Appendix of Tighter Truncated Rectangular Prism Approximation for RNN Robustness Verification

Anonymous submission

The appendix is organized as follows. Appendix A and B provides the proof of volume and surface area for truncated poly-prism, respectively. Appendix C demonstrates how different over-approximations affect verification results, and Appendix D describes the details of the experimental implementation. Appendix E, F, and G present the detailed results for RQ1, RQ2 and RQ3, respectively.

Appendix A: Proof of Volume for Truncated Poly-Prism

First, we present the definition of the truncated poly-prism.

Definition 1. *As shown in Figure 1, let $P_1P_2\cdots P_n$ be a convex polygon in the x-y plane of $\mathbb{R}^3$, where the cooridinate of $P_i \in \mathbb{R}^3$ is $(x_i, y_i, 0)$. Let $Q_i = (x_i, y_i, z_i) \in \mathbb{R}^3\,(i = 1, 2, \dots, n)$ satisfies $z_1, z_2, \dots, z_n > 0$ and $Q_1, Q_2, \dots, Q_n$ in the same plane. The geometry formed by $P_1P_2\cdots P_n; Q_1Q_2\cdots Q_n$ is referred to as "truncated poly-prism", and the polygon $P_1P_2\cdots P_n$ is referred to as the base of the truncated poly-prism.*

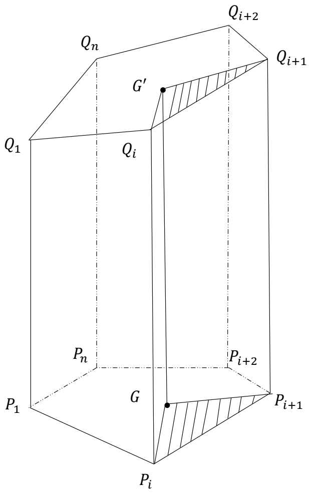


Figure 1: A truncated poly-prism.

In this appendix, we will prove the following Theorem 1.

Theorem 1. *The volume of the truncated poly-prism in Definition 1 is given by:*

$$V = \frac{1}{n}(z_1 + z_2 + \cdots + z_n) \cdot Area(P_1P_2\cdots P_n).$$

We present some lemmas of special cases, and then provide the whole proof of Theorem 1.

Lemma 1. *As shown in Figure 2, let $P_1P_2P_3 \subset \mathbb{R}^2 \times \{0\}$ be a right triangle, with coordinates given by:*

$$P_1 = (0, 0, 0), P_2 = (a, 0, 0), P_3 = (0, b, 0),$$

where $a, b > 0$. Let the cooridinates of $Q_1, Q_2, Q_3 \in \mathbb{R}^3$ be:

$$Q_1 = (0, 0, c), Q_2 = (a, 0, 0), Q_3 = (0, b, d),$$

where $c > 0, d \geq 0$. Then the volume of the truncated triangular prism $P_1P_2P_3 - Q_1Q_2Q_3$ is:

$$V = \frac{1}{3} \cdot (c + d) \cdot Area(P_1P_2P_3).$$

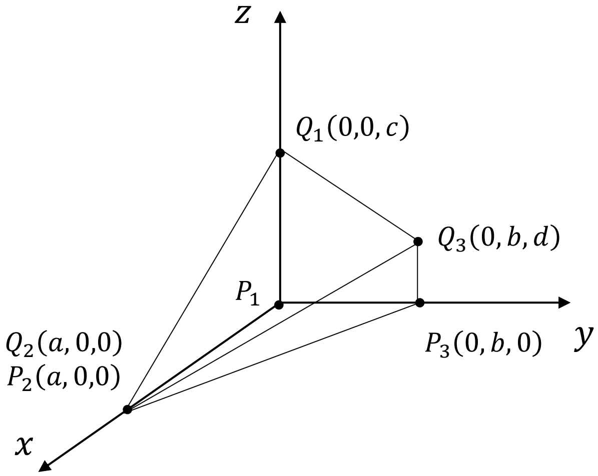


Figure 2: A truncated triangular prism, where $P_1P_2P_3$ is a right triangle.

Proof: The truncated triangular prism $P_1P_2P_3 - Q_1Q_2Q_3$ can be seen as a quadrilateral pyramid, with the base being the trapezoid $P_1P_3Q_3Q_1$ and the height being P_1P_2. Therefore, its volume is:

$$\begin{aligned} V &= \frac{1}{3}Area(P_1P_3Q_3Q_1) \cdot |P_1P_2| \\ &= \frac{1}{3}\left(\frac{1}{2}(c+d) \cdot b\right) \cdot a \\ &= \frac{1}{6}(c+d)ab \\ &= Area(P_1P_2P_3) \cdot \frac{1}{3}(c+d). \end{aligned}$$

Q.E.D. □

Lemma 1 can be extended to a more general case, where the triangle $P_1P_2P_3$ is not necessarily a right triangle. We describe this case in Lemma 2.

Lemma 2. *As shown in Figure 3, let $P_1P_2P_3 \subset \mathbb{R}^2 \times \{0\}$ be a general triangle, with coordinates given by:*

$$P_1 = (0,0,0), P_2 = (a,0,0), P_3 = (x,y,0),$$

where $a, x, y > 0$. Let the coordinates of $Q_1, Q_2, Q_3 \in \mathbb{R}^3$ be:

$$Q_1 = (0,0,c), Q_2 = (a,0,0), Q_3 = (x,y,d),$$

where $c > 0, d \geq 0$. Then, the volume of $P_1P_2P_3 - Q_1Q_2Q_3$ is:

$$V = \frac{1}{3} \cdot (c+d) \cdot Area(P_1P_2P_3).$$

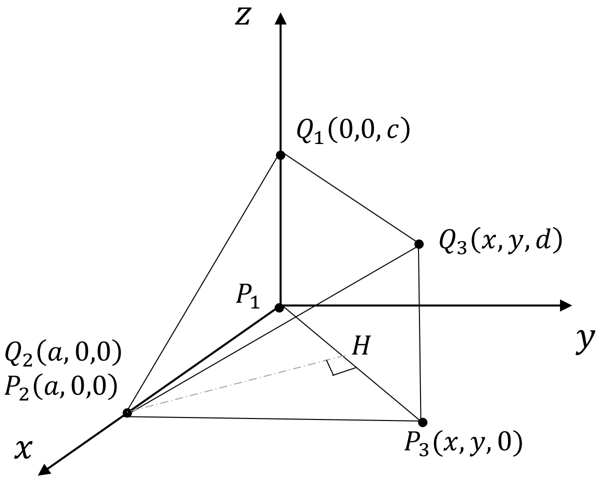


Figure 3: A truncated triangular prism, where $P_1P_2P_3$ is not a right triangle.

Proof: Draw a perpendicular from P_2 to P_1P_3, and let the foot of the perpendicular be H. Then the truncated triangular prism can be seen as a quadrilateral pyramid with a trapezoidal base $P_1P_3Q_3Q_1$ and a height P_2H. Therefore, its volume is:

$$\begin{aligned} V &= \frac{1}{3}Area(P_1P_3Q_3Q_1) \cdot P_2H \\ &= \frac{1}{3} \cdot \left(\frac{1}{2}(c+d) \cdot P_1P_3\right) \cdot P_2H \\ &= \frac{1}{3}(c+d) \cdot \frac{1}{2}P_1P_3 \cdot P_2H \\ &= \frac{1}{3}(c+d) \cdot Area(P_1P_2P_3). \end{aligned}$$

Q.E.D. □

We now prove a more general case as follows.

Lemma 3. *As shown in Figure 4, let $P_1P_2P_3 \subset \mathbb{R}^2 \times \{0\}$ be a general triangle, with coordinates given by:*

$$P_1 = (0,0,0), P_2 = (a,0,0), P_3 = (x,y,0),$$

where $a, x, y > 0$. Let the coordinates of $Q_1, Q_2, Q_3 \in \mathbb{R}^3$ be:

$$Q_1 = (0,0,c), Q_2 = (a,0,e), Q_3 = (x,y,d),$$

where $c > 0, d, e \geq 0$. Then the volume of $P_1P_2P_3 - Q_1Q_2Q_3$ is:

$$V = \frac{1}{3} \cdot (c+d+e) \cdot Area(P_1P_2P_3).$$

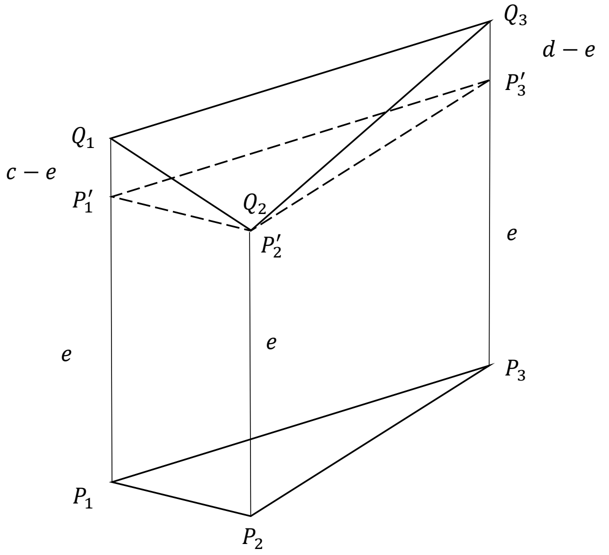


Figure 4: A truncated triangular prism.

Proof: Without loss of generality, assume $e \leq d$. By slicing the truncated triangular prism $P_1P_2P_3 - Q_1Q_2Q_3$ with the plane $z = e$, we obtain two geometric bodies: the triangular prism $P_1P_2P_3 - P_1'P_2'P_3'$ below the plane $z = e$ and the truncated triangular prism $P_1'P_2'P_3' - Q_1Q_2Q_3$ above the plane $z = e$, with the cross-section being the triangle $P_1'P_2'P_3'$. The volumes of these two geometric bodies are:

$$
\begin{aligned}
V_1 &= Area(P_1P_2P_3)\cdot e,\\
V_2 &= \frac{1}{3}\cdot Area(P_1'P_2'P_3')\cdot((c-e)+0+(d-e))\\
&= \frac{1}{3}\cdot Area(P_1P_2P_3)\cdot(d+c-2e).
\end{aligned}
$$

So, the volume of $P_1P_2P_3 - Q_1Q_2Q_3$ is:

$$
\begin{aligned}
V &= V_1 + V_2\\
&= Area(P_1P_2P_3)\cdot e + \frac{1}{3}Area(P_1P_2P_3)\cdot(d+c-2e)\\
&= \frac{1}{3}Area(P_1P_2P_3)\cdot(d+e+c).
\end{aligned}
$$

Q.E.D. □

Finally, we present the whole proof of Theorem 1.

Proof (Theorem 1): As shown in Figure 1, let G be the centroid of the convex polygon $P_1P_2\cdots P_n$, and G' be the centroid of $Q_1Q_2\cdots Q_n$. Let $z_0 = \frac{1}{n}\sum_{i=1}^{n} z_i$. The coordinates of G and G' are:

$$
\begin{aligned}
G &= \left(\frac{1}{n}\sum_{i=1}^{n} x_i, \frac{1}{n}\sum_{i=1}^{n} y_i, 0\right),\\
G' &= \left(\frac{1}{n}\sum_{i=1}^{n} x_i, \frac{1}{n}\sum_{i=1}^{n} y_i, z_0\right).
\end{aligned}
$$

The truncated poly-prism $P_1P_2\cdots P_n - Q_1Q_2\cdots Q_n$ can be partitioned into the union of the following n pairwise disjoint truncated triangular prisms:

$$P_iP_{i+1}G - Q_iQ_{i+1}G', \quad i = 1, 2, \ldots, n.$$

where $P_{n+1} = P_1$ and $Q_{n+1} = Q_1$. The volumes of these n truncated triangular prisms are:

$$V_i = \frac{1}{3}Area(P_iP_{i+1}G)\cdot(z_i + z_{i+1} + z_0),\ i = 1, 2, \ldots, n.$$

The volume of the polyhedron $P_1P_2\cdots P_n - Q_1Q_2\cdots Q_n$ is the sum of the volumes of the n truncated triangular prisms, that is:

$$
\begin{aligned}
V &= V_1 + V_2 + \cdots + V_n\\
&= \frac{1}{3}\sum_{i=1}^{n}\left(Area(P_iP_{i+1}G)\cdot(z_i + z_{i+1} + z_0)\right).
\end{aligned}
$$

Since G is the centroid of the convex polygon $P_1P_2\cdots P_n$, it follows that:

$$\text{Area}(P_iP_{i+1}G) = \frac{1}{n}\cdot\text{Area}(P_1P_2\cdots P_n),$$

where $i = 1, 2, \ldots, n$. Therefore,

$$V = \frac{1}{3}\cdot\frac{1}{n}Area(P_1P_2\cdots P_n)\cdot\sum_{i=1}^{n}(z_i + z_{i+1} + z_0).$$

It is easy to obtain that:

$$\sum_{i=1} n\,(z_i + z_{i+1} + z_0) = 2\sum_{i=1}^{n} z_i + n\cdot z_0 = 3\sum_{i=1}^{n} z_i.$$

Thus, we finally obtain the volume of the truncated poly-prism:

$$
\begin{aligned}
V &= \frac{1}{n}\cdot Area(P_1P_2\cdots P_n)\cdot\sum_{i=1}^{n} z_i\\
&= \frac{1}{n}(z_1 + z_2 + \cdots + z_n)\cdot Area(P_1P_2\cdots P_n).
\end{aligned}
$$

Q.E.D. □

Note that in Theorem 1, when $P_1P_2\cdots P_n$ is a rectangle, it represents the special case needed in the paper.

Appendix B: Proof of Surface Area for Truncated Rectangle Prism

Given a truncated rectangle prism $P_1P_2P_3P_4 - Q_1Q_2Q_3Q_4$, shown in Figure 5, let z_0 be the centroid and z_1, z_2, z_3, z_4 be the four lateral edges of the prism, where $z_1 \geq z_2, z_3$ and $z4 \leq z_2, z_3$. Assume the surface area of the prism is S. We will prove Theorem 2.

Theorem 2. *S is positively correlated with $|z_1 - z_0| + |z_2 - z_0| + |z_3 - z_0| + |z_4 - z_0|$.*

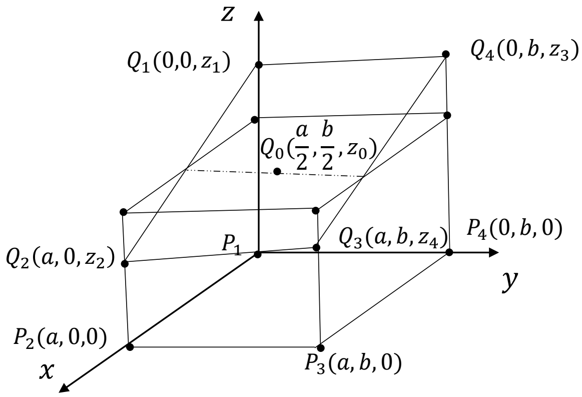


Figure 5: A truncated rectangle prism.

We first present some lemmas and then provide the whole proof of Theorem 2.

Lemma 4. *As shown in Figure 5, the four lateral edges satisfy*

$$z_1 + z_4 = z_2 + z_3.$$

Proof: Let $Q_1Q_2Q_3Q_4 \subset \mathbb{R}^3$ be the top face of the truncated rectangle prism. The coordinates of this four vertices are:

$$
\begin{aligned}
Q_1 = (0, 0, z_1), Q_2 = (a, 0, z_2),\\
Q_3 = (a, b, z_4), Q_4 = (0, b, z_3),
\end{aligned}
$$

where $a, b > 0$. Since Q_1, Q_2, Q_3, Q_4 are coplanar, they satisfy the following determinant:

$$\begin{vmatrix} 0 & 0 & z_1 & 1 \\ a & 0 & z_2 & 1 \\ a & b & z_4 & 1 \\ 0 & b & z_3 & 1 \end{vmatrix} = 0,$$

which can be simplified with

$$z_1 \begin{vmatrix} a & 0 & 1 \\ a & b & 1 \\ 0 & b & 1 \end{vmatrix} - z_2 \begin{vmatrix} 0 & 0 & 1 \\ a & b & 1 \\ 0 & b & 1 \end{vmatrix} + z_4 \begin{vmatrix} 0 & 0 & 1 \\ a & 0 & 1 \\ 0 & b & 1 \end{vmatrix} - z_3 \begin{vmatrix} 0 & 0 & 1 \\ a & 0 & 1 \\ a & b & 1 \end{vmatrix} = 0,$$

$$z_1 \cdot (a(b-b) + (ab-1)) - z_2 \cdot ab + z_4 \cdot ab - z_3 \cdot ab = 0,$$

$$z_1 - z_2 + z_4 - z_3 = 0,$$

i.e.

$$z_1 + z_4 = z_2 + z_3.$$

Q.E.D. □

Next, we will prove Lemma 5.

Lemma 5. *Let the line connecting the centroids of the top and bottom faces be z_0, and the four lateral edges satisfy:*

$$z_1 + z_2 + z_3 + z_4 = 4 \cdot z_0$$

Proof: Let G be the centroid of the quadrilateral $P_1P_2P_3P_4$, and G' be the centroid of $Q_1Q_2Q_3Q_4$. The coordinates of G and G' are:

$$G = \left(\frac{1}{4} \sum_{i=1}^{4} x_i, \frac{1}{4} \sum_{i=1}^{4} y_i, 0 \right),$$

$$G' = \left(\frac{1}{4} \sum_{i=1}^{4} x_i, \frac{1}{4} \sum_{i=1}^{4} y_i, \frac{1}{4} \sum_{i=1}^{4} z_i \right),$$

and the length of GG' is:

$$z_0 = \frac{1}{4} \sum_{i=1}^{4} z_i,$$

i.e.

$$z_1 + z_2 + z_3 + z_4 = 4 \cdot z_0.$$

Q.E.D. □

Finally, we present the whole proof of Theorem 2.

Proof (Theorem 2): The surface area of the truncated rectangular prism is the sum of the areas of six faces, denoted as $S = \sum_{i=1}^{6} S_i$. Among these, the four lateral faces are trapezoidal, and their areas can be calculated as:

$$\begin{aligned} S_1 + S_2 + S_3 + S_4 &= \frac{1}{2} \cdot (z_1 + z_2) \cdot a + \frac{1}{2} \cdot (z_2 + z_3) \cdot b \\ &+ \frac{1}{2} \cdot (z_3 + z_4) \cdot a + \frac{1}{2} \cdot (z_4 + z_1) \cdot b \\ &= \frac{1}{2}(z_1 + z_2 + z_3 + z_4) \cdot a + \frac{1}{2}(z_1 + z_2 + z_3 + z_4) \cdot b \\ &= \frac{1}{2}(z_1 + z_2 + z_3 + z_4) \cdot (a + b) \\ &= 2 \cdot z_0 \cdot (a + b). \end{aligned}$$

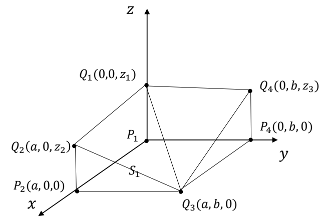


Figure 6: A truncated rectangle prism with the degeneration of $z_4 = 0$.

Therefore, the four lateral faces can be seen as constants, and we only focus on the areas of the top and bottom faces. Next, let the smallest $z_4 = 0$, and calculate the area of the top quadrilateral, as shown in Figure 6.

The coordinates of the top face $Q_1Q_2Q_3Q_4 \subset \mathbb{R}^3$ are given by:

$$Q_1 = (0, 0, z_1), Q_2 = (a, 0, z_2),$$
$$Q_3 = (a, b, 0), Q_4 = (0, b, z_3),$$

where $a, b > 0$. According to Lemma 4, we have $z_1 = z_2 + z_3 = 2z_0$.

We can assume $z_2 \geq z_3$ without loss of generality. So,

$$\begin{aligned} &|z_1 - z_0| + |z_2 - z_0| + |z_3 - z_0| + |0 - z_0| \\ =&|z_1 - \frac{z_1}{2}| + |z_2 - \frac{z_1}{2}| + |z_3 - \frac{z_1}{2}| + |0 - \frac{z_1}{2}| \\ =&\frac{z_1}{2} + z_2 - \frac{z_1}{2} + \frac{z_1}{2} - z_3 + \frac{z_1}{2} \\ =&z_1 - z_3 + z_2 \\ =&2z_2. \end{aligned}$$

Let a, b, c be three sides of a triganle, and Δ be the area of the triangle. It is well known that a highly symmetrical form of Heron's formula (Buchholz 1992) is:

$$(4\Delta)^2 = \begin{bmatrix} a^2 & b^2 & c^2 \end{bmatrix} \begin{bmatrix} -1 & 1 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \end{bmatrix} \begin{bmatrix} a^2 \\ b^2 \\ c^2 \end{bmatrix}.$$

The simplified expression is:

$$16\Delta^2 = (a^2 + b^2 + c^2)^2 - 2(a^4 + b^4 + c^4),$$

and for the triangle $Q_1Q_2Q_3$:

$$\begin{aligned} 16\Delta^2 =&((z_1 - z_2)^2 + a^2 + b^2 + z_2^2 + z_1^2 + a^2 + b^2)^2 \\ &- 2(((z_1 - z_2)^2 + a^2)^2 + (b^2 + z_2^2)^2 \\ &+ (z_1^2 + a^2 + b^2)^2) \\ =&((z_1 - z_2)^2 + a^2 + b^2 + z_2^2 + z_1^2 + a^2 + b^2)^2 \\ &- 2((z_1 - z_2)^2 + 2a^2(z_1 - z_2)^2 + a^4) \\ &- 2(z_2^4 + 2b^2z_2^2 + b^4) \\ &- 2(z_1^4 + 2(a^2 + b^2)z_1^2 + (a^2 + b^2)^2). \end{aligned}$$

The constant term is:

$$\begin{aligned} &4(a^2 + b^2)^2 - 2(a^4 + b^4 + (a^2 + b^2)^2) \\ =&2(a^2 + b^2)^2 - 2(a^4 + b^4) \\ =&4a^2b^2. \end{aligned}$$

The second-order term is:

$$\begin{aligned}&4(a^2+b^2)(z_1^2+z_2^2+(z_1-z_2)^2)\\ &\quad-4a^2(z_1-z_2)^2-4b^2z_2^2-4(a^2+b^2)z_1^2\\ =&4a^2z_2^2+4b^2(z_1-z_2)^2.\end{aligned}$$

The four-order term is:

$$\begin{aligned}&z_1^4+z_2^4+(z_1-z_2)^4\\ &+2z_1^2z_2^2+2z_1^2(z_1-z_2)^2+2z_2^2(z_1-z_2)^2\\ &-2z_1^4-2z_2^4-2(z_1-z_2)^4\\ =&2z_1^2z_2^2+2(z_1^2+z_2^2)(z_1-z_2)^2\\ &-z_1^4-z_2^4-(z_1-z_2)^4\\ =&-(z_1^2-z_2^2)^2+(z_1-z_2)^2(2z_1^2+2z_2^2-(z_1-z_2)^2)\\ =&-(z_1-z_2)^2(z_1+z_2)^2+(z_1-z_2)^2(2z_1^2+2z_2^2-(z_1-z_2)^2)\\ =&(z_1-z_2)^2(-(z_1+z_2)^2+2z_1^2+2z_2^2-(z_1-z_2)^2)\\ =&0.\end{aligned}$$

So,

$$\Delta=\frac{1}{4}\sqrt{4a^2b^2+4a^2z_2^2+4b^2(z_1-z_2)^2}.$$

The area of the triangle $Q_1Q_3Q_4$ can be calculated in the same way. Since $z_1=z_2+z_3$, we have:

$$\begin{aligned}S_{Q_1Q_2Q_3Q_4}=&S_{\triangle Q_1Q_2Q_3}+S_{\triangle Q_1Q_3Q_4}\\ =&\frac{1}{4}\sqrt{4a^2b^2+4a^2z_2^2+4b^2(z_1-z_2)^2}\\ &+\frac{1}{4}\sqrt{4a^2b^2+4a^2z_3^2+4b^2(z_1-z_3)^2}\\ =&\frac{1}{4}\sqrt{4a^2b^2+4a^2z_2^2+4b^2z_3^2}\\ &+\frac{1}{4}\sqrt{4a^2b^2+4a^2z_3^2+4b^2z_2^2}\\ =&\frac{1}{2}(\sqrt{a^2b^2+a^2z_2^2+b^2z_3^2}\\ &+\sqrt{a^2b^2+a^2z_3^2+b^2z_2^2}),\end{aligned}$$

There is an inequality:

$$\begin{aligned}(x+y+z)^2=&x^2+y^2+z^2+2xy+2yz+2xz\\ \le&x^2+y^2+z^2+x^2+y^2+y^2+z^2+x^2+z^2\\ =&3(x^2+y^2+z^2),\end{aligned}$$

i.e.

$$\sqrt{x^2+y^2+z^2}\ge\frac{1}{\sqrt{3}}(x+y+z).$$

Therefore, we have:

$$\begin{aligned}&S_{Q_1Q_2Q_3Q_4}\\ =&\frac{1}{2}(\sqrt{a^2b^2+a^2z_2^2+b^2z_3^2}+\sqrt{a^2b^2+a^2z_3^2+b^2z_2^2})\\ \ge&\frac{1}{2}(\sqrt{a^2b^2+a^2z_2^2}+\sqrt{a^2b^2+b^2z_2^2})\\ \ge&\frac{1}{2\sqrt{3}}(ab+az_2+ab+bz_2)\\ \ge&\frac{1}{\sqrt{3}}ab+\frac{a+b}{2\sqrt{3}}z_2.\end{aligned}$$

Given that $z_2\ge z_3$, we have

$$\begin{aligned}&S_{Q_1Q_2Q_3Q_4}\\ =&\frac{1}{2}(\sqrt{a^2b^2+a^2z_2^2+b^2z_3^2}+\sqrt{a^2b^2+a^2z_3^2+b^2z_2^2})\\ \le&\frac{1}{2}(\sqrt{a^2b^2+a^2z_2^2+b^2z_2^2}+\sqrt{a^2b^2+a^2z_2^2+b^2z_2^2})\\ \le&\sqrt{a^2b^2+(a^2+b^2)z_2^2})\\ \le&ab+\sqrt{a^2+b^2}z_2.\end{aligned}$$

Since the process of calculating the top area is the same as that for the bottom area, the surface area S satisfies the expression:

$$kz_1+b\le S\le k'z_1+b'.$$

So, S is positively correlated with z_2 as well as $|z_1-z_0|+|z_2-z_0|+|z_3-z_0|+|z_4-z_0|$.

Q.E.D. □

Appendix C: An Illustration of Approximation

A more detailed explanation of the volume-area-based method in tighter over-approximation is provided here. As shown in Figure 7, the truncated rectangular prisms formed by the red and yellow planes have the same volume, since they share the same centroid line. However, their surface areas differ, with the red one having a smaller surface area.

Different over-approximations lead to different abstract domains, which ultimately affect the output intervals, as shown in Figure 8. The red region is smaller than the yellow one, indicating a tighter approximation. Moreover, the lower bound of class 1 is higher than the upper bounds of all other classes, making verification successful. In contrast, the yellow region does not satisfy this condition, leading to verification failure.

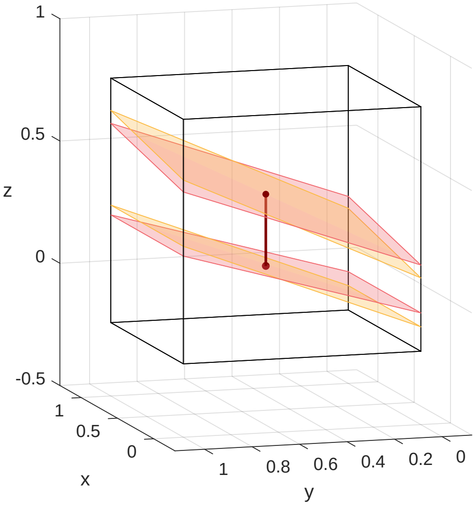


Figure 7: The two truncated rectangular prisms formed by the red and yellow planes have the same volumes but different surface areas. The red prism has a smaller surface area.

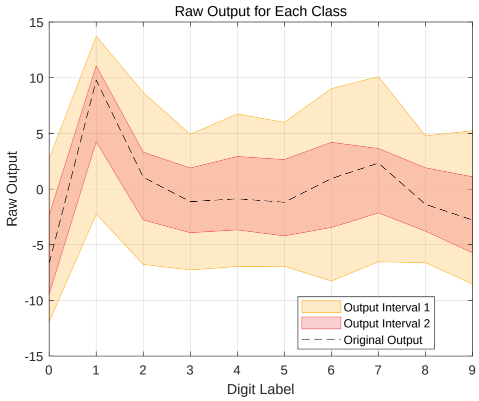


Figure 8: Different output intervals caused by two relaxations. The red prism has a smaller surface area, so the red interval is tighter than the yellow one and closer to the original output.

Appendix D: Experimental Implementation

Evaluation details of the image classification task. In the experiments, we randomly select 100 samples from the test set, check whether these samples are correctly classified, and skip the misclassified ones. Then, we apply perturbations to the originally correctly classified sample and verify whether it remains correctly classified, i.e., the classification is robust. When the verification of a sample exceeds 120 seconds, we consider it a failure. The number of robust samples out of 100 is the verification accuracy, which serves as the evaluation metric in our experiments.

Evaluation details of the speech recognition task. For speech recognition, the original signal is split into several frames, followed by three preprocessing steps, including pre-emphasizing and windowing, the power spectrum of Fast Fourier transform (FFT), and the Mel-filter bank log energy. The nonlinear functions (e.g., log and square) used during preprocessing are considered in the model verification. The magnitude of signal perturbation is measured in decibels (dB), where a smaller dB value indicates a weaker perturbation.

Evaluation details of the sentiment analysis task. We use the GloVe (Pennington, Socher, and Manning 2014) model to map the words into embeddings. GloVe encodes word meanings based on global co-occurrence statistics, positioning semantically similar words close together in the embedding space. Therefore, L_∞-norm perturbations still make sense for the text classification task.

Appendix E: Detailed Results of RQ1

The performance comparison of *RNN-Guard*, *Prover*, and *DeepPrism* under different model parameters is shown in Figure 9. Horizontally, *DeepPrism* outperforms other baselines on accuracy with a slight but acceptable increase in computation time. Vertically, an increase in f and ℓ leads to a decline in the accuracy of the verifier, while an increase in h causes an improvement.

Appendix F: Detailed Results of RQ2

The comparison between the single-plane and multi-plane verification methods across four datasets can be found in Figure 10 (MNIST, image classification), Figure 11 (GSC, speech recognition), Figure 12 (FSDD, speech recognition) and Figure 13 (RT, sentiment analysis), respectively. The results of multi-plane approximation are better than those of single-plane approximation in all models, and the running time increases as the accuracy decreases because verification failures require iterating all epochs.

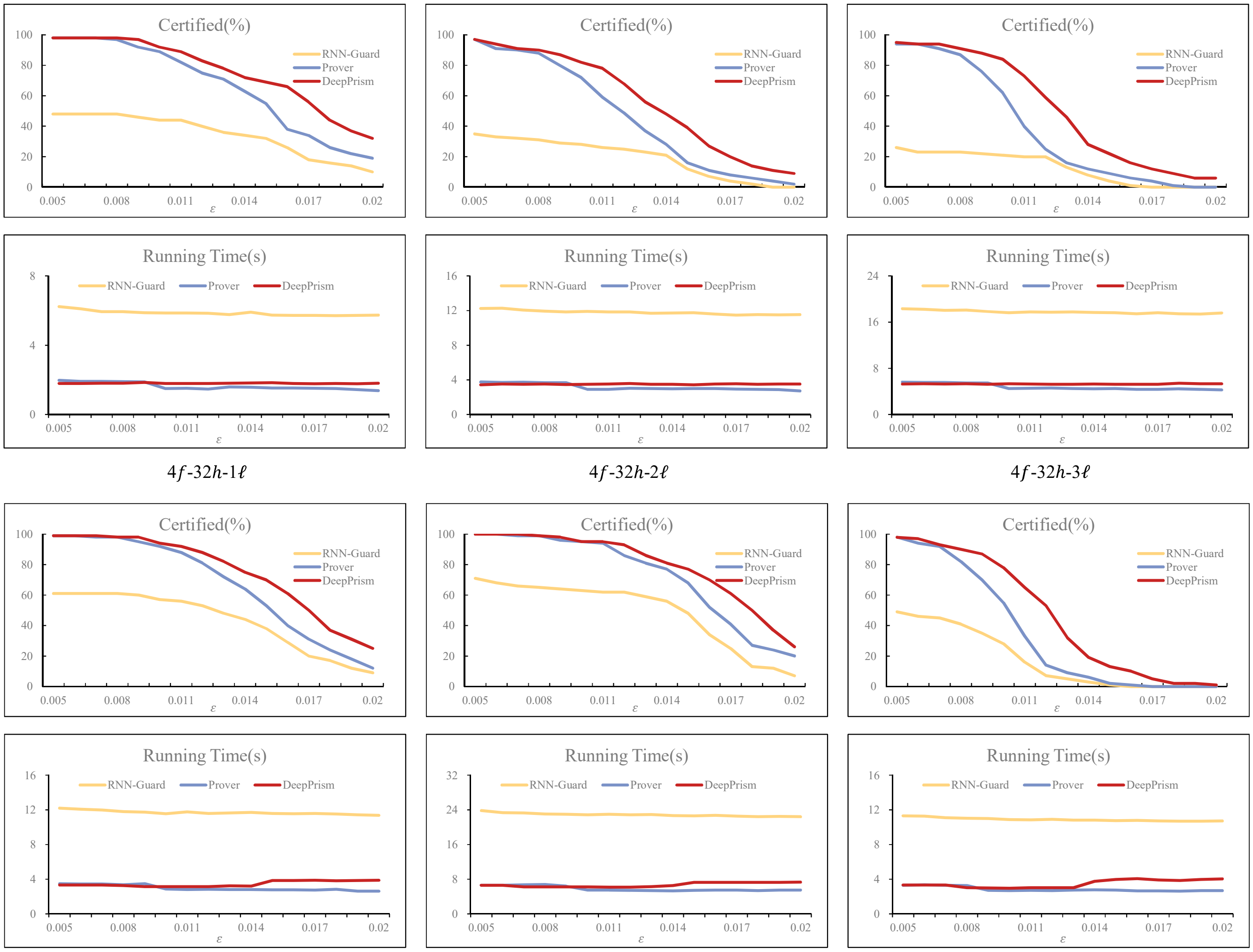
Appendix G: Detailed Results of RQ3

The comparison between different divisons across four datasets can be found in Table 1–6 (MNIST, image classification), Table 7 (GSC, speech recognition), Table 8 (FSDD, speech recognition), Table 9 (RT, sentiment analysis), respectively. Under low perturbations, all models perform similarly under all divisions, where division ③ 4-tri and ⑥ 4-rec slightly outperform others. Under high perturbations, division ⑥ 4-rec clearly outperforms others, maintaining higher accuracy. In the cases of ⑦ 9-rec and ⑧ 16-rec divisions, the verification accuracy has a significant improvement compared to that of ⑥ 4-rec division.

References

Buchholz, R. H. 1992. Perfect pyramids. *Bulletin of the Australian Mathematical Society*, 45(3): 353–368.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.



$4f$-32h-1ℓ

$4f$-32h-2ℓ

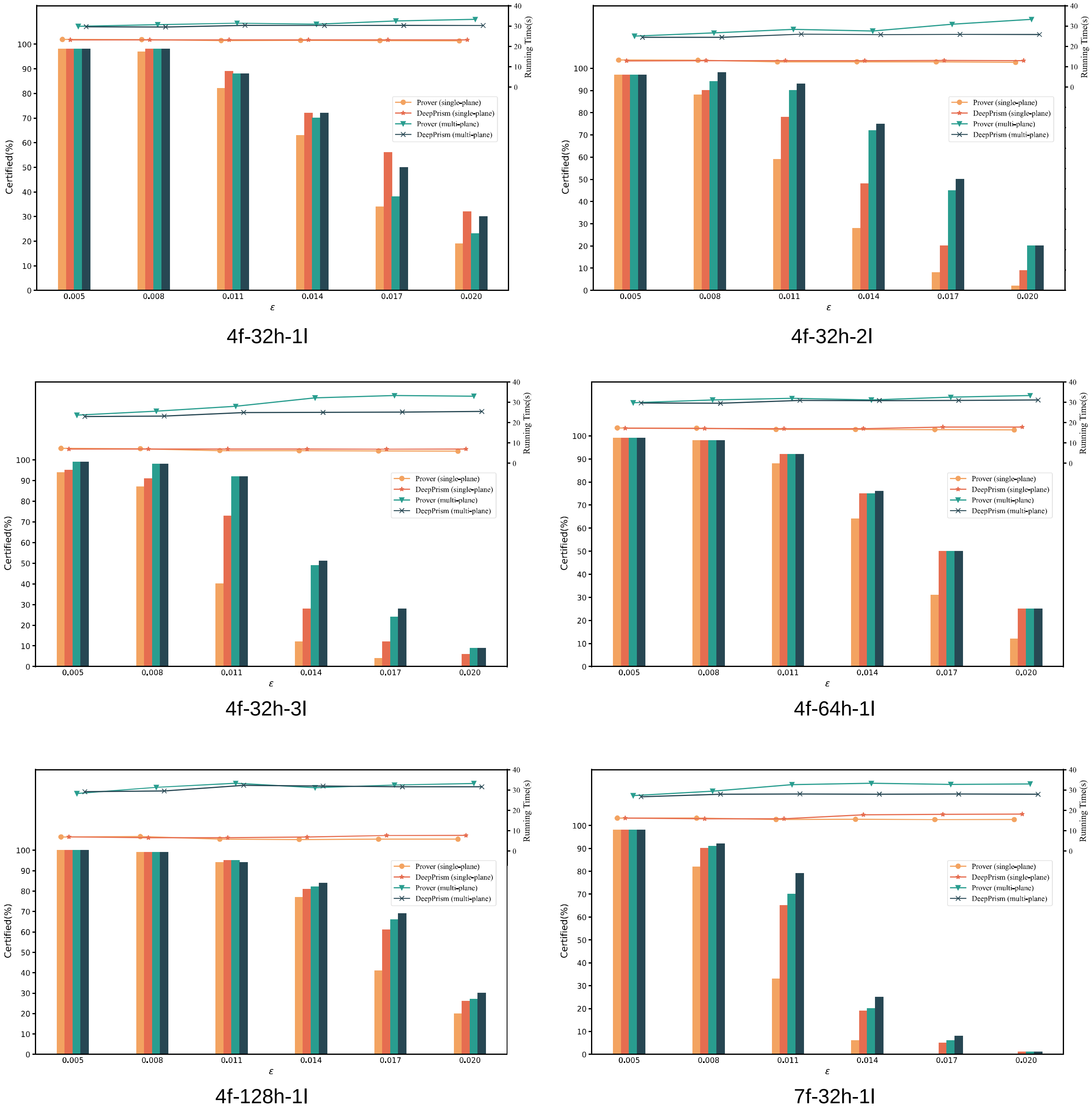
$4f$-32h-3ℓ

$4f$-64h-1ℓ

$4f$-128h-1ℓ

$7f$-32h-1ℓ

Figure 9: Results on MNIST with different perturbations and models. *DeepPrism*(red solid line) the highest certified accuracy and short running time in all models.



4f-32h-1l

4f-32h-2l

4f-32h-3l

4f-64h-1l

4f-128h-1l

7f-32h-1l

Figure 10: Comparison of two verification methods (single-plane, multi-plane) on two models (*Prover* and *DeepPrism*) evaluated on MNIST. Certified accuracy (bar) and running time (line) are shown in the same plot. Experimental results show that the multi-plane method significantly outperforms the single-plane method. Under the multi-plane setting, *DeepPrism* surpasses *Prover*, with the performance gap widening under larger perturbations.

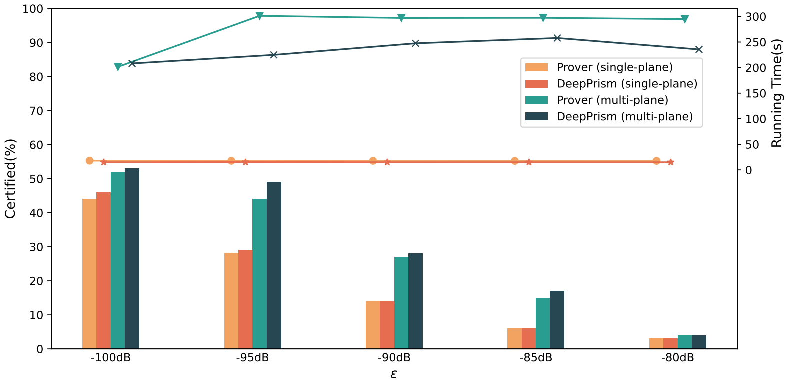


Figure 11: Comparison of two verification methods (single-plane, multi-plane) on two models (*Prover* and *DeepPrism*) evaluated on GSC. Certified accuracy (bar) and running time (line) are shown in the same plot. *DeepPrism* with the multi-plane method outperforms other approaches in certified accuracy.

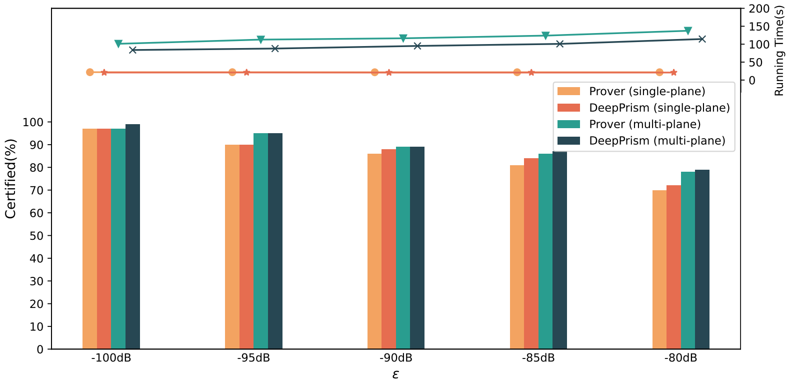


Figure 12: Comparison of two verification methods (single-plane, multi-plane) on two models (*Prover* and *DeepPrism*) evaluated on FDSS. Certified accuracy (bar) and running time (line) are shown in the same plot. *DeepPrism* with the multi-plane method outperforms other approaches in certified accuracy.

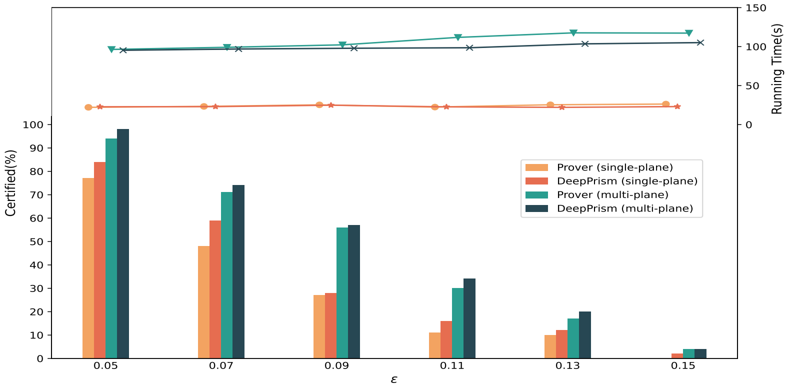


Figure 13: Comparison of two verification methods (single-plane, multi-plane) on two models (*Prover* and *DeepPrism*) evaluated on RT. Certified accuracy (bar) and running time (line) are shown in the same plot. *DeepPrism* with the multi-plane method outperforms other approaches in certified accuracy.

	① 2-tri-up		② 2-tri-down		③ 4-tri		④ 2-rec-vec		⑤ 2-rec-hor		⑥ 4-rec		⑦ 9-rec		⑧ 16-rec	
ϵ	Acc.	Time	Acc.	Time	Acc.	Time	Acc.	Time	Acc.	Time	Acc.	Time	Acc.	Time	Acc.	Time
0.005	**98**	3.82	**98**	16.72	**98**	7.28	**98**	5.20	**98**	5.60	**98**	9.21	**98**	17.71	**98**	23.94
0.008	**98**	3.94	**98**	16.71	**98**	7.95	**98**	5.43	**98**	5.34	**98**	9.21	**98**	20.39	**98**	24.38
0.011	87	4.83	88	19.18	88	8.59	85	6.03	87	5.84	88	9.40	**89**	20.90	**89**	24.63
0.014	70	5.42	71	16.07	70	8.19	68	8.20	71	7.42	70	9.67	**72**	21.77	**72**	24.85
0.017	37	6.00	37	16.10	**38**	9.46	37	8.85	37	8.02	**38**	11.62	**38**	23.38	**38**	24.98
0.020	21	6.94	21	10.06	**23**	10.11	21	8.46	21	8.37	**23**	10.19	**23**	22.41	**23**	25.12

Table 1: Verification accuracy of different divisons on MNIST under different perturbations where $f = 4, h = 32$ and $\ell = 1$.

	① 2-tri-up		② 2-tri-down		③ 4-tri		④ 2-rec-vec		⑤ 2-rec-hor		⑥ 4-rec		⑦ 9-rec		⑧ 16-rec	
ϵ	Acc.	Time	Acc.	Time	Acc.	Time	Acc.	Time	Acc.	Time	Acc.	Time	Acc.	Time	Acc.	Time
0.005	**97**	7.79	**97**	7.80	**97**	14.58	**97**	10.81	**97**	8.42	**97**	18.69	**97**	28.18	**97**	38.02
0.008	94	8.82	94	9.14	94	15.97	93	11.17	94	10.93	95	18.98	**97**	28.40	**97**	37.85
0.011	88	9.84	87	9.55	90	17.56	81	12.42	95	11.55	**97**	19.00	**97**	29.89	**97**	37.85
0.014	68	12.13	72	13.68	72	16.87	54	19.62	86	17.26	93	19.74	95	33.09	**96**	38.69
0.017	44	13.08	39	16.88	45	19.86	23	19.17	64	18.98	65	21.42	71	33.15	**76**	37.64
0.020	18	14.47	18	21.91	20	22.07	7	17.06	28	19.53	54	21.72	57	31.68	**64**	38.39

Table 2: Verification accuracy of different divisons on MNIST under different perturbations where $f = 4, h = 32$ and $\ell = 2$.

	① 2-tri-up		② 2-tri-down		③ 4-tri		④ 2-rec-vec		⑤ 2-rec-hor		⑥ 4-rec		⑦ 9-rec		⑧ 16-rec	
ϵ	Acc.	Time	Acc.	Time	Acc.	Time	Acc.	Time	Acc.	Time	Acc.	Time	Acc.	Time	Acc.	Time
0.005	**99**	11.70	**99**	50.08	**99**	21.70	**99**	15.24	**99**	14.76	**99**	28.18	**99**	60.13	**99**	71.71
0.008	97	13.55	98	51.99	98	23.65	97	16.34	**99**	15.71	**99**	28.40	**99**	56.24	**99**	72.26
0.011	83	16.37	82	68.98	92	25.95	63	21.59	89	19.27	**95**	29.89	**95**	64.10	**95**	71.91
0.014	34	19.82	33	23.66	49	30.10	25	27.78	46	31.83	71	33.09	80	72.79	**83**	72.46
0.017	16	19.26	14	24.22	24	31.15	7	26.63	21	28.10	33	33.15	35	65.34	**46**	71.83
0.020	1	21.44	1	31.40	9	30.90	0	24.06	5	26.09	**15**	31.68	**15**	73.36	**15**	72.13

Table 3: Verification accuracy of different divisons on MNIST under different perturbations where $f = 4, h = 32$ and $\ell = 3$.

	① 2-tri-up		② 2-tri-down		③ 4-tri		④ 2-rec-vec		⑤ 2-rec-hor		⑥ 4-rec		⑦ 9-rec		⑧ 16-rec	
ϵ	Acc.	Time	Acc.	Time	Acc.	Time	Acc.	Time	Acc.	Time	Acc.	Time	Acc.	Time	Acc.	Time
0.005	**99**	7.52	**99**	10.87	**99**	14.40	**99**	10.38	**99**	9.88	**99**	19.09	**99**	40.04	**99**	48.26
0.008	**98**	8.50	**98**	10.62	**98**	15.63	**98**	10.15	**98**	9.92	**98**	19.01	**98**	40.30	**98**	48.56
0.011	91	8.87	91	11.28	**92**	16.28	91	10.91	91	10.73	**92**	19.19	**92**	37.60	**92**	48.37
0.014	71	9.56	71	10.86	75	15.61	68	12.87	71	13.07	78	20.90	80	42.55	**81**	48.94
0.017	33	10.15	33	12.97	50	16.78	33	13.73	33	14.01	58	19.39	59	39.81	**62**	48.89
0.020	18	11.25	18	15.58	25	17.49	16	13.22	18	12.95	28	19.72	34	44.14	**38**	50.42

Table 4: Verification accuracy of different divisons on MNIST under different perturbations where $f = 4, h = 64$ and $\ell = 1$.

	① 2-tri-up		② 2-tri-down		③ 4-tri		④ 2-rec-vec		⑤ 2-rec-hor		⑥ 4-rec		⑦ 9-rec		⑧ 16-rec	
ϵ	Acc.	Time	Acc.	Time	Acc.	Time	Acc.	Time	Acc.	Time	Acc.	Time	Acc.	Time	Acc.	Time
0.005	**100**	15.01	**100**	18.81	**100**	27.90	**100**	21.51	**100**	17.52	**100**	38.02	**100**	80.22	**100**	96.43
0.008	**99**	16.43	**99**	18.84	**99**	30.89	**99**	19.70	**99**	18.99	**99**	37.85	**99**	77.75	**99**	97.53
0.011	94	16.94	94	17.74	**95**	32.86	94	21.38	94	20.26	**95**	37.85	**95**	71.62	**95**	96.38
0.014	78	17.76	78	20.15	82	30.74	77	24.43	78	23.04	83	38.69	87	83.78	**88**	95.85
0.017	50	18.27	50	23.14	66	32.05	47	25.23	50	24.01	66	37.64	**69**	75.57	**69**	97.12
0.020	22	20.63	22	26.45	27	32.83	21	22.32	22	23.37	27	38.39	34	80.95	**38**	96.56

Table 5: Verification accuracy of different divisons on MNIST under different perturbations where $f = 4, h = 128$ and $\ell = 1$.

	① 2-tri-up		② 2-tri-down		③ 4-tri		④ 2-rec-vec		⑤ 2-rec-hor		⑥ 4-rec		⑦ 9-rec		⑧ 16-rec	
ϵ	Acc.	Time	Acc.	Time	Acc.	Time	Acc.	Time	Acc.	Time	Acc.	Time	Acc.	Time	Acc.	Time
0.005	**98**	7.29	**98**	11.08	**98**	13.16	**98**	9.63	**98**	9.02	**98**	16.69	**98**	36.70	**98**	43.66
0.008	88	8.36	88	13.08	91	15.06	88	10.69	88	10.38	**92**	16.96	**92**	37.76	**92**	44.92
0.011	37	11.09	38	13.97	70	17.96	37	15.17	37	14.25	**71**	18.33	**71**	41.17	**71**	44.51
0.014	7	12.78	7	14.45	20	18.50	7	18.25	7	18.24	27	19.24	**32**	44.45	**32**	45.21
0.017	0	11.93	0	16.41	6	18.06	0	19.17	0	17.04	8	18.77	**10**	40.60	**10**	44.77
0.020	0	13.10	0	19.40	1	18.26	1	13.12	1	15.31	**3**	18.92	**3**	40.76	**3**	45.31

Table 6: Verification accuracy of different divisons on MNIST under different perturbations where $f = 7$, $h = 32$ and $\ell = 1$.

	① 2-tri-up		② 2-tri-down		③ 4-tri		④ 2-rec-vec		⑤ 2-rec-hor		⑥ 4-rec		⑦ 9-rec		⑧ 16-rec	
ϵ	Acc.	Time	Acc.	Time	Acc.	Time	Acc.	Time	Acc.	Time	Acc.	Time	Acc.	Time	Acc.	Time
-100	31	213.78	31	218.91	52	201.46	50	206.33	51	210.88	55	205.84	64	435.15	**67**	457.99
-95	18	219.62	18	263.13	44	301.25	41	224.08	41	219.81	48	299.48	50	486.18	**52**	517.85
-90	14	254.91	13	216.15	27	297.09	35	232.41	35	241.06	29	303.30	**30**	561.35	**30**	579.64
-85	2	233.72	3	227.14	15	297.39	10	219.36	10	223.75	15	295.24	18	611.89	**23**	624.35
-80	0	286.03	0	286.77	4	294.80	3	272.17	3	298.10	4	303.92	4	644.22	**6**	681.95

Table 7: Verification accuracy of different divisons on GSC dataset under different perturbations.

	① 2-tri-up		② 2-tri-down		③ 4-tri		④ 2-rec-vec		⑤ 2-rec-hor		⑥ 4-rec		⑦ 9-rec		⑧ 16-rec	
ϵ	Acc.	Time	Acc.	Time	Acc.	Time	Acc.	Time	Acc.	Time	Acc.	Time	Acc.	Time	Acc.	Time
-100	**97**	89.48	**97**	91.40	**97**	101.11	**97**	87.99	**97**	84.54	**97**	96.43	**97**	203.35	**97**	235.52
-95	**95**	94.57	**95**	91.81	**95**	112.79	**95**	96.64	**95**	94.86	**95**	95.66	**95**	207.19	**95**	253.08
-90	87	105.98	87	110.32	89	116.45	87	108.57	87	104.45	89	112.23	**91**	223.49	**91**	251.12
-85	84	109.01	85	108.14	86	123.89	85	110.83	84	114.50	86	118.59	**89**	241.20	**89**	289.95
-80	70	118.04	68	120.22	78	137.39	71	115.96	61	117.85	81	133.43	82	258.86	**85**	291.99

Table 8: Verification accuracy of different divisons on FSDD dataset under different perturbations.

	① 2-tri-up		② 2-tri-down		③ 4-tri		④ 2-rec-vec		⑤ 2-rec-hor		⑥ 4-rec		⑦ 9-rec		⑧ 16-rec	
ϵ	Acc.	Time	Acc.	Time	Acc.	Time	Acc.	Time	Acc.	Time	Acc.	Time	Acc.	Time	Acc.	Time
0.05	94	81.73	94	74.41	94	96.40	94	73.23	94	71.89	**95**	93.19	**95**	181.42	**95**	222.67
0.07	68	78.59	64	74.50	71	99.14	70	77.38	70	77.95	78	94.47	82	189.20	**88**	233.38
0.09	39	72.92	42	83.98	56	102.29	48	81.56	49	80.12	61	104.13	66	205.32	**69**	253.22
0.11	21	93.63	18	86.26	30	111.85	21	86.12	21	87.90	39	113.60	46	222.91	**54**	273.01
0.13	10	89.78	10	90.65	17	117.66	10	88.75	9	86.03	19	114.73	22	227.63	**28**	281.49
0.15	0	96.15	0	92.01	**4**	117.34	1	92.34	0	94.76	**4**	116.32	**4**	236.38	**4**	293.80

Table 9: Verification accuracy of different divisons on RT dataset under different perturbations.