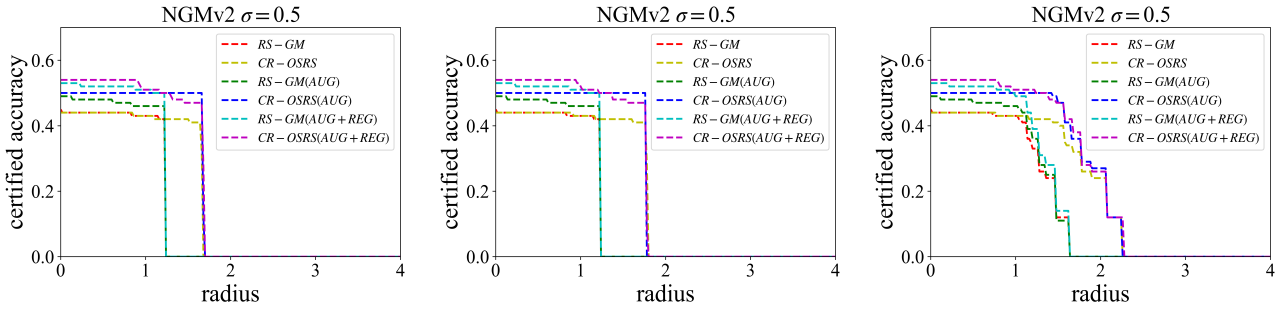
This section reports the results on the Pascal VOC dataset and SPair-71k dataset under keypoint position perturbations. The results under image pixel perturbations as well as on the Willow ObjectClass dataset are presented in Appendix F.

5.2.1 Sampling Evaluation. We use the sampling method presented in Sec. 4.4 to estimate the size of the certified space, where a larger space signifies stronger certified robustness. Specifically, we first randomly generate 1,000 pairs of (δ_1, δ_2) from a uniform distribution $\mathcal{U}(\sigma, \sigma)$. Then we insert the pairs into Eq. 9 and calculate the probability of pairs that satisfy Eq. 9. This probability for CR-OSRS with data augmentation is 83.5% and is 40.7% for RS-GM with data augmentation when $\sigma = 0.5$, $s = 0.9$ in Eq. 5, $\beta = 0.01$ and $n = 2$ in Eq. 11. This indicates that the certified space derived by CR-OSRS is larger than that derived by RS-GM, i.e., CR-OSRS achieves a better robustness guarantee.

5.2.2 Marginal Radii Evaluation. To evaluate the three marginal radii ($\|\delta\|_{\text{lower}}$, $\|\delta\|_{\text{upper}}$, and $\|\delta\|_{\text{volume}}$) proposed in Sec. 4.4, we propose two evaluation criteria: certified accuracy (CA) and average certified radius (ACR). Inspired by CA for classification [8], we define CA for GM as follows:

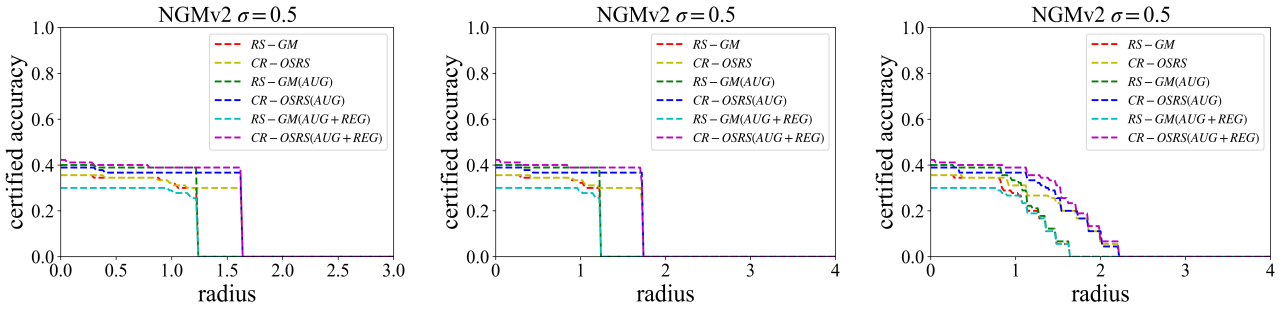
$$CA(R) = \mathbb{E}_{(x,\mathbf{X}_{gt})}\left[\mathbb{I}(\|\delta_1\| \ge R)\mathbb{I}(\|\delta_2\| \ge R)\mathbb{I}(g_0(x) = 1)\mathbb{I}(\mathbf{X_c} = \mathbf{X}_{gt})\right], \tag{16}$$

where $\mathbb{I}$ is an indicator function, $\|\delta_1\|$ and $\|\delta_2\|$ denote the radii calculated by Eq. 13, Eq. 14, or Eq. 15, R is the scale, g_0 represents the smoothed function defined in Eq. 7, x denotes an element in the test set. Meanwhile, to measure the certified robustness of the entire test set, we refer to the ACR mentioned in Zhai et al. [46] to



(a) CA and $\|\delta\|_{\text{lower}}$ (b) CA and $\|\delta\|_{\text{upper}}$ (c) CA and $\|\delta\|_{\text{volume}}$

Figure 2: CA achieved by CR-OSRS and RS-GM for NGMv2 on Pascal VOC when perturbing keypoint positions. "AUG" denotes data augmentation and "REG" denotes the regularizer in Eq. 11. It shows the result for original $\sigma = 0.5$, $s = 0.9$ in Eq. 5, $\beta = 0.01$ and $n = 2$ in Eq. 11.



(a) CA and $\|\delta\|_{\text{lower}}$ (b) CA and $\|\delta\|_{\text{upper}}$ (c) CA and $\|\delta\|_{\text{volume}}$

Figure 3: CA of RS-GM and CR-OSRS for NGMv2 on SPair-71k dataset when perturbing keypoint positions. It shows the result with original $\sigma = 0.5$, $s = 0.9$ in Eq. 5, $\beta = 0.01$ and $n = 2$ in Eq. 11.

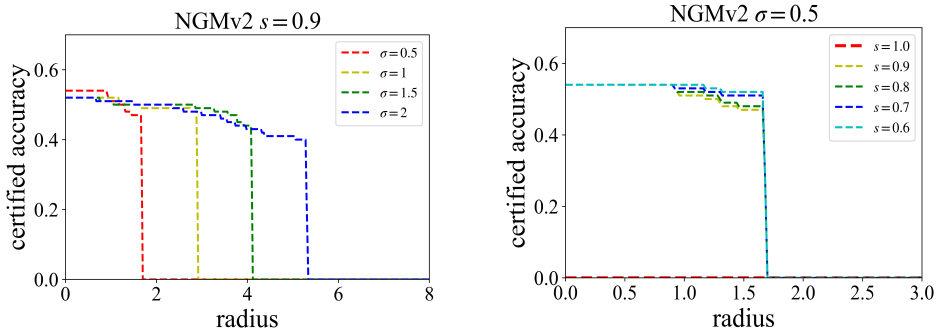
Table 1: ACR of CR-OSRS and RS-GM for six GM solvers on Pascal VOC under keypoint position perturbations. "AUG" denotes data augmentation and "REG" denotes the regularizer in Eq. 11. It shows the result for $\sigma = 0.5$, $s = 0.9$, $\beta = 0.01$ and $n = 2$.

	CR-OSRS+AUG+REG			RS-GM+AUG+REG		
	$\|\delta\|_{\text{lower}}$	$\|\delta\|_{\text{upper}}$	$\|\delta\|_{\text{volume}}$	$\|\delta\|_{\text{lower}}$	$\|\delta\|_{\text{upper}}$	$\|\delta\|_{\text{volume}}$
COMMON [24]	1.550	1.751	1.900	0.952	0.952	1.069
ASAR [29]	1.541	1.648	1.968	0.683	0.683	0.841
NGMv2 [36]	1.425	1.586	1.934	0.778	0.778	1.010
CIE-H [43]	0.987	1.167	1.354	0.572	0.572	0.731
PCA-GM [35]	0.954	1.158	1.340	0.546	0.546	0.686
GMN [45]	0.899	1.076	1.253	0.514	0.514	0.617

propose the ACR for GM as follows:

$$ACR = \mathbb{E}_{(x,\mathbf{X}_{gt})}\left[\|\delta_1\|\|\delta_2\|\mathbb{I}(g_0(x) = 1)\mathbb{I}(\mathbf{X_c} = \mathbf{X}_{gt})\right]. \quad (17)$$

We examine the relationship between CA and three marginal radii in Fig. 2 on the Pascal VOC dataset and Fig. 3 on the SPair-71k dataset. Specifically, we evaluate the performance of RS-GM and CR-OSRS under three training conditions: without data augmentation and regularizer, with data augmentation only, and with both data augmentation and regularizer, as defined in Eq. 11. Note that the performance under the training conditions with only the regularizer is identical to that without the data augmentation and regularizer. This is because, in the absence of data augmentation, the outputs corresponding to all copy data described in Sec. 4.3 are the same, and thus the regularizer is always zero. Consequently, the certification result is consistent with the baseline case RS-GM and CR-OSRS.



(a) CA and $\|\delta\|_{\text{lower}}$ when varying original σ. σ determines the scale of Σ_1 and Σ_2 that controls the trade-off between certified robustness and certified accuracy.

(b) CA and $\|\delta\|_{\text{lower}}$ when varying the original s. Reducing s enhances the certified robustness, as it enlarges the output subspace in Eq. 5 and relaxes the constraints on the output.

Figure 4: Projections for the certified accuracy given larger or smaller original σ and similarity threshold s. It shows the result for CR-OSRS trained by the data augmentation and regularizer with $\beta = 0.01$ and $n = 2$ for NGMv2 on Pascal VOC.

In Fig. 2, the curve of CR-OSRS is almost always above RS-GM in Fig. 2, implying higher certified robustness and matching accuracy. At the same time, it also demonstrates that the proposed data augmentation and regularizer are effective. In Fig. 3, the curve of CR-OSRS is also almost always above RS-GM, implying greater certified robustness and matching accuracy. However, we notice that applying both data augmentation and regularizer to RS-GM worsens the baseline RS-CM result without them. We hypothesize that this phenomenon may result from the excessive dispersion of the smoothed outputs, which prevents the loss after adding the regularizer from converging properly. By integrating Fig. 5, Fig. 3 and Table 1, we conclude that our method is applicable to various datasets and GM solvers.

To assess the overall provable robustness of the entire dataset, we compute ACR of CR-OSRS and RS-GM for six GM solvers in Table 1. It is evident that the ACR of CR-OSRS is higher than that of RS-GM and hence the overall provable robustness of CR-OSRS is superior to that of RS-GM for various GM solvers. Furthermore, we observe a positive correlation between the performance of the base and the smoothed models: the smoothed model demonstrates higher certified robustness as the performance of the base model improves.

5.3 Hyperparameter Analysis

Our method introduces the following hyperparameters: original σ, similarity threshold s for subspace construction as defined in Eq. 5, the constraint hyperparameter κ, number of copies n and regularization hyperparameter β as shown in Eq. 11 as well as k for Monte Carlo sampling. This subsection examines the effect of σ and s, and refers to Appendix F for the other hyperparameters. σ is varied from $\sigma \in \{0.5, 1.0, 1.5, 2.0\}$ and the certified accuracy with each σ is plotted in Fig. 4(a). Generally, a lower σ results in a higher certified accuracy and lower certified radii, while a higher σ allows for larger certified radii but a lower certified accuracy. s is varied from $s \in \{0.6, 0.7, 0.8, 0.9, 1.0\}$ and the certified accuracy achieved by CR-OSRS with each s is plotted in Fig. 4(b). When $s = 1$, the subspace in Eq. 5 degenerates into a single matrix, which implies a stringent robustness guarantee that the output remains invariant under any perturbation. However, as shown in Fig. 4(b), when $s = 1$, the accuracy is always zero, in line with the discussion in Sec. 4.1. The certification may fail or yield a small certified space due to the absence of a dominant output.

6 CONCLUSION AND OUTLOOK

This paper has introduced the first definition of certified robustness for visual graph matching and proposes a novel method, named CR-OSRS. This method uses the correlation between keypoints to construct a joint smoothing distribution and devises a global optimization method to determine the optimal smoothing range that balances provable robustness and matching performance. Furthermore, it presents a data augmentation technique based on the joint Gaussian distribution and a regularizer based on output similarity to improve model performance during the training phase. Then it derives an ℓ_2-norm certified space and suggests two quantitative methods (sampling and marginal radii) to address the challenge of quantifying the certified space. Finally, it conducts experiments on different GM solvers and datasets and achieves state-of-the-art robustness certification.

Future work. A significant direction is to enable robustness certification on combinatorial solvers whereby GM is one of such cases. We expect that this work can inspire subsequent research in this promising area i.e. learning for combinatorial optimization where theoretical results are welcomed given recent intensive empirical studies, e.g., [2, 41]. This requirement is especially pronounced in the setting of predict-then-optimization [14, 38], whereby the perception or forecasting model needs to be jointly learned with the backend decision model, and it remains open for certifying such complex yet practical machine learning systems, and we hope our work is a beneficial exploration in this area.

REFERENCES

[1] Motasem Alfarra, Adel Bibi, Philip Torr, and Bernard Ghanem. 2022. Data Dependent Randomized Smoothing. In *The 38th Conference on Uncertainty in Artificial Intelligence*.

[2] Yoshua Bengio, Andrea Lodi, and Antoine Prouvost. 2021. Machine learning for combinatorial optimization: a methodological tour d'horizon. *European Journal of Operational Research* 290, 2 (2021), 405–421.

[3] Aleksandar Bojchevski, Johannes Klicpera, and Stephan Günnemann. 2020. Efficient robustness certificates for discrete data: Sparsity-aware randomized smoothing for graphs, images and more. In *International Conference on Machine Learning*. PMLR, 1003–1013.

[4] Lubomir Bourdev and Jitendra Malik. 2009. Poselets: Body part detectors trained using 3d human pose annotations. In *2009 IEEE 12th International Conference on Computer Vision*. IEEE, 1365–1372.

[5] Ping-yeh Chiang, Michael Curry, Ahmed Abdelkader, Aounon Kumar, John Dickerson, and Tom Goldstein. 2020. Detection as regression: Certified object detection with median smoothing. *Advances in Neural Information Processing Systems* 33 (2020), 1275–1286.

[6] Minsu Cho, Karteek Alahari, and Jean Ponce. 2013. Learning Graphs to Match. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[7] Minsu Cho, Jungmin Lee, and Kyoung Mu Lee. 2010. Reweighted random walks for graph matching. In *European conference on Computer vision*. Springer, 492–505.

[8] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. 2019. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*. PMLR, 1310–1320.

[9] Francisco Eiras, Motasem Alfarra, M Pawan Kumar, Philip HS Torr, Puneet K Dokania, Bernard Ghanem, and Adel Bibi. 2021. Ancer: Anisotropic certification via sample-wise volume maximization. *arXiv preprint arXiv:2107.04570* (2021).

[10] Frank Emmert-Streib, Matthias Dehmer, and Yongtang Shi. 2016. Fifty years of graph matching, network alignment and network comparison. *Information sciences* 346 (2016), 180–197.

[11] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision* 88, 2 (2010), 303–338.

[12] Marc Fischer, Maximilian Baader, and Martin Vechev. 2021. Scalable certified segmentation via randomized smoothing. In *International Conference on Machine Learning*. PMLR, 3340–3351.

[13] Simon Geisler, Johanna Sommer, Jan Schuchardt, Aleksandar Bojchevski, and Stephan Günnemann. 2021. Generalization of Neural Combinatorial Solvers Through the Lens of Adversarial Robustness. *arXiv preprint arXiv:2110.10942* (2021).

[14] H Geng, H Ruan, R Wang, Y Li, Y Wang, L Chen, and J Yan. 2023. Rethinking and Benchmarking Predict-then-Optimize Paradigm for Combinatorial Optimization Problems. *arXiv preprint arXiv:2311.07633* (2023).

[15] Kenneth Hung and William Fithian. 2019. Rank verification for exponential families. *The Annals of Statistics* 47, 2 (2019), 758–782.

[16] Jinyuan Jia, Binghui Wang, Xiaoyu Cao, and Neil Zhenqiang Gong. 2020. Certified robustness of community detection against adversarial structural perturbation via randomized smoothing. In *Proceedings of The Web Conference 2020*. 2718–2724.

[17] Zetian Jiang, Tianzhe Wang, and Junchi Yan. 2021. Unifying Offline and Online Multi-graph Matching via Finding Shortest Paths on Supergraph. *TPAMI* 43, 10 (2021), 3648–3663.

[18] Maurice G Kendall. 2004. *A Course in the Geometry of n Dimensions*. Courier Corporation.

[19] Aounon Kumar and Tom Goldstein. 2021. Center Smoothing: Certified Robustness for Networks with Structured Outputs. *Advances in Neural Information Processing Systems* 34 (2021), 5560–5575.

[20] Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. 2019. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, 656–672.

[21] D. T. Lee and B. J. Schachter. 1980. Two Algorithms for Constructing a Delaunay Triangulation. *International Journal of Parallel Programming* 9, 3 (1980), 219–242.

[22] Guang-He Lee, Yang Yuan, Shiyu Chang, and Tommi Jaakkola. 2019. Tight certificates of adversarial robustness for randomly smoothed classifiers. *Advances in Neural Information Processing Systems* 32 (2019).

[23] Alexander J Levine and Soheil Feizi. 2021. Improved, deterministic smoothing for L_1 certified robustness. In *International Conference on Machine Learning*. PMLR, 6254–6264.

[24] Yijie Lin, Mouxing Yang, Jun Yu, Peng Hu, Changqing Zhang, and Xi Peng. 2023. Graph matching with bi-level noisy correspondence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 23362–23371.

[25] Eliane Maria Loiola, Nair Maria Maia de Abreu, Paulo Oswaldo Boaventura-Netto, Peter Hahn, and Tania Querido. 2007. A survey for the quadratic assignment problem. *European journal of operational research* 176, 2 (2007), 657–690.

[26] Han Lu, Zenan Li, Runzhong Wang, Qibing Ren, Junchi Yan, and Xiaokang Yang. 2021. Mind Your Solver! On Adversarial Attack and Defense for Combinatorial Optimization. *arXiv preprint arXiv:2201.00402* (2021).

[27] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. 2019. Spair-71k: A large-scale benchmark for semantic correspondence. *arXiv preprint arXiv:1908.10543* (2019).

[28] Jiaxiang Ren, Zijie Zhang, Jiayin Jin, Xin Zhao, Sixing Wu, Yang Zhou, Yelong Shen, Tianshi Che, Ruoming Jin, and Dejing Dou. 2021. Integrated defense for resilient graph matching. In *International Conference on Machine Learning*. PMLR, 8982–8997.

[29] Qibing Ren, Qingquan Bao, Runzhong Wang, and Junchi Yan. 2022. Appearance and Structure Aware Robust Deep Visual Graph Matching: Attack, Defense and Beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15263–15272.

[30] Michal Rolínek, Vít Musil, Anselm Paulus, Marin Vlastelica, Claudio Michaelis, and Georg Martius. 2020. Optimizing rank-based metrics with blackbox differentiation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7620–7630.

[31] Yu Rong, Wenbing Huang, Tingyang Xu, and Junzhou Huang. 2019. Dropedge: Towards deep graph convolutional networks on node classification. *arXiv preprint arXiv:1907.10903* (2019).

[32] Huaqing Shao, Lanjun Wang, and Junchi Yan. 2023. Robustness Certification for Structured Prediction with General Inputs via Safe Region Modeling in the Semimetric Output Space. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2010–2022.

[33] M. Vento. 2015. A long trip in the charming world of graphs for Pattern Recognition. *Pattern Recognition* (2015).

[34] Vladimir, G., Kim, Wilmot, Li, Niloy, J., Mitra, Stephen, and DiVerdi. 2012. Exploring collections of 3D models using fuzzy correspondences. *ACM Transactions on Graphics (TOG) - SIGGRAPH 2012 Conference Proceedings* 31, 4 (2012), 1–11.

[35] Runzhong Wang, Junchi Yan, and Xiaokang Yang. 2019. Learning combinatorial embedding networks for deep graph matching. In *Proceedings of the IEEE/CVF international conference on computer vision*. 3056–3065.

[36] Runzhong Wang, Junchi Yan, and Xiaokang Yang. 2021. Neural graph matching network: Learning lawler's quadratic assignment problem with extension to hypergraph and multiple-graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).

[37] Runzhong Wang, Junchi Yan, and Xiaokang Yang. 2023. Combinatorial Learning of Robust Deep Graph Matching: an Embedding based Approach. *TPAMI* 45, 6 (2023), 6984–7000.

[38] R. Wang, Y. Zhang, Z. Guo, T. Chen, X. Yang, and J. Yan. 2013. inSATNet: The Positive Linear Satisfiability Neural Networks.

[39] Eric Wong and Zico Kolter. 2018. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning*. PMLR, 5286–5295.

[40] Yu Xiong, Qingqiu Huang, Lingfeng Guo, Hang Zhou, Bolei Zhou, and Dahua Lin. 2019. A graph-based framework to bridge movies and synopses. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4592–4601.

[41] Junchi Yan, Shuang Yang, and Edwin Hancock. 2020. Learning Graph Matching and Related Combinatorial Optimization Problems. In *IJCAI*.

[42] Greg Yang, Tony Duan, J Edward Hu, Hadi Salman, Ilya Razenshteyn, and Jerry Li. 2020. Randomized smoothing of all shapes and sizes. In *International Conference on Machine Learning*. PMLR, 10693–10705.

[43] Tianshu Yu, Runzhong Wang, Junchi Yan, and Baoxin Li. 2019. Learning deep graph matching with channel-independent embedding and hungarian attention. In *International conference on learning representations*.

[44] Yu-Feng Yu, Guoxia Xu, Min Jiang, Hu Zhu, Dao-Qing Dai, and Hong Yan. 2019. Joint Transformation Learning via the L 2, 1-Norm Metric for Robust Graph Matching. *IEEE transactions on cybernetics* 51, 2 (2019), 521–533.

[45] Andrei Zanfir and Cristian Sminchisescu. 2018. Deep learning of graph matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2684–2693.

[46] Runtian Zhai, Chen Dan, Di He, Huan Zhang, Boqing Gong, Pradeep Ravikumar, Cho-Jui Hsieh, and Liwei Wang. 2020. Macer: Attack-free and scalable robust training via maximizing certified radius. *arXiv preprint arXiv:2001.02378* (2020).

[47] Shaofeng Zhang, Meng Liu, Junchi Yan, Hengrui Zhang, Lingxiao Huang, Xiaokang Yang, and Pinyan Lu. 2022. M-Mix: Generating Hard Negatives via Multi-sample Mixing for Contrastive Learning. In *Proceedings of Knowledge Discovery and Data Mining Conference*.

[48] Zijie Zhang, Zeru Zhang, Yang Zhou, Yelong Shen, Ruoming Jin, and Dejing Dou. 2020. Adversarial attacks on deep graph matching. *Advances in Neural Information Processing Systems* 33 (2020), 20834–20851.

[49] Daniel Zügner and Stephan Günnemann. 2020. Certifiable robustness of graph convolutional networks under structure perturbations. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 1656–1665.

A PROOFS FOR l_2 NORM

In this section, we present the full proofs for Theorem. 4.1. The main tool for our proofs is the Neyman-Pearson lemma for two variables, which we establish in Appendix A.1. Based on this lemma, we obtain the certified result in Appendix A.2. Finally, we provide the details of the linear transformation used for certification in Appendix A.3.

A.1 Neyman-Pearson for Two Variables

Lemma A.1 (Neyman-Pearson for two variables). *Let X_1 and Y_1 be random variables in $\mathbb{R}^d$ with densities μ_{X_1} and μ_{Y_1}. Then, let X_2 and Y_2 be random variables in $\mathbb{R}^d$ with densities μ_{X_2} and μ_{Y_2}. Let $h : \mathbb{R}^d \times \mathbb{R}^d \to \{0, 1\}$ be any deterministic or random function with an input pair. Then:*

1. If $\mathcal{S}_1 \times \mathcal{S}_2 = \left\{z_1 \in \mathbb{R}^d, z_2 \in \mathbb{R}^d : \frac{\mu_{Y_1}(z_1)\mu_{Y_2}(z_2)}{\mu_{X_1}(z_1)\mu_{X_2}(z_2)} \leq t\right\}$ for some $t > 0$ and $P(h(X_1, X_2) = 1) \geq P((X_1, X_2) \in \mathcal{S}_1 \times \mathcal{S}_2)$, then $P(h(Y_1, Y_2) = 1) \geq P((Y_1, Y_2) \in \mathcal{S}_1 \times \mathcal{S}_2)$.

2. If $\mathcal{S}_1 \times \mathcal{S}_2 = \left\{z_1 \in \mathbb{R}^d, z_2 \in \mathbb{R}^d : \frac{\mu_{Y_1}(z_1)\mu_{Y_2}(z_2)}{\mu_{X_1}(z_1)\mu_{X_2}(z_2)} \geq t\right\}$ for some $t > 0$ and $P(h(X_1, X_2) = 1) \leq P((X_1, X_2) \in \mathcal{S}_1 \times \mathcal{S}_2)$, then $P(h(Y_1, Y_2) = 1) \leq P((Y_1, Y_2) \in \mathcal{S}_1 \times \mathcal{S}_2)$.

Proof. We denote the complement of $\mathcal{S}_1 \times \mathcal{S}_2$ as $\mathcal{S}^c$.

$$
\begin{aligned}
&P(h(Y_1, Y_2) = 1) - P((Y_1, Y_2) \in \mathcal{S}_1 \times \mathcal{S}_2) \\
&= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} h(1 \mid z_1, z_2)\mu_{Y_1}(z_1)\mu_{Y_2}(z_2)dz_1dz_2 - \int\int_{\mathcal{S}_1\times\mathcal{S}_2} \mu_{Y_1}(z_1)\mu_{Y_2}(z_2)dz_1dz_2 \\
&= \left[\int\int_{\mathcal{S}_1\times\mathcal{S}_2} h(1 \mid z_1, z_2)\mu_{Y_1}(z_1)\mu_{Y_2}(z_2)dz_1dz_2 + \int\int_{\mathcal{S}^c} h(1 \mid z_1, z_2)\mu_{Y_1}(z_1)\mu_{Y_2}(z_2)dz_1dz_2\right] \\
&- \left[\int\int_{\mathcal{S}_1\times\mathcal{S}_2} h(1 \mid z_1, z_2)\mu_{Y_1}(z_1)\mu_{Y_2}(z_2)dz_1dz_2 + \int\int_{\mathcal{S}_1\times\mathcal{S}_2} h(0 \mid z_1, z_2)\mu_{Y_1}(z_1)\mu_{Y_2}(z_2)dz_1dz_2\right] \\
&= \int\int_{\mathcal{S}^c} h(1 \mid z_1, z_2)\mu_{Y_1}(z_1)\mu_{Y_2}(z_2)dz_1dz_2 - \int\int_{\mathcal{S}_1\times\mathcal{S}_2} h(0 \mid z_1, z_2)\mu_{Y_1}(z_1)\mu_{Y_2}(z_2)dz_1dz_2 \\
&\geq t\left[\int\int_{\mathcal{S}^c} h(1 \mid z_1, z_2)\mu_{X_1}(z_1)\mu_{X_2}(z_2)dz_1dz_2 - \int\int_{\mathcal{S}_1\times\mathcal{S}_2} h(0 \mid z_1, z_2)\mu_{X_1}(z_1)\mu_{X_2}(z_2)dz_1dz_2\right] \\
&= t[\int\int_{\mathcal{S}^c} h(1 \mid z_1, z_2)\mu_{X_1}(z_1)\mu_{X_2}(z_2)dz_1dz_2 + \int\int_{\mathcal{S}_1\times\mathcal{S}_2} h(1 \mid z_1, z_2)\mu_{X_1}(z_1)\mu_{X_2}(z_2)dz_1dz_2 \\
&- \int\int_{\mathcal{S}_1\times\mathcal{S}_2} h(1 \mid z_1, z_2)\mu_{X_1}(z_1)\mu_{X_2}(z_2)dz_1dz_2 - \int\int_{\mathcal{S}_1\times\mathcal{S}_2} h(0 \mid z_1, z_2)\mu_{X_1}(z_1)\mu_{X_2}(z_2)dz_1dz_2] \\
&= t\left[\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} h(1 \mid z_1, z_2)\mu_{X_1}(z_1)\mu_{X_2}(z_2)dz_1dz_2 - \int\int_{\mathcal{S}_1\times\mathcal{S}_2} \mu_{X_1}(z_1)\mu_{X_2}(z_2)dz_1dz_2\right] \\
&= t[P(h(X_1, X_2) = 1) - P((X_1, X_2) \in \mathcal{S}_1 \times \mathcal{S}_2)] \\
&\geq 0
\end{aligned}
$$

□

Next, we prove Lemma A.2, which is a special case of Lemma A.1 and states the Neyman-Pearson lemma for two joint Gaussian noise variables.

Lemma A.2 (Neyman-Pearson for Two Joint Gaussian Noise). *Let $X_1 \sim \mathcal{N}(x_1, \Sigma_1)$, $X_2 \sim \mathcal{N}(x_2, \Sigma_2)$ and $Y_1 \sim \mathcal{N}(x_1 + \delta_1, \Sigma_1)$, $Y_2 \sim \mathcal{N}(x_2 + \delta_2, \Sigma_2)$. Let $h : \mathbb{R}^d \times \mathbb{R}^d \to \{0, 1\}$ be any deterministic or random function. Then:*

1. If $\mathcal{S}_1 \times \mathcal{S}_2 = \left\{z_1 \in \mathbb{R}^d, z_2 \in \mathbb{R}^d : \delta_1^\top\Sigma_1^{-1}z_1 + \delta_2^\top\Sigma_2^{-1}z_2 \leq \beta\right\}$ for some β and $P(h(X_1, X_2) = 1) \geq P((X_1, X_2) \in \mathcal{S}_1 \times \mathcal{S}_2)$, then $P(h(Y_1, Y_2) = 1) \geq P((Y_1, Y_2) \in \mathcal{S}_1 \times \mathcal{S}_2)$.

2. If $\mathcal{S}_1 \times \mathcal{S}_2 = \left\{z_1 \in \mathbb{R}^d, z_2 \in \mathbb{R}^d : \delta_1^\top\Sigma_1^{-1}z_1 + \delta_2^\top\Sigma_2^{-1}z_2 \geq \beta\right\}$ for some β and $P(h(X_1, X_2) = 1) \leq P((X_1, X_2) \in \mathcal{S}_1 \times \mathcal{S}_2)$, then $P(h(Y_1, Y_2) = 1) \leq P((Y_1, Y_2) \in \mathcal{S}_1 \times \mathcal{S}_2)$.

Proof. This lemma is the special case of Neyman-Pearson for two variables when X_1, X_2, Y_1, and Y_2 are joint Gaussian noises. It suffices to simply show that for any β, there is some $t > 0$ for which:

$$
\begin{aligned}
\left\{z_1, z_2 : \delta_1^\top\Sigma_1^{-1}z_1 + \delta_2^\top\Sigma_2^{-1}z_2 \leq \beta\right\} &= \left\{z_1, z_2 : \frac{\mu_{Y_1}(z_1)\mu_{Y_2}(z_2)}{\mu_{X_1}(z_1)\mu_{X_2}(z_2)} \leq t\right\}, \\
\left\{z_1, z_2 : \delta_1^\top\Sigma_1^{-1}z_1 + \delta_2^\top\Sigma_2^{-1}z_2 \geq \beta\right\} &= \left\{z_1, z_2 : \frac{\mu_{Y_1}(z_1)\mu_{Y_2}(z_2)}{\mu_{X_1}(z_1)\mu_{X_2}(z_2)} \geq t\right\}.
\end{aligned} \tag{18}
$$

For ease of representation, we use $M_1 \in \mathbb{R}^{d\times d}$ (with element $m_{1_{ij}}$) instead of Σ_1^{-1} and $M_2 \in \mathbb{R}^{d\times d}$ (with element $m_{2_{ij}}$) instead of Σ_2^{-1}. The likelihood ratio for this choice of X_1, X_2, Y_1 and Y_2 turns out to be:

$$
\begin{aligned}
&\frac{\mu_{Y_1}(z_1)\mu_{Y_2}(z_2)}{\mu_{X_1}(z_1)\mu_{X_2}(z_2)} \\
&= \frac{\exp\left(-\frac{1}{2}(z_1-(x_1+\delta_1))^\top\Sigma_1^{-1}(z_1-(x_1+\delta_1))\right)}{\exp\left(-\frac{1}{2}(z_1-x_1)^\top\Sigma_1^{-1}(z_1-x_1)\right)} \times \frac{\exp\left(-\frac{1}{2}(z_2-(x_2+\delta_2))^\top\Sigma_2^{-1}(z_2-(x_2+\delta_2))\right)}{\exp\left(-\frac{1}{2}(z_2-x_2)^\top\Sigma_2^{-1}(z_2-x_2)\right)} \\
&= \frac{\exp\left(-\frac{1}{2}\sum_i^d\sum_j^d\left(z_{1_i}-\left(x_{1_i}+\delta_{1_i}\right)\right)m_{1_{ij}}\left(z_{1_j}-\left(x_{1_j}+\delta_{1_j}\right)\right)\right)}{\exp\left(-\frac{1}{2}\sum_i^d\sum_j^d\left(z_{1_i}-x_{1_i}\right)m_{1_{ij}}\left(z_{1_j}-x_{1_j}\right)\right)} \\
&\times \frac{\exp\left(-\frac{1}{2}\sum_i^d\sum_j^d\left(z_{2_i}-\left(x_{2_i}+\delta_{2_i}\right)\right)m_{2_{ij}}\left(z_{2_j}-\left(x_{2_j}+\delta_{2_j}\right)\right)\right)}{\exp\left(-\frac{1}{2}\sum_i^d\sum_j^d\left(z_{2_i}-x_{2_i}\right)m_{2_{ij}}\left(z_{2_j}-x_{2_j}\right)\right)} \\
&= \exp\left(\delta_1^\top\Sigma_1^{-1}z_1-\delta_1^\top\Sigma_1^{-1}x_1-\frac{1}{2}\delta_1^\top\Sigma_1^{-1}\delta_1\right)\times\exp\left(\delta_2^\top\Sigma_2^{-1}z_2-\delta_2^\top\Sigma_2^{-1}x_2-\frac{1}{2}\delta_2^\top\Sigma_2^{-1}\delta_2\right) \\
&= \exp\left(\delta_1^\top\Sigma_1^{-1}z_1+\delta_2^\top\Sigma_2^{-1}z_2-\delta_1^\top\Sigma_1^{-1}x_1-\frac{1}{2}\delta_1^\top\Sigma_1^{-1}\delta_1-\delta_2^\top\Sigma_2^{-1}x_2-\frac{1}{2}\delta_2^\top\Sigma_2^{-1}\delta_2\right) \\
&= \exp\left(\delta_1^\top\Sigma_1^{-1}z_1+\delta_2^\top\Sigma_2^{-1}z_2+b\right)\le t,
\end{aligned}
$$

where b is a constant, specifically $b=-\delta_1^\top\Sigma_1^{-1}x_1-\frac{1}{2}\delta_1^\top\Sigma_1^{-1}\delta_1-\delta_2^\top\Sigma_2^{-1}x_2-\frac{1}{2}\delta_2^\top\Sigma_2^{-1}\delta_2$. Therefore given any β, we may take $t=\exp(\beta+b)$ and obtain this correlation:

$$
\begin{aligned}
&\delta_1^\top\Sigma_1^{-1}z_1+\delta_2^\top\Sigma_2^{-1}z_2\le\beta \Longleftrightarrow \exp\left(\delta_1^\top\Sigma_1^{-1}z_1+\delta_2^\top\Sigma_2^{-1}z_2+b\right)\le t, \\
&\delta_1^\top\Sigma_1^{-1}z_1+\delta_2^\top\Sigma_2^{-1}z_2\ge\beta \Longleftrightarrow \exp\left(\delta_1^\top\Sigma_1^{-1}z_1+\delta_2^\top\Sigma_2^{-1}z_2+b\right)\ge t.
\end{aligned}
\tag{19}
$$

□

A.2 Proof of the Certified Robustness

This subsection presents the logic for proving robustness guarantees and derives the certified spaces for these guarantees in Eq. 9.

To show that $g_0\left(\mathbf{c}^1,\mathbf{c}^2,\mathbf{z}^1+\delta_1,\mathbf{z}^2+\delta_2\right)=1$, it follows from the definition of g_0 that we need to show that:

$$
P(f\left(\mathbf{c}^1,\mathbf{c}^2,\mathbf{z}^1+\varepsilon_1+\delta_1,\mathbf{z}^2+\varepsilon_2+\delta_2\right)\in\mathcal{X}')\ge P(f\left(\mathbf{c}^1,\mathbf{c}^2,\mathbf{z}^1+\varepsilon_1+\delta_1,\mathbf{z}^2+\varepsilon_2+\delta_2\right)\notin\mathcal{X}').
$$

We define two random variables:

$$
\begin{aligned}
I &:= \left(\mathbf{c}^1,\mathbf{c}^2,\mathbf{z}^1+\varepsilon_1,\mathbf{z}^2+\varepsilon_2\right)=\left(\mathbf{c}^1,\mathbf{c}^2,\mathcal{N}\left(\mathbf{z}^1,\Sigma_1\right),\mathcal{N}\left(\mathbf{z}^2,\Sigma_2\right)\right) \\
O &:= \left(\mathbf{c}^1,\mathbf{c}^2,\mathbf{z}^1+\varepsilon_1+\delta_1,\mathbf{z}^2+\varepsilon_2+\delta_2\right)=\left(\mathbf{c}^1,\mathbf{c}^2,\mathcal{N}\left(\mathbf{z}^1+\delta_1,\Sigma_1\right),\mathcal{N}\left(\mathbf{z}^2+\delta_2,\Sigma_2\right)\right).
\end{aligned}
$$

We know that:

$$
P(f(I)\in\mathcal{X}')\ge\underline{p}. \tag{20}
$$

Our goal is to show that

$$
P(f(O)\in\mathcal{X}')>P(f(O)\notin\mathcal{X}'). \tag{21}
$$

According to lemma A.2, we can define the half-spaces:

$$
\begin{aligned}
\mathcal{A} &= \left\{z_1,z_2:\delta_1^\top\Sigma_1^{-1}(z_1-\mathbf{z}^1)+\delta_2^\top\Sigma_2^{-1}(z_2-\mathbf{z}^2)\le\|\delta_1^\top\Sigma_1^{-1}\mathbf{B}_1+\delta_2^\top\Sigma_2^{-1}\mathbf{B}_2\|\Phi^{-1}\left(\underline{p}\right)\right\}, \\
\mathcal{B} &= \left\{z_1,z_2:\delta_1^\top\Sigma_1^{-1}(z_1-\mathbf{z}^1)+\delta_2^\top\Sigma_2^{-1}(z_2-\mathbf{z}^2)\ge\|\delta_1^\top\Sigma_1^{-1}\mathbf{B}_1+\delta_2^\top\Sigma_2^{-1}\mathbf{B}_2\|\Phi^{-1}\left(\underline{p}\right)\right\}.
\end{aligned}
$$

Claim 1 shows that $P(I\in\mathcal{A})=\underline{p}$, therefore we can obtain $P(f(I)\in\mathcal{X}')\ge P(I\in\mathcal{A})$. Hence we may apply Lemma A.2 to conclude:

$$
P(f(O)\in\mathcal{X}')\ge P(O\in\mathcal{A}). \tag{22}
$$

Similarly, we obtain $P(f(I)\notin\mathcal{X}')\le P(I\in\mathcal{B})$. Hence we may apply Lemma A.2 to conclude:

$$
P(f(O)\notin\mathcal{X}')\le P(O\in\mathcal{B}). \tag{23}
$$

Combining Eq. 22 and 23, we can obtain the conditions of Eq. 21:

$$
P(f(O)\in\mathcal{X}')\ge P(O\in\mathcal{A})>P(O\in\mathcal{B})\ge P(f(O)\notin\mathcal{X}'). \tag{24}
$$

According to Claim 3 and Claim 4, we can obtain $P(O \in \mathcal{A})$ and $P(O \in \mathcal{B})$ as:

$$\begin{aligned} P(O \in \mathcal{A}) &= \Phi\left(\Phi^{-1}\left(\underline{p}\right) - \frac{\delta_1^\top \Sigma_1^{-1}\delta_1 + \delta_2^\top \Sigma_2^{-1}\delta_2}{\|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|}\right), \\ P(O \in \mathcal{B}) &= \Phi\left(-\Phi^{-1}\left(\underline{p}\right) + \frac{\delta_1^\top \Sigma_1^{-1}\delta_1 + \delta_2^\top \Sigma_2^{-1}\delta_2}{\|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|}\right). \end{aligned} \tag{25}$$

Finally, we obtain that $P(O \in \mathcal{A}) > P(O \in \mathcal{B})$ if and only if:

$$\frac{\delta_1^\top \Sigma_1^{-1}\delta_1 + \delta_2^\top \Sigma_2^{-1}\delta_2}{\|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|} < \Phi^{-1}\left(\underline{p}\right).$$

A.3 Linear Transformation and Derivation

This subsection begins with Lemma A.3, which is the main tool for deriving all claims. Then, we present the proof process of claims, which is applied in Sec. A.2.

Lemma A.3 (Joint Gaussian Distribution). *If there is a random matrix $X \sim \mathcal{N}(\mu, \Sigma)$, where $\mu \in \mathbb{R}^n$ is the mean matrix. A positive semi-definite real symmetric matrix $\Sigma \in \mathbb{S}_{++}^{n\times n}$ is the covariance matrix of X. There is a full rank matrix $\mathbf{B} \in \mathbb{R}^{n\times n}$, which makes $X = \mathbf{B}Z + \mu$, $Z \sim \mathcal{N}(\mathbf{0}, I)$ and $\mathbf{B}^\top\mathbf{B} = \Sigma$.*

We obtain four claims based on linear transformation:

Claim 1. $P(I \in \mathcal{A}) = \underline{p}$

Proof. Recall that $\mathcal{A} = \left\{z_1, z_2 : \delta_1^\top \Sigma_1^{-1}(z_1 - \mathbf{z}^1) + \delta_2^\top \Sigma_2^{-1}(z_2 - \mathbf{z}^2) \le \|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|\Phi^{-1}\left(\underline{p}\right)\right\}$, according to lemma A.3, we can obtain:

$$\begin{aligned} P(I \in \mathcal{A}) &= P\left(\delta_1^\top \Sigma_1^{-1}(\mathcal{N}\left(\mathbf{z}^1, \Sigma_1\right) - \mathbf{z}^1) + \delta_2^\top \Sigma_2^{-1}(\mathcal{N}\left(\mathbf{z}^2, \Sigma_2\right) - \mathbf{z}^2) \le \|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|\Phi^{-1}\left(\underline{p}\right)\right) \\ &= P\left(\delta_1^\top \Sigma_1^{-1}\mathcal{N}\left(0, \Sigma_1\right) + \delta_2^\top \Sigma_2^{-1}\mathcal{N}\left(0, \Sigma_2\right) \le \|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|\Phi^{-1}\left(\underline{p}\right)\right) \\ &= P\left(\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1\mathcal{N}\left(0, I\right) + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\mathcal{N}\left(0, I\right) \le \|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|\Phi^{-1}\left(\underline{p}\right)\right) \\ &= P\left(\|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|\mathcal{N}\left(0, 1\right) \le \|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|\Phi^{-1}\left(\underline{p}\right)\right) \\ &= \Phi\left(\Phi^{-1}\left(\underline{p}\right)\right) \\ &= \underline{p}. \end{aligned}$$

□

Claim 2. $P(I \in \mathcal{B}) = 1 - \underline{p}$

Proof. Recall that $\mathcal{B} = \left\{z_1, z_2 : \delta_1^\top \Sigma_1^{-1}(z_1 - \mathbf{z}^1) + \delta_2^\top \Sigma_2^{-1}(z_2 - \mathbf{z}^2) \ge \|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|\Phi^{-1}\left(\underline{p}\right)\right\}$, according to lemma A.3, we can obtain:

$$\begin{aligned} P(I \in \mathcal{B}) &= P\left(\delta_1^\top \Sigma_1^{-1}(\mathcal{N}\left(\mathbf{z}^1, \Sigma_1\right) - \mathbf{z}^1) + \delta_2^\top \Sigma_2^{-1}(\mathcal{N}\left(\mathbf{z}^2, \Sigma_2\right) - \mathbf{z}^2) \ge \|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|\Phi^{-1}\left(\underline{p}\right)\right) \\ &= P\left(\delta_1^\top \Sigma_1^{-1}\mathcal{N}\left(0, \Sigma_1\right) + \delta_2^\top \Sigma_2^{-1}\mathcal{N}\left(0, \Sigma_2\right) \ge \|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|\Phi^{-1}\left(\underline{p}\right)\right) \\ &= P\left(\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1\mathcal{N}\left(0, I\right) + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\mathcal{N}\left(0, I\right) \ge \|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|\Phi^{-1}\left(\underline{p}\right)\right) \\ &= P\left(\|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|\mathcal{N}\left(0, 1\right) \ge \|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|\Phi^{-1}\left(\underline{p}\right)\right) \\ &= 1 - \Phi\left(\Phi^{-1}\left(\underline{p}\right)\right) \\ &= 1 - \underline{p}. \end{aligned}$$

□

Claim 3. $P(O \in \mathcal{A}) = \Phi\left(\Phi^{-1}\left(\underline{p}\right) - \frac{\delta_1^\top \Sigma_1^{-1}\delta_1 + \delta_2^\top \Sigma_2^{-1}\delta_2}{\|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|}\right)$

Proof. Recall that $\mathcal{A} = \left\{z_1, z_2 : \delta_1^\top \Sigma_1^{-1}(z_1 - \mathbf{z}^1) + \delta_2^\top \Sigma_2^{-1}(z_2 - \mathbf{z}^2) \le \|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|\Phi^{-1}\left(\underline{p}\right)\right\}$ and $O \sim \left(\mathbf{c}^1, \mathbf{c}^2, \mathcal{N}\left(\mathbf{z}^1 + \delta_1, \Sigma_1\right), \mathcal{N}\left(\mathbf{z}^2 + \delta_2, \Sigma_2\right)\right)$, according to lemma A.3, we can obtain:

$$
\begin{aligned}
&P(O \in \mathcal{A}) \\
&= P\left(\delta_1^\top \Sigma_1^{-1}(\mathcal{N}\left(\mathbf{z}^1 + \delta_1, \Sigma_1\right) - \mathbf{z}^1) + \delta_2^\top \Sigma_2^{-1}(\mathcal{N}\left(\mathbf{z}^2 + \delta_2, \Sigma_2\right) - \mathbf{z}^2) \le \|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|\Phi^{-1}\left(\underline{p}\right)\right) \\
&= P\left(\delta_1^\top \Sigma_1^{-1}\mathcal{N}\left(\delta_1, \Sigma_1\right) + \delta_2^\top \Sigma_2^{-1}\mathcal{N}\left(\delta_2, \Sigma_2\right) \le \|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|\Phi^{-1}\left(\underline{p}\right)\right) \\
&= P\left(\delta_1^\top \Sigma_1^{-1}(\mathbf{B}_1\mathcal{N}\left(0, I\right) + \delta_1) + \delta_2^\top \Sigma_2^{-1}(\mathbf{B}_2\mathcal{N}\left(0, I\right) + \delta_2) \le \|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|\Phi^{-1}\left(\underline{p}\right)\right) \\
&= P\left(\|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|\mathcal{N}\left(0, 1\right) + \delta_1^\top \Sigma_1^{-1}\delta_1 + \delta_2^\top \Sigma_2^{-1}\delta_2 \le \|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|\Phi^{-1}\left(\underline{p}\right)\right) \\
&= P\left(\mathcal{N}\left(0, 1\right) \le \Phi^{-1}\left(\underline{p}\right) - \frac{\delta_1^\top \Sigma_1^{-1}\delta_1 + \delta_2^\top \Sigma_2^{-1}\delta_2}{\|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|}\right) \\
&= \Phi\left(\Phi^{-1}\left(\underline{p}\right) - \frac{\delta_1^\top \Sigma_1^{-1}\delta_1 + \delta_2^\top \Sigma_2^{-1}\delta_2}{\|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|}\right).
\end{aligned}
$$

□

Claim 4. $P(O \in \mathcal{B}) = \Phi\left(-\Phi^{-1}\left(\underline{p}\right) + \frac{\delta_1^\top \Sigma_1^{-1}\delta_1 + \delta_2^\top \Sigma_2^{-1}\delta_2}{\|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|}\right)$

Proof. Recall that $\mathcal{B} = \left\{z_1, z_2 : \delta_1^\top \Sigma_1^{-1}(z_1 - \mathbf{z}^1) + \delta_2^\top \Sigma_2^{-1}(z_2 - \mathbf{z}^2) \ge \|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|\Phi^{-1}\left(\underline{p}\right)\right\}$ and $O \sim \left(\mathbf{c}^1, \mathbf{c}^2, \mathcal{N}\left(\mathbf{z}^1 + \delta_1, \Sigma_1\right), \mathcal{N}\left(\mathbf{z}^2 + \delta_2, \Sigma_2\right)\right)$, according to lemma A.3, we can obtain:

$$
\begin{aligned}
&P(O \in \mathcal{B}) \\
&= P\left(\delta_1^\top \Sigma_1^{-1}(\mathcal{N}\left(\mathbf{z}^1 + \delta_1, \Sigma_1\right) - \mathbf{z}^1) + \delta_2^\top \Sigma_2^{-1}(\mathcal{N}\left(\mathbf{z}^2 + \delta_2, \Sigma_2\right) - \mathbf{z}^2) \ge \|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|\Phi^{-1}\left(\underline{p}\right)\right) \\
&= P\left(\delta_1^\top \Sigma_1^{-1}\mathcal{N}\left(\delta_1, \Sigma_1\right) + \delta_2^\top \Sigma_2^{-1}\mathcal{N}\left(\delta_2, \Sigma_2\right) \ge \|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|\Phi^{-1}\left(\underline{p}\right)\right) \\
&= P\left(\delta_1^\top \Sigma_1^{-1}(\mathbf{B}_1\mathcal{N}\left(0, I\right) + \delta_1) + \delta_2^\top \Sigma_2^{-1}(\mathbf{B}_2\mathcal{N}\left(0, I\right) + \delta_2) \ge \|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|\Phi^{-1}\left(\underline{p}\right)\right) \\
&= P\left(\|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|\mathcal{N}\left(0, 1\right) + \delta_1^\top \Sigma_1^{-1}\delta_1 + \delta_2^\top \Sigma_2^{-1}\delta_2 \ge \|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|\Phi^{-1}\left(\underline{p}\right)\right) \\
&= P\left(\mathcal{N}\left(0, 1\right) \ge \Phi^{-1}\left(\underline{p}\right) - \frac{\delta_1^\top \Sigma_1^{-1}\delta_1 + \delta_2^\top \Sigma_2^{-1}\delta_2}{\|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|}\right) \\
&= \Phi\left(-\Phi^{-1}\left(\underline{p}\right) + \frac{\delta_1^\top \Sigma_1^{-1}\delta_1 + \delta_2^\top \Sigma_2^{-1}\delta_2}{\|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|}\right).
\end{aligned}
$$

□

B PROOF FOR l_1 NORM

In this section, we present the full proofs for the robustness guarantee for l_1 norm. The main tool for our proofs is the Neyman-Pearson lemma for two variables, which we establish in Lemma A.1. Next, we prove Lemma B.1, which is a special case of Lemma A.1 and states the Neyman-Pearson lemma for two Laplace noise variables. Based on this lemma, we obtain the certified result in Appendix B.1.

Lemma B.1 (Neyman-Pearson for Two Laplace Noise). *Let $X_1 \sim x_1 + \mathcal{L}(\lambda_1)$, $X_2 \sim x_2 + \mathcal{L}(\lambda_2)$ and $Y_1 \sim x_1 + \mathcal{L}(\lambda_1) + \delta_1$, $Y_2 \sim x_2 + \mathcal{L}(\lambda_2) + \delta_2$. Let $h : \mathbb{R}^d \times \mathbb{R}^d \to \{0, 1\}$ be any deterministic or random function. Then:*

1. If $\mathcal{S}_1 \times \mathcal{S}_2 = \left\{z_1 \in \mathbb{R}^d, z_2 \in \mathbb{R}^d : \frac{1}{\lambda_1}(\|\boldsymbol{z_1} - \boldsymbol{\delta_1}\|_1 - \|\boldsymbol{z_1}\|_1) + \frac{1}{\lambda_2}(\|\boldsymbol{z_2} - \boldsymbol{\delta_2}\|_1 - \|\boldsymbol{z_2}\|_1)\right) \ge \beta\}$ for some β and $P(h(X_1, X_2) = 1) \ge P((X_1, X_2) \in \mathcal{S}_1 \times \mathcal{S}_2)$, then $P(h(Y_1, Y_2) = 1) \ge P((Y_1, Y_2) \in \mathcal{S}_1 \times \mathcal{S}_2)$.

2. If $\mathcal{S}_1 \times \mathcal{S}_2 = \left\{z_1 \in \mathbb{R}^d, z_2 \in \mathbb{R}^d : \frac{1}{\lambda_1}(\|\boldsymbol{z_1} - \boldsymbol{\delta_1}\|_1 - \|\boldsymbol{z_1}\|_1) + \frac{1}{\lambda_2}(\|\boldsymbol{z_2} - \boldsymbol{\delta_2}\|_1 - \|\boldsymbol{z_2}\|_1)\right) \le \beta\}$ for some β and $P(h(X_1, X_2) = 1) \le P((X_1, X_2) \in \mathcal{S}_1 \times \mathcal{S}_2)$, then $P(h(Y_1, Y_2) = 1) \le P((Y_1, Y_2) \in \mathcal{S}_1 \times \mathcal{S}_2)$.

Proof. *This lemma is the special case of Neyman-Pearson for two variables when X_1, X_2, Y_1, and Y_2 are Laplace noises. It suffices to simply show that for any β, there is some $t > 0$ for which:*

$$\begin{aligned}
&\left\{z_1, z_2 : \frac{1}{\lambda_1}(\|z_1 - \boldsymbol{\delta_1}\|_1 - \|z_1\|_1) + \frac{1}{\lambda_2}(\|z_2 - \boldsymbol{\delta_2}\|_1 - \|z_2\|_1) \geq \beta\right\} = \left\{z_1, z_2 : \frac{\mu_{Y_1}(z_1)\mu_{Y_2}(z_2)}{\mu_{X_1}(z_1)\mu_{X_2}(z_2)} \leq t\right\}, \\
&\left\{z_1, z_2 : \frac{1}{\lambda_1}(\|z_1 - \boldsymbol{\delta_1}\|_1 - \|z_1\|_1) + \frac{1}{\lambda_2}(\|z_2 - \boldsymbol{\delta_2}\|_1 - \|z_2\|_1) \leq \beta\right\} = \left\{z_1, z_2 : \frac{\mu_{Y_1}(z_1)\mu_{Y_2}(z_2)}{\mu_{X_1}(z_1)\mu_{X_2}(z_2)} \geq t\right\}.
\end{aligned} \tag{26}$$

$$\begin{aligned}
&\frac{\mu_{Y_1}(z_1)\mu_{Y_2}(z_2)}{\mu_{X_1}(z_1)\mu_{X_2}(z_2)} \\
&= \frac{\exp\left(-\frac{1}{\lambda_1}\|z_1 - \boldsymbol{\delta_1}\|_1\right)\exp\left(-\frac{1}{\lambda_2}\|z_2 - \boldsymbol{\delta_2}\|_1\right)}{\exp\left(-\frac{1}{\lambda_1}\|z_1\|_1\right)\exp\left(-\frac{1}{\lambda_2}\|z_2\|_1\right)} \\
&= \exp\left(-\frac{1}{\lambda_1}(\|z_1 - \boldsymbol{\delta_1}\|_1 - \|z_1\|_1) - \frac{1}{\lambda_2}(\|z_2 - \boldsymbol{\delta_2}\|_1 - \|z_2\|_1)\right)
\end{aligned}$$

By choosing $\beta = -\log(t)$, we can derive that

$$\begin{aligned}
&\frac{1}{\lambda_1}(\|z_1 - \boldsymbol{\delta_1}\|_1 - \|z_1\|_1) + \frac{1}{\lambda_2}(\|z_2 - \boldsymbol{\delta_2}\|_1 - \|z_2\|_1) \geq \beta \Longleftrightarrow \frac{\mu_{Y_1}(z_1)\mu_{Y_2}(z_2)}{\mu_{X_1}(z_1)\mu_{X_2}(z_2)} \leq t, \\
&\frac{1}{\lambda_1}(\|z_1 - \boldsymbol{\delta_1}\|_1 - \|z_1\|_1) + \frac{1}{\lambda_2}(\|z_2 - \boldsymbol{\delta_2}\|_1 - \|z_2\|_1) \leq \beta \Longleftrightarrow \frac{\mu_{Y_1}(z_1)\mu_{Y_2}(z_2)}{\mu_{X_1}(z_1)\mu_{X_2}(z_2)} \geq t.
\end{aligned}$$

□

B.1 Proof of the Certified Robustness for l_1 norm

Theorem B.2 (ℓ_1 **norm certified space for visual GM**). *Let f be a matching function, f_0 and g_0 be defined as in Eq. 6 and Eq. 7, $\varepsilon_1 \sim \mathcal{L}(\lambda_1)$, $\varepsilon_2 \sim \mathcal{L}(\lambda_2)$. Suppose $\underline{p} \in (\frac{1}{2}, 1]$ satisfy:*

$$\begin{aligned}
&P(f_0\left(\mathbf{c}^1, \mathbf{c}^2, \mathbf{z}^1 + \varepsilon_1, \mathbf{z}^2 + \varepsilon_2\right) = 1) = \\
&P(f(\mathbf{c}^1, \mathbf{c}^2, \mathbf{z}^1 + \varepsilon_1, \mathbf{z}^2 + \varepsilon_2) \in \mathcal{X}') = p \geq \underline{p}.
\end{aligned} \tag{27}$$

Then we obtain the ℓ_1 norm certified space for the perturbation pair (δ_1, δ_2):

$$\frac{\|\delta_1\|_1}{\lambda_1} + \frac{\|\delta_2\|_1}{\lambda_2} \leq -\log\left[2\left(1 - \underline{p}\right)\right], \tag{28}$$

which guarantees $g_0\left(\mathbf{c}^1, \mathbf{c}^2, \mathbf{z}^1 + \delta_1, \mathbf{z}^2 + \delta_2\right) = 1$.

To show that $g_0\left(\mathbf{c}^1, \mathbf{c}^2, \mathbf{z}^1 + \delta_1, \mathbf{z}^2 + \delta_2\right) = 1$, it follows from the definition of g_0 that we need to show that:

$$P(f\left(\mathbf{c}^1, \mathbf{c}^2, \mathbf{z}^1 + \varepsilon_1 + \delta_1, \mathbf{z}^2 + \varepsilon_2 + \delta_2\right) \in \mathcal{X}') \geq P(f\left(\mathbf{c}^1, \mathbf{c}^2, \mathbf{z}^1 + \varepsilon_1 + \delta_1, \mathbf{z}^2 + \varepsilon_2 + \delta_2\right) \notin \mathcal{X}').$$

We define two random variables:

$$\begin{aligned}
I &:= \left(\mathbf{c}^1, \mathbf{c}^2, \mathbf{z}^1 + \varepsilon_1, \mathbf{z}^2 + \varepsilon_2\right) \\
O &:= \left(\mathbf{c}^1, \mathbf{c}^2, \mathbf{z}^1 + \varepsilon_1 + \delta_1, \mathbf{z}^2 + \varepsilon_2 + \delta_2\right).
\end{aligned}$$

We know that:

$$P(f(I) \in \mathcal{X}') \geq \underline{p}. \tag{29}$$

Our goal is to show that

$$P(f(O) \in \mathcal{X}') > P(f(O) \notin \mathcal{X}'). \tag{30}$$

Denote $T(\mathbf{z}^1, \mathbf{z}^2) = \frac{1}{\lambda_1}(\|\mathbf{z}^1 - \boldsymbol{\delta_1}\|_1 - \|\mathbf{z}^1\|_1) + \frac{1}{\lambda_2}(\|\mathbf{z}^2 - \boldsymbol{\delta_2}\|_1 - \|\mathbf{z}^2\|_1)$. Use Triangle Inequality we can derive a bound for $T(\mathbf{z}^1, \mathbf{z}^2)$:

$$-\frac{\|\delta_1\|_1}{\lambda_1} - \frac{\|\delta_2\|_1}{\lambda_2} \leq T(\mathbf{z}^1, \mathbf{z}^2) \leq \frac{\|\delta_1\|_1}{\lambda_1} + \frac{\|\delta_2\|_1}{\lambda_2}. \tag{31}$$

Pick β' such that there exists $B' \subseteq \{z_1, z_2 : T(z_1, z_2) = \beta'\}$, and

$$P\left(I \in \left\{z_1, z_2 : T(z_1, z_2) < \beta'\right\} \cup B'\right) = 1 - \underline{p} = P(f(I) \notin \mathcal{X}')). \tag{32}$$

Define

$$S := \left\{z_1, z_2 : T(z_1, z_2) < \beta'\right\} \cup B', \tag{33}$$

so we also have $P(X \notin S) = p = P(f(I) \notin \mathcal{X}')$. Plug into Lemma B.1, we can get

$$\begin{aligned} P(Y \notin S) &\leq P(f(O) \in \mathcal{X}'), \\ P(Y \in S) &\geq P(f(O) \notin \mathcal{X}'). \end{aligned} \tag{34}$$

Then we can obtain

$$\begin{aligned} \mathbb{P}(Y \in S) &= \int \int_S [2\lambda_1]^{-d} [2\lambda_2]^{-d} \exp\left(-\frac{\|\mathbf{z}^1 - \boldsymbol{\delta}_1\|_1}{\lambda_1}\right) \exp\left(-\frac{\|\mathbf{z}^2 - \boldsymbol{\delta}_2\|_1}{\lambda_2}\right) \mathrm{d}\mathbf{z}^1 \mathrm{d}\mathbf{z}^2 \\ &= \int \int_S [2\lambda_1]^{-d} [2\lambda_2]^{-d} \exp\left(-\frac{\|\mathbf{z}^1\|_1}{\lambda_1}\right) \left(-\frac{\|\mathbf{z}^2\|_1}{\lambda_2}\right) \exp\left(-T(\mathbf{z}^1, \mathbf{z}^2)\right) \mathrm{d}\mathbf{z}^1 \mathrm{d}\mathbf{z}^2 \\ &\leq \exp\left(\frac{\|\delta_1\|_1}{\lambda_1} + \frac{\|\delta_2\|_1}{\lambda_2}\right) \int \int_S [2\lambda_1]^{-d} [2\lambda_2]^{-d} \exp\left(-\frac{\|\mathbf{z}^1\|_1}{\lambda_1}\right) \left(-\frac{\|\mathbf{z}^2\|_1}{\lambda_2}\right) \mathrm{d}\mathbf{z}^1 \mathrm{d}\mathbf{z}^2 \\ &= \exp\left(\frac{\|\delta_1\|_1}{\lambda_1} + \frac{\|\delta_2\|_1}{\lambda_2}\right) (1 - \underline{p}). \end{aligned} \tag{35}$$

Thus, if $\frac{\|\delta_1\|_1}{\lambda_1} + \frac{\|\delta_2\|_1}{\lambda_2} \leq -\log\left[2\left(1 - \underline{p}\right)\right]$, it holds that

$$\begin{aligned} P(Y \in S) &\leq \exp\left(\frac{\|\delta_1\|_1}{\lambda_1} + \frac{\|\delta_2\|_1}{\lambda_2}\right) (1 - \underline{p}) \\ &\exp\left(-\log\left[2\left(1 - \underline{p}\right)\right]\right) (1 - \underline{p}) \\ &= \frac{1}{2}. \end{aligned} \tag{36}$$

C SUMMARY OF RELATED METHODS

To present various methods of graph matching and certified robustness more clearly, we have categorized and reviewed the mainstream methods. Deep visual GM solvers aim to align the corresponding keypoints from different images based on node-to-node and edge-to-edge correlations. We introduce and compare the mainstream methods in Tab. 2. We then present some of the representative RS-based methods for certified robustness in Tab. 3, along with their applicable scenarios and features.

D METHODOLOGY SUPPLEMENT

In this section, we provide a supplement to the method described in Sec. 4. We first present the algorithm of the entire process, and then explain the construction and optimization of the joint Gaussian distribution under pixel perturbations.

D.1 Algorithm of the Entire Process

Alg. 1 consists of training and testing parts. In the training part, we use data augmentation and a regularizer based on the output similarity as introduced in Sec. 4.3 to train a model. In the testing part, we employ Monte Carlo sampling to estimate the certification result in practice. First, we construct and optimize the smoothing joint Gaussian distribution according to Sec. 4.2 and construct the smoothed model g_0. Second, we sample $(\varepsilon_1, \varepsilon_2)$ with k_0 times and obtain the core output $\mathbf{X}_c$ in Eq. 4 and subspace $\mathcal{X}'$ in Eq. 5. Then we sample $(\varepsilon_1, \varepsilon_2)$ with k times, and count how many outputs fall into the subspace $\mathcal{X}'$ to obtain the probability $\underline{p}$ in Eq. 8 and the certified space in Eq. 9. Finally, we use two quantitative methods as in Sec. 4.4 to obtain evaluation results.

Alg. 2 summarizes the updates for optimizing by solving Eq. 10 with K steps of stochastic gradient ascent. $\underline{p}$ is approximated by the Monte Carlo sampling algorithm in the subsequent certification process, as described in Alg. 1, but we simplify its approximation in the optimization method. As we only require a favorable trend rather than a very accurate $\underline{p}$ value here, we estimate it by sampling once. This method enhances the efficiency of the optimization method and also reduces the high variability in the gradient estimation due to multiple sampling. We fix the number of iterations for optimization to $K = 10$, the size of data used for optimization to $L = 100$, and set the original correlation parameter $b = 0.01$. Therefore, the entire optimization process is relatively fast and can be relatively easily applied to various visual GM models.

D.2 Smoothing Distribution for Perturbing Image Pixels

Due to the large number of image pixels, which far exceeds the number of keypoints, constructing correlation matrices between pixel points as in Sec. 4.2 is computationally expensive, not to mention that mining the correlation between pixels is not trivial. We, therefore, simplify $\mathbf{B}$ to σ as in Cohen et al. [8] under pixel perturbations and modify the optimization problem accordingly:

$$\arg\max_{\sigma} \Phi^{-1}\left(\underline{p}\right)\sigma. \tag{37}$$

Algorithm 2 Algorithm for optimization.

Input: L data; base function f; original σ; original b; iteration times K.
Output: $\mathbf{B}_1, \mathbf{B}_2, \Sigma_1, \Sigma_2$.
Initialize: $\sigma^0 \leftarrow \sigma, b^0 \leftarrow b$.
for $k = 0 \dots K-1$ **do**
  Calculate $\mathbf{B}_1^k, \mathbf{B}_2^k, \Sigma_1^k, \Sigma_2^k$ using σ^k and b^k according to Sec. 4.
  Initialize the sum of optimization goal O.
  for $l = 0 \dots L-1$ **do**
    Initialize k^{th} data.
    Sample $\varepsilon_1 \sim \mathcal{N}(0, \mathbf{B}_1^k), \varepsilon_2 \sim \mathcal{N}(0, \mathbf{B}_2^k)$.
    Calculate $\underline{p}$ according to Eq. 8 and eigenvalues of $\mathbf{B}_1^k, \mathbf{B}_2^k$, then calculate the optimization goal O^l as in Eq. 10.
    $O \leftarrow O + O^l$.
  end for
  $\sigma^{k+1}, b^{k+1} \leftarrow \nabla_{\sigma^k, b^k} O$.
end for
Calculate $\mathbf{B}_1$, $\mathbf{B}_2$, Σ_1, Σ_2 using σ^{K-1} and b^{K-1} according to Sec. 4.
return $\mathbf{B}_1, \mathbf{B}_2, \Sigma_1, \Sigma_2$.

D.3 Quantify Certification for l_1 norm

Moreover, by fixing one of δ_1 and δ_2 which is similar to in Sec. 4.4.2, we simplify the joint space in Eq. 28 to a marginal space, which facilitates robustness evaluation. Specifically, we set one of δ_1 and δ_2 to be a zero matrix and derive a simple expression for Eq. 28. As an example, we consider the case of setting δ_2 to a zero matrix as follows:

$$\|\delta_1\|_1 \leq -\lambda_1 \log\left[2\left(1-\underline{p}\right)\right]. \tag{38}$$

E ADDITIONAL EXPERIMENT SETTINGS

This section provides the details of the baseline for certification, GM solvers, which are supplementary to Sec. 5.1.

E.1 Baseline for Certification

This paper adopts modified RS [8] as the baseline method for the proposed CR-OSRS strategy, which is referred to as RS-GM. Unless otherwise stated, we follow the same experimental parameter settings as RS. We use the hypothesis test [15] as in [8] by using α to represent the probability of obtaining incorrect matching results. In this study, we set $\alpha = 0.001$, which ensures a high probability (99.9%) of certification. α can be arbitrarily small, so in theory our method is highly reliable. We choose the Monte Carlo sample number k in Alg. 1 to be 1000, which is smaller than the sample number for classifier certification, due to the low efficiency of the GM solver. Theoretically, increasing k would improve the certification results, but at the expense of the efficiency of the GM solver. We reveal the impact of different k on the experimental results in Appendix F.1.

E.2 Deep Graph Matching Solvers

This paper evaluates the proposed method on the Pascal VOC dataset [11] with Berkeley annotations [4], the Willow ObjectClass

Table 2: Summary of main existing literature in learning GM.

Method	Introduction
GMN [45]	The seminal work that employs a convolutional neural network to extract node features and constructs an end-to-end model with spectral matching.
PCA-GM [35]	Leveraging intra-graph and cross-graph structural information using graph convolutional networks.
CIE-H [43]	Enhancing end-to-end training by edge embedding and Hungarian-based attention mechanism.
NGMv2 [36]	Developing a matching-aware graph convolution scheme with Sinkhorn iteration.
ASAR [29]	An appearance-aware regularizer is employed to explicitly increase the dissimilarities between similar keypoints and improve model robustness through adversarial attacks.
COMMON [24]	Integrating the momentum distillation strategy to balance the quadratic contrastive loss and reduce the impact of bi-level noisy correspondence.

Table 3: Summary of main existing literature in RS-type methods for robustness certification.

Method	Introduction
RS [8]	A pioneering work on certified robustness for classification tasks, demonstrating that Gaussian smoothing distributions can provide a provable ℓ_2 perturbation bound.
DSSN [23]	Providing a novel non-additive smoothing robustness certificate for the ℓ_1 threat model.
Median Smoothing [5]	Developing a new variant of smoothing specifically for detection based on the medians of the smoothed predictions.
RS for Segmentation [12]	Presenting a scalable certification method for image and point cloud segmentation based on randomized smoothing.
RS for Community Detection [16]	Building a new smoothed community detection method via randomly perturbing the graph structure.

dataset [6] and SPair-71k dataset [27] for visual graph matching. Following the protocol of Wang et al. [36], for the Pascal VOC dataset, we exclude images with poor annotations. Then we use 100 inputs (about 650 keypoints) from 20 categories in the dataset to test the proposed method on six representative deep GM methods: GMN [45], PCA-GM [35], CIE-H [43], NGMv2 [36], ASAR [29] and COMMON [24], using the checkpoints of these GM models provided by ThinkMatch (https://github.com/Thinklab-SJTU/ThinkMatch). For the Willow ObjectClass dataset, we use 100 inputs from 5 categories to test the method on the NGMv2 solver. For the SPair-71k dataset, we use 90 inputs from 5 categories to test the method on the NGMv2 solver.

F EXPERIMENTAL RESULTS

This section first presents the certification results on the Willow ObjectClass dataset and the certification results for ASAR [29] and COMMON [24] solvers. Moreover, it reports additional results on how the parameters n, κ, β, and k affect the certified robustness and model performance in Appendix F.1. Finally, it shows the results under pixel perturbations in Appendix F.2.

F.1 Additional Experimental Results under Keypoint Position Perturbations

First, we investigate the relationship of CA and three marginal radii ($\|\delta\|_{\text{lower}}$, $\|\delta\|_{\text{upper}}$, and $\|\delta\|_{\text{volume}}$) for RS-GM and CR-OSRS on the Willow ObjectClass dataset in Fig. 5.We also evaluate the performance of RS-GM and CR-OSRS under three training conditions: without data augmentation and regularizer, with data augmentation only, and with both data augmentation and regularizer, as defined in Eq. 11. In Fig. 5, the curve of CR-OSRS is almost always above RS-GM, indicating that CR-OSRS corresponds to larger radii for the same certified accuracy and corresponds to higher accuracy for the same radii, which implies greater certified robustness. However, we observe that the improvement of model performance by data augmentation and regularizer is not as significant as on the Pascal VOC dataset. We conjecture that this is because Willow is less sensitive to perturbations for keypoint positions. Therefore, data augmentation and regularizer have little effect on the "majority decision" of RS and even cause the model to underfit.

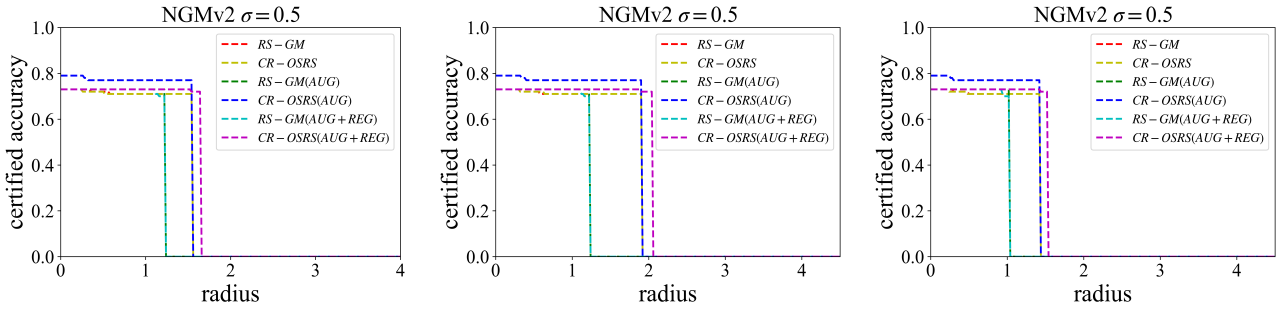
Second, we examine the relationship of CA and three marginal radii for RS-GM and CR-OSRS on ASAR [29] and COMMON [24] in Fig. 6 and Fig. 7. The curve of CR-OSRS is almost always above RS-GM, which implies greater certified robustness and matching

accuracy. At the same time, it also demonstrates that the proposed data augmentation and regularizer are effective.

Third, we further examine the effect of the number of copies n, the constraint hyperparameter κ, the regularization hyperparameter β in Eq. 11 as well as the Monte Carlo sample number k for Monte Carlo sampling on the certification results, which were not examined in Sec. 5.2. We vary n from $n \in \{1, 2, 3, 4\}$ and plot the certified accuracy with each n in Fig. 8 which indicates that choosing appropriate values of n is crucial for improving the model performance. We vary κ from $\kappa \in \{0, \frac{1}{300}, \frac{1}{200}, \frac{1}{100}\}$ and plot the certified accuracy with each κ in Fig. 9. Fig. 9 shows that κ had little overall influence on the outcomes, but a larger κ results in a larger $\|\delta\|_{\text{volume}}$ and $\|\delta\|_{\text{upper}}$ as well as a smaller $\|\delta\|_{\text{lower}}$. We vary β from $\beta \in \{0.005, 0.01, 0.02\}$ and plot the certified accuracy with each β in Fig. 10 which indicates that choosing appropriate values of β will help balance the trade-off between matching performance and certified robustness. Furthermore, we vary k from $k \in \{1000, 2000, 3000, 4000, 5000\}$ and plot the certified accuracy with each k in Fig. 11, which projects how the certified accuracy would change when using more samples k (under the assumption $k = 10k_0$). We observe that when k increases, the robustness can be certified to be stronger, which is influenced by the Monte Carlo sampling algorithm.

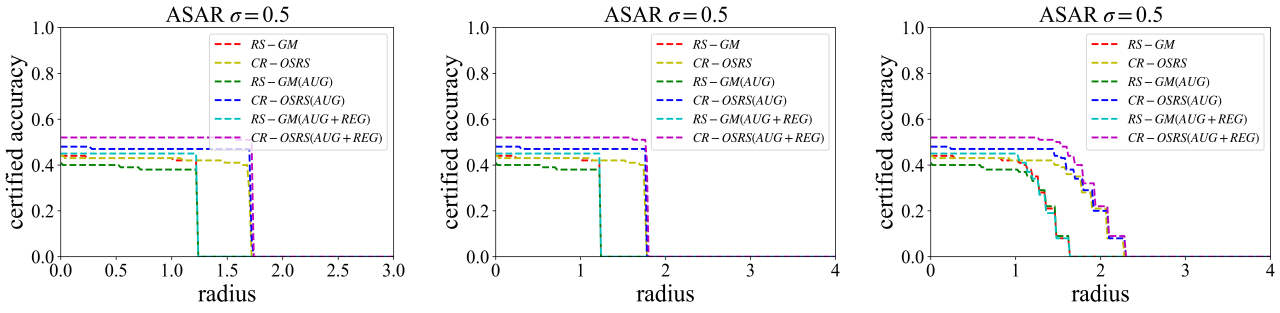
F.2 Experimental Results on Image Pixel Perturbations

For perturbing image pixels, we plot the relationship of certified accuracy (CA) and three marginal radii ($\|\delta\|_{\text{lower}}$, $\|\delta\|_{\text{upper}}$, and $\|\delta\|_{\text{volume}}$) in Fig. 12 with the original $\sigma = 0.5$, $\beta = 0.01$ and $n = 2$. As discussed in Section D.2, constructing a correlation matrix between pixels is computationally expensive due to the large number of image pixels. Moreover, it is challenging to extract the correlation between pixels. Hence, we employ RS-GM to achieve robustness certification under pixel perturbations. Fig. 12 demonstrates the effectiveness of data augmentation and regularizer. Data augmentation has a significant effect, but the regularizer does not improve performance in this case. We hypothesize that this outcome results from the low variability of the output distribution of a fixed input under multiple perturbations, which renders the regularizer insignificant.



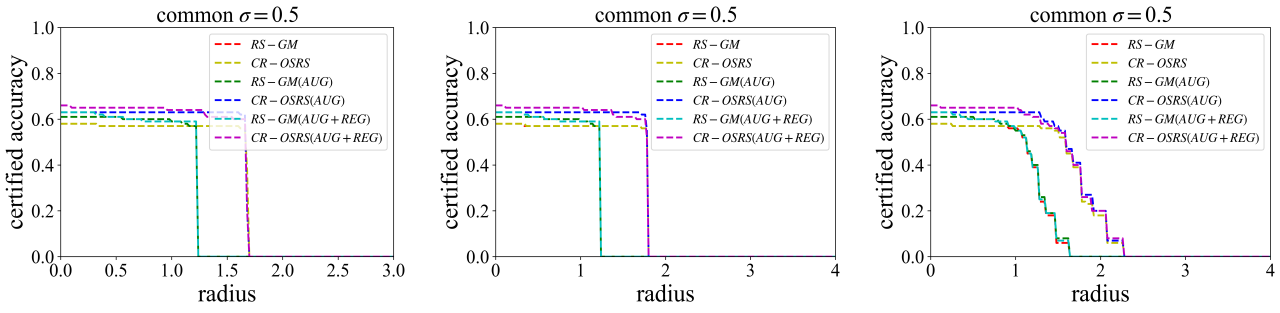
(a) CA and $\|\delta\|_{\text{lower}}$ (b) CA and $\|\delta\|_{\text{upper}}$ (c) CA and $\|\delta\|_{\text{volume}}$

Figure 5: CA achieved by RS-GM and CR-OSRS for NGMv2 on Willow ObjectClass dataset when perturbing keypoint positions. Fig. 5 shows the result with original $\sigma = 0.5$, $s = 0.9$ in Eq. 5, $\beta = 0.01$ and $n = 2$ in Eq. 11.



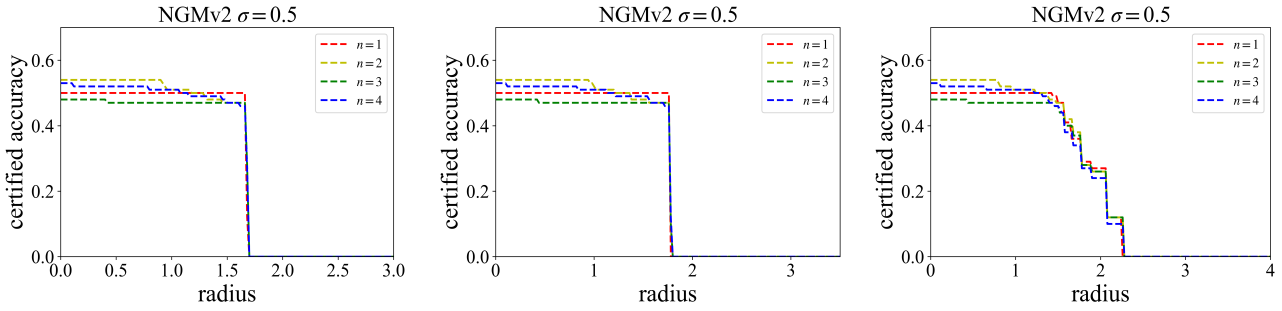
(a) CA and $\|\delta\|_{\text{lower}}$ (b) CA and $\|\delta\|_{\text{upper}}$ (c) CA and $\|\delta\|_{\text{volume}}$

Figure 6: CA achieved by CR-OSRS and RS-GM for ASAR on Pascal VOC when perturbing keypoint positions. It shows the result for original $\sigma = 0.5$, $s = 0.9$ in Eq. 5, $\beta = 0.01$ and $n = 2$ in Eq. 11.



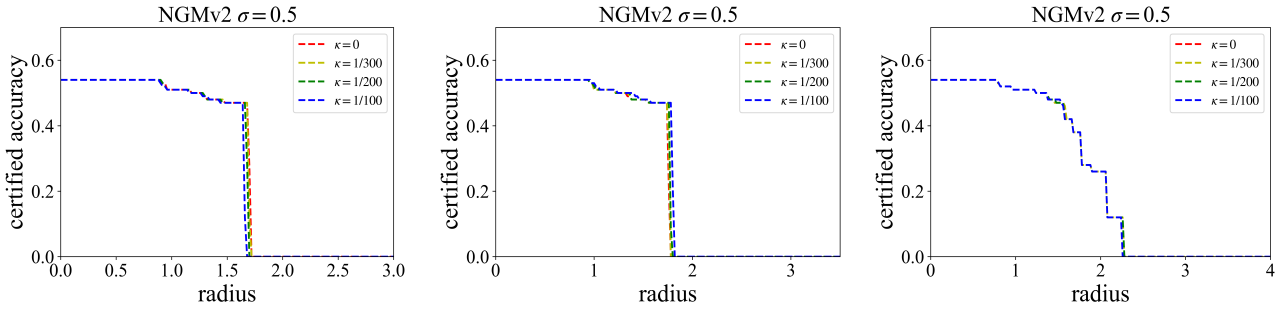
(a) CA and $\|\delta\|_{\text{lower}}$ (b) CA and $\|\delta\|_{\text{upper}}$ (c) CA and $\|\delta\|_{\text{volume}}$

Figure 7: CA achieved by CR-OSRS and RS-GM for COMMON on Pascal VOC when perturbing keypoint positions. Fig. 7 shows the result for original $\sigma = 0.5$, $s = 0.9$ in Eq. 5, $\beta = 0.01$ and $n = 2$ in Eq. 11.



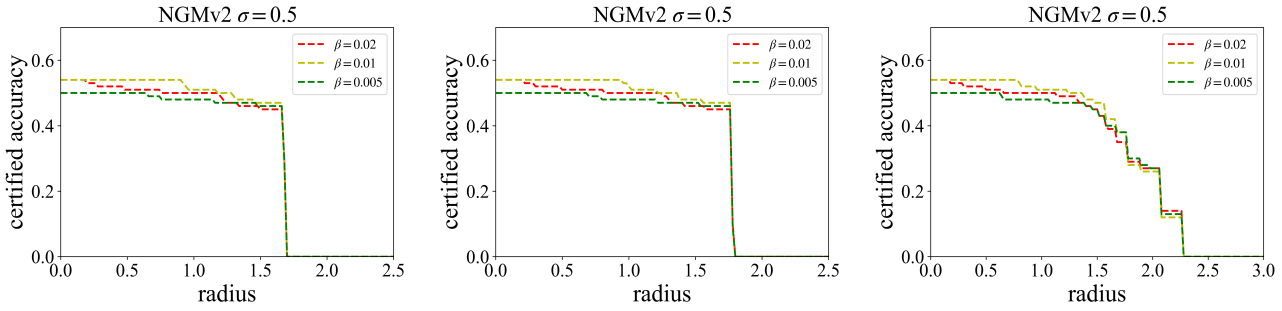
(a) CA and $\|\delta\|_{\text{lower}}$ (b) CA and $\|\delta\|_{\text{upper}}$ (c) CA and $\|\delta\|_{\text{volume}}$

Figure 8: Projections for the certified accuracy if the loss function parameter n had been larger or smaller. Fig. 8 shows the result for CR-OSRS trained by the data augmentation and regularizer with $\sigma = 0.5$, $s = 0.9$ and $\beta = 0.01$ for NGMv2 on Pascal VOC.



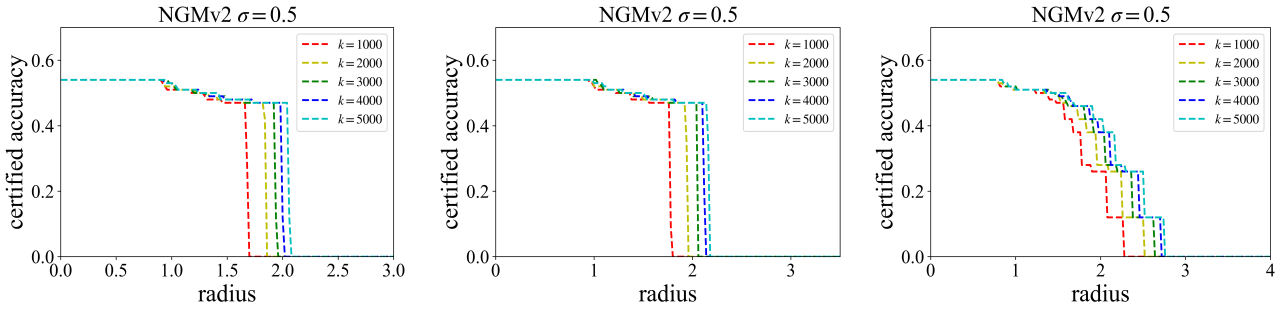
(a) CA and $\|\delta\|_{\text{lower}}$ (b) CA and $\|\delta\|_{\text{upper}}$ (c) CA and $\|\delta\|_{\text{volume}}$

Figure 9: Projections for the certified accuracy if the constraint hyperparameter κ had been larger or smaller. Fig. 9 shows the result for CR-OSRS trained by the data augmentation and regularizer with $\sigma = 0.5$, $s = 0.9$ and $\beta = 0.01$ for NGMv2 on Pascal VOC.



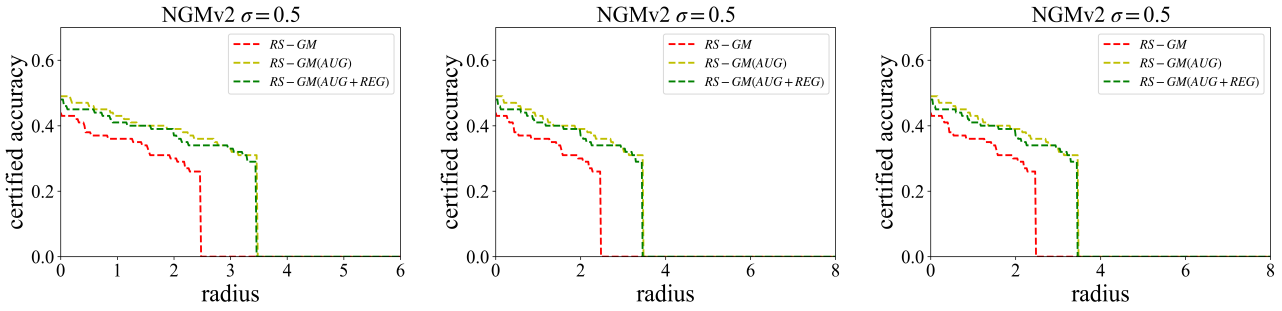
(a) CA and $\|\delta\|_{\text{lower}}$ (b) CA and $\|\delta\|_{\text{upper}}$ (c) CA and $\|\delta\|_{\text{volume}}$

Figure 10: Projections for the certified accuracy if the regularization hyperparameter β had been larger or smaller. Fig. 10 shows the result for CR-OSRS trained by the data augmentation and regularizer with $\sigma = 0.5$, $s = 0.9$ and $n = 2$ for NGMv2 on Pascal VOC.



(a) CA and $\|\delta\|_{\text{lower}}$ (b) CA and $\|\delta\|_{\text{upper}}$ (c) CA and $\|\delta\|_{\text{volume}}$

Figure 11: Projections for the certified accuracy if the Monte Carlo sample number k had been larger or smaller. Fig. 11 shows the result for CR-OSRS trained by the data augmentation and regularizer with $\sigma = 0.5$, $s = 0.9$, $\beta = 0.01$ and $n = 2$ for NGMv2 on Pascal VOC.



(a) CA and $\|\delta\|_{\text{lower}}$ (b) CA and $\|\delta\|_{\text{upper}}$ (c) CA and $\|\delta\|_{\text{volume}}$

Figure 12: CA achieved by RS-GM for NGMv2 on Pascal VOC under pixel perturbations.