SUPPLEMENTARY MATERIAL FOR THE ICIP 2025 PAPER : AN END-TO-END CLASS-AWARE AND ATTENTION-GUIDED MODEL FOR OBJECT STATE CLASSIFICATION

1. SUPPLEMENTARY SECTION

1.1. Dataset Details

Table 1 presents the details for the 4 benchmark datasets used in the ablation study and experimental evaluation.

1.2. Additional Results for Ablation study

Table 2 and Table 3 present the results of the ablation study that due to space consideration could not be included in the main paper. Specifically, Table 2 shows the results for the 5 different values of the number of attention heads and Table 3 shows the results for the 11 different values of the γ parameter.

1.3. Additional Experimental Results with different backbones

Table 4-Table 6 present the experiments conducted with 3 different backbones: Vgg16[1], Vit_b16[2] and Vit_b32[2]. Specifically, Table 4 shows the results for Vgg16, Table 5 shows the results for Vit_b16 and Table 6 shows the results for Vit_b32 respectively.

Dataset	Train	Val	Test	S	O	VO	TO	A
OSDD	6,977	1,124	5,275	9	14	35	126	2.36
CGQA	509	65	806	5	18	41	90	1.71
MIT	171	42	349	5	14	20	70	1.57
VAW	2,752	516	1,584	9	23	51	207	2.61

Table 1. Dataset details . Train/Val/Test: Number of Training/Validation/Testing Images. S: Number State classes. O: Number of Object classes. VO/TO: Valid/Total Object-State combinations. A: Average number of states per object.

Heads	Train	**OSDD**	**CGQA**	**MIT**	**VAW**
2		60.0/58.2	34.9/32.1	44.6/37.5	33.4/33.9
4		60.4/58.4	34.8/34.3	46.5/38.6	33.3/33.4
8	**OSDD**	59.8/57.6	33.0/ 33.4	44.4/36.7	32.4/33.6
16		60.1/57.7	32.4/32.5	44.2/36.3	32.1/33.4
32		60.0/ 58.4	30.7/32.2	44.6/37.1	31.5/33.2
2		41.3/42.8	66.8/44.4	43.8/39.3	52.5/41.4
4		42.1/ 44.4	66.0/45.0	44.9/39.8	52.4/42.2
8	**CGQA**	40.9/43.0	66.6/45.8	44.0/39.5	52.4/42.0
16		42.2/ 43.8	66.4/44.9	43.5/38.8	53.0/42.6
32		41.4/43.8	65.5/45.8	45.1/40.5	52.6/43.0
2		30.0/43.3	30.9/36.9	76.4/74.2	30.9/46.7
4		30.0/43.4	30.5/28.0	74.8/72.5	31.7/45.9
8	**MIT**	30.4/43.2	29.1/30.9	75.7/73.7	30.0/45.2
16		29.5/42.6	28.6/32.1	75.2/73.3	29.9/46.4
32		29.9/43.1	32.2/37.0	75.9/73.7	32.2/47.6
2		32.8/32.6	61.9/47.5	38.3/35.0	58.5/50.0
4		31.2/32.0	61.4/47.9	37.7/33.9	57.7/49.8
8	**VAW**	32.8/32.2	59.9/47.3	37.4/33.8	57.7/49.1
16		33.1/ 33.1	59.6/47.3	38.2/35.2	57.4/50.4
32		32.0/32.7	61.3/46.3	36.0/33.7	58.1/50.2

Table 2. Ablation study for the number of attention heads. The reported values have been averaged over the 11 different values of the γ parameter

γ	Train	**OSDD**	**CGQA**	**MIT**	**VAWQ**
0		13.3/12.5	12.2/12.4	10.4/9.9	11.4/12.8
0.1		63.7/60.5	36.4/33.4	49.9/41.5	34.6/35.9
0.2		62.6/60.2	33.0/35.4	46.5/37.4	32.7/35.3
0.3		61.2/60.1	35.8/34.2	46.5/37.9	33.6/35.5
0.4		61.8/59.1	32.7/36.2	45.8/37.6	33.4/35.3
0.5	**OSDD**	60.4/59.1	36.4/34.5	47.5/39.8	34.3/35.8
0.6		60.3/58.4	34.5/31.7	46.8/38.3	34.3/34.4
0.7		60.0/58.5	32.4/30.7	43.1/35.4	32.8/33.8
0.8		58.3/57.2	33.5/32.8	44.4/37.6	33.0/32.8
0.9		55.7/53.2	32.1/30.7	38.1/31.1	29.8/28.5
1		56.5/54.4	24.7/29.6	40.2/35.6	26.6/27.7
0		19.8/18.9	19.4/18.2	20.5/20.4	18.1/18.5
0.1		38.9/40.9	68.1/47.3	43.0/38.5	53.9/42.6
0.2		45.5/45.1	67.2/47.9	44.5/39.2	53.6/43.2
0.3		44.6/45.6	67.5/48.2	43.9/38.9	54.3/44.5
0.4		41.9/43.8	66.9/46.7	43.0/38.8	54.1/45.4
0.5	**CGQA**	42.5/44.8	66.7/46.4	44.5/40.2	54.0/44.3
0.6		40.8/42.8	66.6/45.4	42.9/38.7	53.4/43.5
0.7		43.7/45.2	66.2/45.3	46.4/41.3	53.0/43.6
0.8		42.5/45.7	65.1/45.6	46.1/41.6	52.5/43.0
0.9		42.4/44.5	67.1/43.1	44.3/39.4	50.8/39.4
1		33.0/37.1	61.2/36.2	43.9/39.4	46.3/33.1
0		19.8/16.6	21.4/22.7	18.5/18.4	22.0/20.0
0.1		32.7/45.0	36.6/35.5	77.2/76.0	35.0/50.8
0.2		31.7/44.5	32.1/32.9	77.3/76.5	32.8/55.8
0.3		31.7/43.8	32.2/29.8	76.4/74.4	32.5/49.9
0.4		32.0/44.0	30.6/34.5	76.2/74.4	31.7/45.6
0.5	**MIT**	30.1/44.1	31.8/35.3	76.0/75.1	33.0/46.3
0.6		28.3/42.6	29.0/31.8	76.6/75.2	28.5/40.3
0.7		28.5/42.6	28.9/32.1	77.7/75.4	29.4/41.8
0.8		27.4/42.1	24.9/35.5	76.7/73.3	27.4/44.1
0.9		29.4/41.2	27.9/31.3	72.3/69.7	28.8/42.3
1		27.6/41.0	28.6/31.0	69.7/64.9	30.4/46.7
0		10.4/11.2	12.6/12.5	10.5/10.1	11.1/10.7
0.1		35.5/33.3	64.5/51.3	44.7/38.5	62.4/54.1
0.2		35.1/33.7	61.1/50.5	41.8/36.2	60.3/52.7
0.3		35.0/34.1	62.5/50.8	35.8/32.5	59.8/53.0
0.4		35.1/34.3	61.9/49.0	41.5/37.4	59.6/52.1
0.5	**VAW**	33.5/33.1	61.4/49.0	42.0/39.0	58.8/51.1
0.6		31.8/32.1	61.6/45.0	35.8/33.1	56.7/47.5
0.7		29.1/31.6	61.1/44.8	34.2/32.5	57.0/48.8
0.8		30.2/31.5	61.0/44.6	34.6/32.4	57.3/48.4
0.9		30.0/32.1	59.5/45.7	34.7/32.5	57.3/49.3
1		28.5/29.0	53.7/41.8	30.2/29.1	49.7/42.2

Table 3. Ablation study for the parameter γ. The reported values have been averaged over the 5 different values of the number of attention heads.

Metric	Train \ Test	OSDD	CGQA	MIT	VAW
WA	**OSDD**	57.6 / 61.5	31.9 / 31.0	45.0 / 34.7	28.6 / 35.6
AA		57.4 / 60.4	39.1 / 32.7	36.3 / 32.5	35.0 / 36.1
AA	**CGQA**	32.6 / 37.1	57.3 / 65.3	53.6 / 23.5	48.6 / 49.3
AA		40.1 / 20.9	43.6 / 39.5	47.2 / 28.3	45.0 / 21.7
WA	**MIT**	31.6 / 35.8	33.3 / 19.7	80.6 / 80.6	34.0 / 37.9
AA		44.9 / 21.2	38.4 / 36.5	76.9 / 78.3	62.9 / 23.8
WA	**VAW**	30.7 / 32.0	55.5 / 61.5	54.2 / 41.8	54.8 / 61.7
AA		26.7 / 31.7	38.2 / 58.0	51.2 / 37.0	49.3 / 55.7

Table 4. Experimental results for our method and the three baseline methods using Vgg16 as backbone. 1st/2nd of our method/OA-SC. WA: Weighted Accuracy. AA: Average Accuracy.

Metric	Train \ Test	OSDD	CGQA	MIT	VAW
WA	**OSDD**	52.4 / 65.2	25.9 / 34.3	42.5 / 27.6	30.4 / 37.5
AA		43.3 / 62.5	33.3 / 43.1	37.8 / 27.0	31.7 / 39.0
AA	**CGQA**	9.0 / 34.3	61.0 / 70.4	35.7 / 24.5	42.5 / 53.2
AA		20.0 / 19.7	20.0 / 46.2	33.3 / 31.5	20.1 / 22.9
WA	**MIT**	22.5 / 37.5	20.2 / 24.9	71.4 / 86.7	18.7 / 32.5
AA		38.1 / 20.3	29.4 / 41.4	60.2 / 81.5	24.8 / 20.1
WA	**VAW**	28.9 / 33.7	67.8 / 67.6	41.7 / 43.9	55.8 / 66.3
AA		25.1 / 31.0	49.7 / 54.6	32.3 / 40.8	37.5 / 58.4

Table 5. Experimental results for our method and the three baseline methods using Vit_B16 as backbone. 1st/2nd of our method/OA-SC. WA: Weighted Accuracy. AA: Average Accuracy.

Metric	Train \ Test	OSDD	CGQA	MIT	VAW
WA	**OSDD**	52.2 / 63.7	20.0 / 38.0	42.5 / 42.9	31.5 / 39.4
AA		48.6 / 61.5	28.6 / 40.7	39.4 / 41.2	36.2 / 40.5
AA	**CGQA**	9.2 / 33.7	62.0 / 68.5	35.7 / 22.4	43.0 / 51.6
AA		20.1 / 19.2	23.4 / 43.7	33.3 / 25.9	20.9 / 22.1
WA	**MIT**	18.0 / 38.2	17.5 / 30.5	64.3 / 78.6	18.3 / 36.6
AA		35.1 / 21.3	23.7 / 41.4	53.6 / 72.4	21.7 / 20.9
WA	**VAW**	19.3 / 33.2	61.6 / 69.5	37.5 / 41.8	46.7 / 67.1
AA		20.2 / 32.6	33.2 / 58.9	30.0 / 41.4	25.2 / 59.1

Table 6. Experimental results for our method and the three baseline methods using Vit_B32 as backbone. 1st/2nd of our method/OA-SC. WA: Weighted Accuracy. AA: Average Accuracy.

2. REFERENCES

[1] Karen Simonyan, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[2] Alexey Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.