VERAM: View-Enhanced Recurrent Attention Model for 3D Shape Classification Supplemental materials

Songle Chen, Lintao Zheng, Yan Zhang, Zhixin Sun and Kai Xu*

1 Additional material for learning with view confidence

As described in subsection 3.2.3, the classification subnetwork of VERAM is easier to train than the view estimation one, and learned attention policy can be easily trapped in local optimization. In practice, with the same network, we find the accuracy of each category from different trained BoundaryRAM model fluctuates widely.

In table 1, columns from 2 to 3 show the min, max, mean and variance of the class-level accuracy of the 5 trained BoundaryRAM models with $T = 4$ on ModelNet10, and columns from 4 to 5 show that of the other 5 trained BoundaryRAM models with $T = 6$. When the time step T is set to 4, the average class-level accuracy is 93.6% and the average variance is 1.66. However, when the number of time step T increases to 6, the average variance rises to 3.26, and average class-level accuracy drops to 90.1%. Obviously, with the number of time steps increases, the model trained by BoundaryRAM will be more unstable.

To alleviate this problem, based on BoundaryRAM, in subsection 3.2.3, we propose a method of learning with view confidence for REINFORCE to provide effective guidance to agent on where to deploy the model's attention, here called ConfRAM. In table 1, the right two columns show the min, max, mean and variable of the average class-level accuracy of 5 trained ConfRAM models on ModelNet10 with 6 time steps. ConfRAM achieves 95.0% average class-level accuracy and the variance is only 0.84. The variance is smaller than that of BoundaryRAM with 4 time steps, and much smaller than that of BoundaryRAM with 6 time steps.

TABLE 1: Statistical variables of class-level accuracy of 5 trained BoundaryRAM and ConfRAM models on ModelNet10 with different time steps T, taking Resnet as CNN and linear mapping as RNN.

ModelNet10	BoundaryRAM T=4		BoundaryRAM T=6		ConfRAM T=6	
	Min Max	Mean Var	Min Max	Mean Var	Min max	Mean Var
bathtub	96.0 98.0	97.5 0.87	78.0 96.0	90.0 7.35	96.0 98.0	97.3 0.94
bed	97.0 100.0	98.5 1.12	90.0 98.0	93.8 3.03	98.0 99.0	98.3 0.47
chair	98.0 99.0	98.5 0.50	93.0 99.0	96.8 2.28	99.0 100	99.7 0.47
desk	83.7 90.7	86.9 2.78	79.1 91.9	86.3 5.66	89.5 90.7	90.3 0.55
dresser	90.7 95.4	92.2 1.91	80.2 89.5	85.8 3.43	86.1 91.9	89.1 2.39
monitor	96.0 99.0	98.0 1.23	95.0 98.0	97.3 1.30	99.0 100	99.3 0.47
night stand	72.1 84.9	78.2 5.60	66.3 73.3	70.9 2.73	83.7 84.9	84.5 0.55
sofa	95.0 97.0	96.3 0.83	93.0 95.0	93.8 0.83	96.0 97.0	96.7 0.47
table	87.0 92.0	89.8 1.79	84.0 94.0	88.8 4.76	93.0 97.0	95.0 1.63
toilet	100 100	100 0.00	97.0 100	98.0 1.23	99.0 100	99.7 0.47
Average	- -	**93.6** **1.66**	- -	**90.1** **3.26**	- -	**95.0** **0.84**

This indicates that learning with view confidence can force our model to select more discriminative views and avoid trapping in local minima with low discriminative views.

Fig. 1 gives a detail example. The 3D shape is $night\_stand\_0224$. The first row shows the visited view of each time step when using a trained BoundaryRAM model with $T = 6$. The predication probability of this shape belonging to $night\_stand$ is only 0.39 and it is misclassified as $dresser$. The second row is the visited view of each time step when using another trained BoundaryRAM model but $T = 4$. The shape is predicated correctly with probability 0.75. The third row shows the visited view of each time

- *Corresponding author: Kai Xu (kevin.kai.xu@gmail.com).*
- *Songle Chen is with Jiangsu High Technology Research Key Laboratory for Wireless Sensor Networks, Nanjing University of Posts and Telecommunications. Email: chensongle@njupt.edu.cn.*
- *Lintao Zheng is with School of Computer, National University of Defense Technology. Email: lintaozheng1991@gmail.com.*
- *Yan Zhang is with Department of Computer Science and Technology, Nanjing University. Email: zhangyannju@nju.edu.cn.*
- *Zhixin Sun is with Jiangsu High Technology Research Key Laboratory for Wireless Sensor Networks, Nanjing University of Posts and Telecommunications. Email: sunzx@njupt.edu.cn.*
- *Kai Xu is with School of Computer, National University of Defense Technology. Email: kevin.kai.xu@gmail.com.*

step when using a trained ConfRAM model. The shape is classified correctly and the probability reaches to 0.93. This means ConfRAM has collected more positive information for the predication.

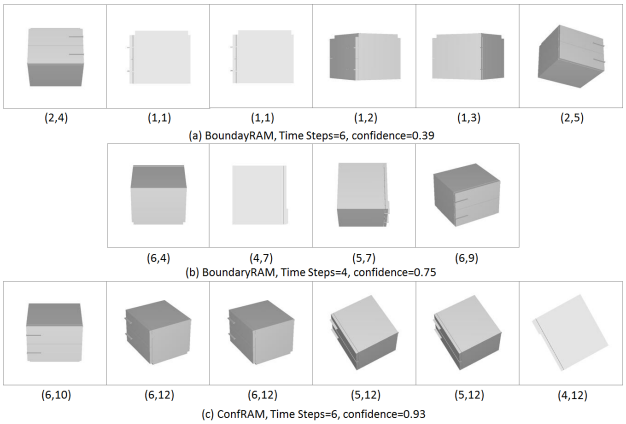

Fig. 1: Visited view of each time step when predicating 3D shape $night\_stand\_0224$ in the testing set of ModelNet10, with (a) BoundaryRAM ($T = 6$), (b) BoundaryRAM ($T = 4$), and (c) ConfRAM ($T = 6$).

2 Additional Material for Learning with Location Constrains

ConfRAM can achieve a stable and fairly good performance. However, the visited view location of each time step may overlap. As shown in Fig. 1 (c), both location (6, 12) and (5, 12) repeat two times. In subsection 3.2.4, based on BoundaryRAM and ConfRAM, a weak regular term is adopted by VERAM to keep the visited views separated from each other, here called LocRAM, equally to the whole VERAM.

Correspondingly to Fig. 1, Fig. 2 shows the visited view of each time step when using a trained LocRAM model with $T = 6$ for $night\_stand\_0224$ in ModelNet10, it can be seen that the regular term has taken effect.

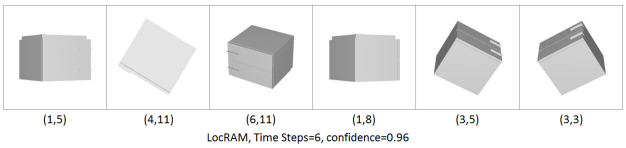

Fig. 2: Visited view of each time step when predicating 3D shape $night\_stand\_0224$ in the testing set of ModelNet10, with LocRAM ($T = 6$).

3 Heat Maps of View Location Frequency

Taking Resnet as CNN and linear mapping as RNN, VERAM achieves class-level accuracy 91.5% (with 6 views) on Modelnet40. Fig. 3 shows the heat maps of view location frequency of applying this trained model to all airplane shapes in the testing set of ModelNet40.

The most frequent locations of step 1 to step 6 are (1,8), (9,11), (8,10), (6,7), (7,5) and (8,4). We select two 3D shapes

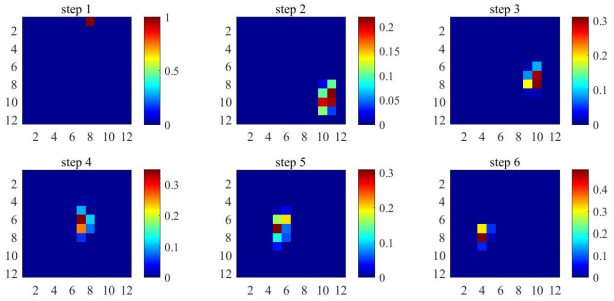

Fig. 3: Heat maps of the view location frequency of the trained VERAM model applied to all airplane shapes in the testing set of ModelNet40.

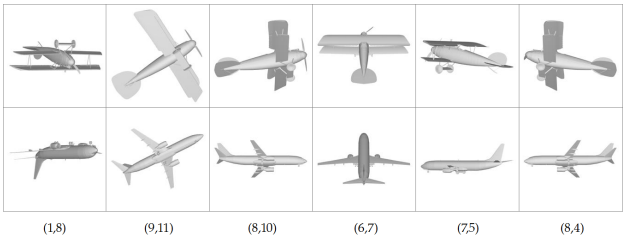

Fig. 4: 2D images of two airplanes rendered from the viewpoints at the most frequent locations.

from the category airplane and show the rendered images corresponding to the most frequent locations in Fig. 4.

Fig. 5 shows the heat maps of view location frequency of applying the trained VERAM model to all bookshelf shapes in the testing set of ModelNet40.

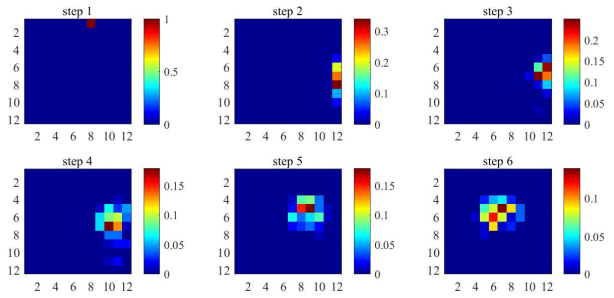

Fig. 5: Heat maps of the view location frequency of the trained VERAM model applied to all bookshelf shapes in the testing set of ModelNet40.

The most frequent view locations of step 1 to step 6 are (1,8), (8,12), (7,11), (7,10), (5,9) and (5,7). We select two 3D shapes from the category bookshelf and show the rendered images corresponding to the most frequent locations in Fig. 6.

4 Classification Examples

VERAM can quickly converge within a few time steps. Taking Resnet as CNN and linear mapping as RNN, with only 3 time steps, the best class-level accuracy VERAM obtained on ModelNet10 is 95.64%. For each category, we

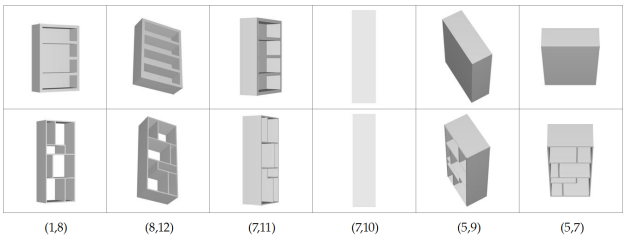

Fig. 6: 2D images of two bookshelves rendered from the viewpoints at the most frequent locations.

select the first 3D shape in the testing set and show the visited view of each time step in Fig. 7.

All these shapes are correctly classified except the first shape of category $sofa$, which is misclassified as $night\_stand$ with the category probability 79.0%. We notice that the first component of location is within 5 to 8, and 7, corresponding to the front face, only appears 3 times. This indicates the learned model prefers the inclined views. The second component is within 4 to 12 and more widespread.

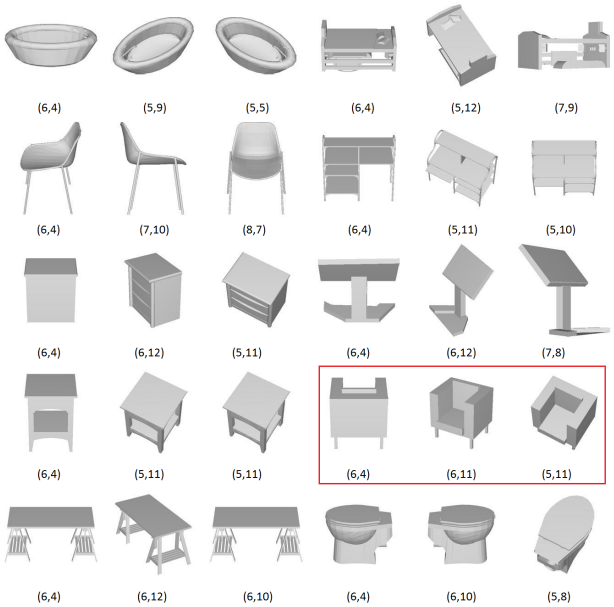

Fig. 7: Visited view of each time step of the first 3D shape in the testing set of ModelNet10 when using the trained VERAM model with 3 time steps.

Taking Resnet as CNN and linear mapping as RNN, with 7 time steps, the best class-level VERAM achieved on ModelNet10 is 96.11%. The accuracy increases 0.47% but the time steps T also increase 4. Using this best model, we also select the first 3D shape in the testing set of each category and show the visited views in Fig. 8. The first shape of category $sofa$ misclassified in Fig. 7 is correctly classified. The situation is a little trickier, for the first shape of category $dresser$ correctly classified in Fig. 7 is misclassified as $night\_stand$ with probability 99.6%.

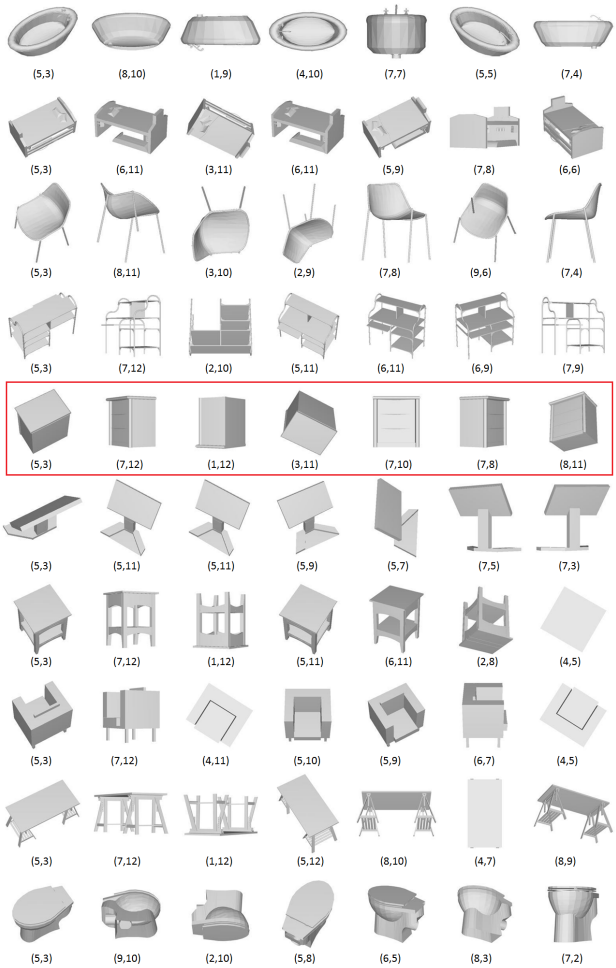

Fig. 8: Visited view of each time step of the first 3D shape in the testing set of ModelNet10 when using the trained VERAM model with 7 time steps.