Supplementary Material of "Global2Local: Efficient Structure Search for Video Action Segmentation"

Shang-Hua Gao[1*] Qi Han[1*] Zhong-Yu Li[1]
Pai Peng[2] Liang Wang[3] Ming-Ming Cheng[1 †]
TKLNDST, CS, Nankai University[1] Tencent[2] NLPR[3]
http://mmcheng.net/g2lsearch

GTEA	F@0.1	F@0.25	F@0.5	Edit	Acc
Spatial CNN	41.8	36.0	25.1	-	54.1
Bi-LSTM	66.5	59.0	43.6	-	55.5
Dilated TCN	58.8	52.2	42.2	-	58.3
ST-CNN	58.7	54.4	41.9	-	60.6
TUnet	67.1	63.7	51.9	60.3	59.9
ED-TCN	72.2	69.3	56.0	-	64.0
TResNet	74.1	69.9	57.6	64.4	65.8
TricorNet	76.0	71.1	59.2	-	64.8
TRN	77.4	71.3	59.1	72.2	67.8
TDRN	79.2	74.4	62.7	74.1	70.1
MS-TCN	87.5	85.4	74.6	81.4	79.2
Reproduce	87.1	83.6	70.4	81.1	75.5
Ours-MS-TCN	89.9	87.3	75.8	84.6	78.5

Table 1. Comparison with state-of-the-art methods on the GTEA dataset.

1. Performance on the GTEA dataset

We also compare our proposed global-to-local search with existing action segmentation methods on the small scale GTEA dataset, as shown in Tab. 1. Based on the MS-TCN architecture, the global-to-local searched structure surpasses the human-designed baseline with 2.4% on F@0.1. Also, our global-to-local based MS-TCN has a considerable performance gain compared with existing methods.

2. Searching Cost

We report the cost of our proposed global-to-local search method. When cooperating with MS-TCN, the size of the receptive field combination search space is 1024^{40}. The cost of searching on such a huge space using existing methods is unaffordable. Our proposed global-to-local search decomposes the searching process into the global and local search to find the combination in a coarse-to-fine manner.

GPU Hours	*BreakFast*	*50Salads*	*GTEA*
Global Search	144h	9h	1h
Local Search	2.2h	0.15h	0.05h

Table 2. GPU hours of the global and local search on each fold of different datasets using the RTX 2080Ti GPU.

Actual Searching Hours	*BreakFast*	*50Salads*	*GTEA*
Global Search (8 GPU)	42h	11h	5h
Local Search (1 GPU)	3.3h	0.2h	0.07h

Table 3. The actual searching time of the global and local search on each fold of different datasets using the RTX 2080Ti GPU×8 server.

In the global search process, we search structures on each fold of datasets. Since the main bottleneck of the search method is the GPU resources, we report the GPU hours of the proposed global-to-local search in Tab. 2. In our work, 1 GPU hour means using one RTX 2080Ti GPU for an hour. For each fold in the dataset, the global search takes 144 GPU hours on the BreakFast dataset, 9 GPU hours on the 50Salads dataset, and 1 GPU hours on the GTEA dataset. Through the global search, we can find multiple new well-performed structures that have different patterns and achieve better performance than human-designed patterns. The local search further fine-tunes the global searched structures in the dense but local search space. Based on one of the global-searched structures, the local search takes about 2.2 GPU hours on the BreakFast dataset, 0.15 GPU hours on the 50Salads dataset, and 0.05 hours on the GTEA dataset.

Due to the CPU and disk IO speed limitation in actual experiments, the search time is longer than the GPU hour. We report the actual searching time on an RTX 2080Ti GPU×8 server in Tab. 3. For each fold of the dataset, the global search takes 42 hours on the BreakFast dataset, 11 hours on

[*] Equal contribution

[†] M.M. Cheng (cmm@nankai.edu.cn) is the corresponding author.

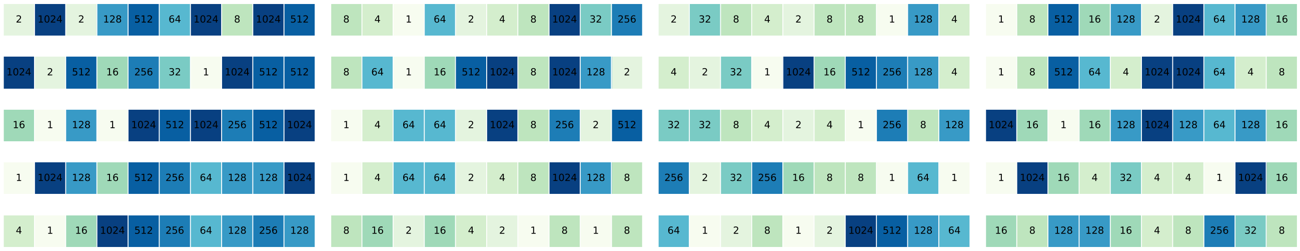
the 50Salads dataset, and 5 hours on the GTEA dataset on an RTX 2080Ti GPU×8 server. The local search takes 3.3 hours on the BreakFast dataset, 12 minutes on the 50Salads dataset, and 4 minutes on the GTEA dataset, with a single RTX 2080Ti GPU.

3. Common Patterns in the Searched Structures

The global search objective is to find more well-performed receptive field combinations that have different patterns than human-designings. We visualize the top-5 well-performed searched structures on each fold of different datasets. The visualization of searched structures of the BreakFast dataset is shown in Fig. 1, Fig. 2, Fig. 3, and Fig. 4. The visualization of searched structures of the 50salads dataset is shown in Fig. 5, Fig. 6, Fig. 7, Fig. 8, and Fig. 9. The visualization of searched structures of the GTEA dataset is shown in Fig. 10, Fig. 11, Fig. 12, and Fig. 13. We visualize the common receptive field combination patterns in the well-performed searched structures in Tab. 4. Common receptive field combination patterns vary among different datasets. Also, different stages in the MS-TCN have different common patterns.

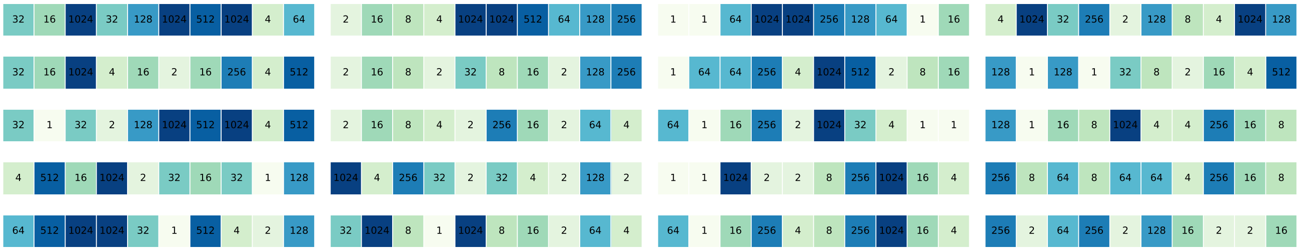
Datasets	stage1	stage2	stage3	stage4
BreakFast	512 64 256	4 1 64 2 1024 8 256	4 32 8 2 2 8 8	1024 16 1 16 128
	2 128 512 32 1024 8 64	64 512 1024 8 1024 128	2 32 8 4	128 1024 128 64 128
	2 128 512 64 256 32	8 4 1 16 512 1024	32 8 2 2	1 16 128 1024 128 64 128
50Salads	32 512 1024	4 1 64 2 1024	1 4 64 32 4 8 1	512 16 128 2 32 1 4
	32 1024 128 32 512 1024	1 64 2 1024 8 256	256 1 1 4 64	128 2 32 1
	2 512 512 1024 256 1024 256	4 1 64 2 1024 8 256	128 4 64 32	2 32 1 4 64 128
GTEA	512 64 256 512 16	512 32 2 1 4 2 16	32 4 128 1 16 16 512	4 8 2 32 16
	1 2 1 32 512 64 256	1 1 32 32 16 2	1 256 16 512 4	32 8 16 1 1024 32
	512 8 256 64 16	1 64 1 1 32	8 128 1 16	2 32 16 1 512 64 32

Table 4. Common receptive field combination patterns in the well-performed searched structures.



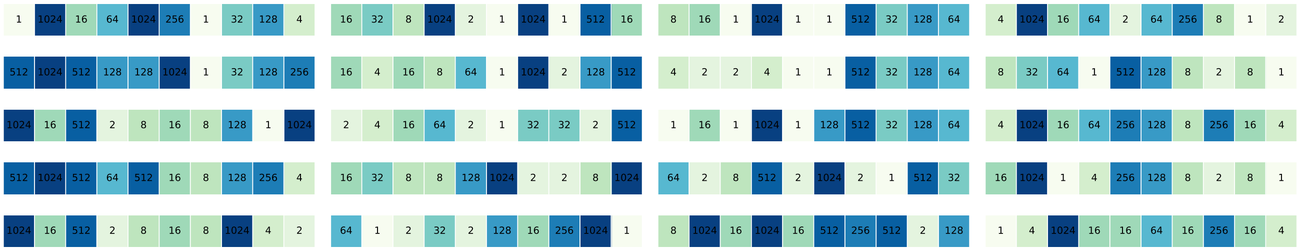
2 1024 2 128 512 64 1024 8 1024 512	8 4 1 64 2 4 8 1024 32 256	2 32 8 4 2 8 8 1 128 4	1 8 512 16 128 2 1024 64 128 16
1024 2 512 16 256 32 1 1024 512 512	8 64 1 16 512 1024 8 1024 128 2	4 2 32 1 1024 16 512 256 128 4	1 8 512 64 4 1024 1024 64 4 8
16 1 128 1 1024 512 1024 256 512 1024	1 4 64 64 2 1024 8 256 2 512	32 32 8 4 2 4 1 256 8 128	1024 16 1 16 128 1024 128 64 128 16
1 1024 128 16 512 256 64 128 128 1024	1 4 64 64 2 4 8 1024 128 8	256 2 32 256 16 8 8 1 64 1	1 1024 16 4 32 4 4 1 1024 16
4 1 16 1024 512 256 64 128 256 128	8 16 2 16 4 2 1 8 1 8	64 1 2 8 1 2 1024 512 128 64	16 8 128 128 16 4 8 256 32 8

Figure 1. Well-performed structures on fold 1 of the BreakFast dataset.



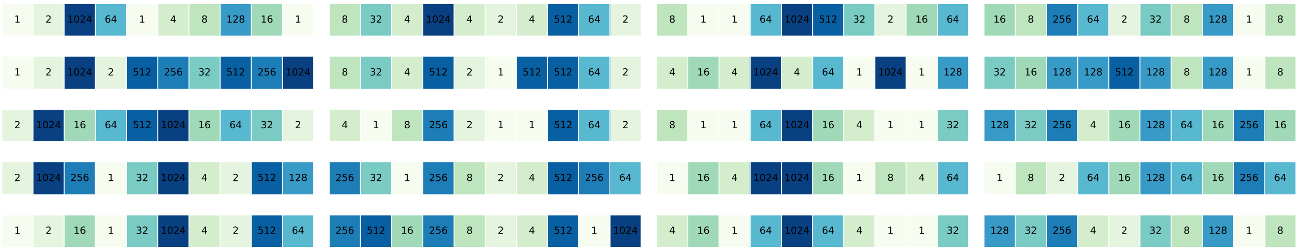
32 16 1024 32 128 1024 512 1024 4 64	2 16 8 4 1024 1024 512 64 128 256	1 1 64 1024 1024 256 128 64 1 16	4 1024 32 256 2 128 8 4 1024 128
32 16 1024 4 16 2 16 256 4 512	2 16 8 2 32 8 16 2 128 256	1 64 64 256 4 1024 512 2 8 16	128 1 128 1 32 8 2 16 4 512
32 1 32 2 128 1024 512 1024 4 512	2 16 8 4 2 256 16 2 64 4	64 1 16 256 2 1024 32 4 1 1	128 1 16 8 1024 4 4 256 16 8
4 512 16 1024 2 32 16 32 1 128	1024 4 256 32 2 32 4 2 128 2	1 1 1024 2 2 8 256 1024 16 4	256 8 64 8 64 64 4 256 16 8
64 512 1024 1024 32 1 512 4 2 128	32 1024 8 1 1024 8 16 2 64 4	64 1 16 256 4 8 256 1024 16 4	256 2 64 256 2 128 16 2 2 16

Figure 2. Well-performed structures on fold 2 of the BreakFast dataset.



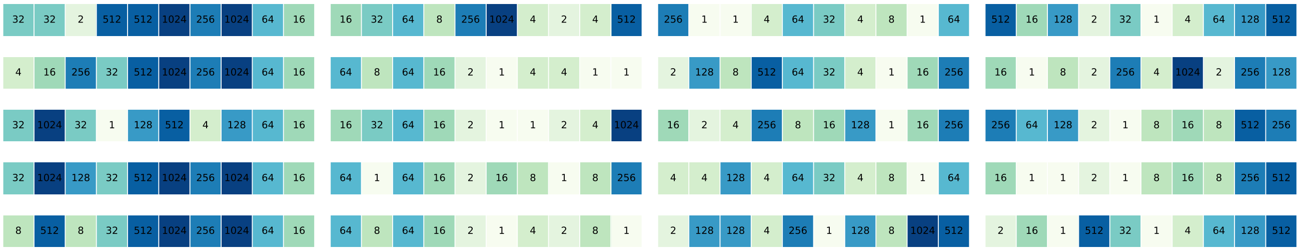
1 1024 16 64 1024 256 1 32 128 4	16 32 8 1024 2 1 1024 1 512 16	8 16 1 1024 1 1 512 32 128 64	4 1024 16 64 2 64 256 8 1 2
512 1024 512 128 128 1024 1 32 128 256	16 4 16 8 64 1 1024 2 128 512	4 2 2 4 1 1 512 32 128 64	8 32 64 1 512 128 8 2 8 1
1024 16 512 2 8 16 8 128 1 1024	2 4 16 64 2 1 32 32 2 512	1 16 1 1024 1 128 512 32 128 64	4 1024 16 64 256 128 8 256 16 4
512 1024 512 64 512 16 8 128 256 4	16 32 8 8 128 1024 2 2 8 1024	64 2 8 512 2 1024 2 1 512 32	16 1024 1 4 256 128 8 2 8 1
1024 16 512 2 8 16 8 1024 4 2	64 1 2 32 2 128 16 256 1024 1	8 1024 16 1024 16 512 256 512 2 128	1 4 1024 16 16 64 16 256 16 4

Figure 3. Well-performed structures on fold 3 of the BreakFast dataset.



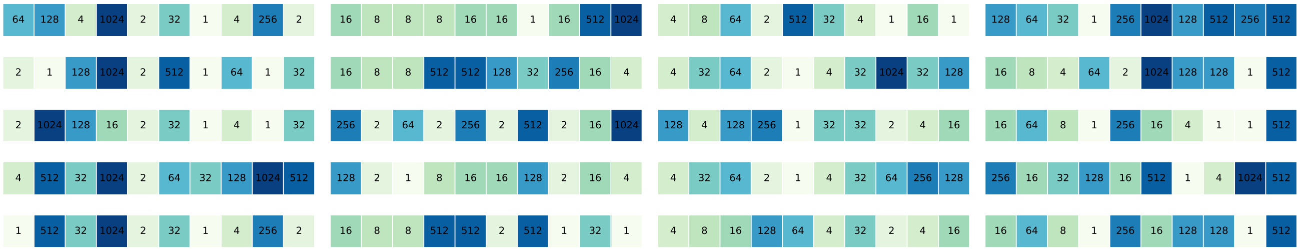
1	2	1024	64	1	4	8	128	16	1	8	32	4	1024	4	2	4	512	64	2	8	1	1	64	1024	512	32	2	16	64	16	8	256	64	2	32	8	128	1	8
1	2	1024	2	512	256	32	512	256	1024	8	32	4	512	2	1	512	512	64	2	4	16	4	1024	4	64	1	1024	1	128	32	16	128	128	512	128	8	128	1	8
2	1024	16	64	512	1024	16	64	32	2	4	1	8	256	2	1	1	512	64	2	8	1	1	64	1024	16	4	1	1	32	128	32	256	4	16	128	64	16	256	16
2	1024	256	1	32	1024	4	2	512	128	256	32	1	256	8	2	4	512	256	64	1	16	4	1024	1024	16	1	8	4	64	1	8	2	64	16	128	64	16	256	64
1	2	16	1	32	1024	4	2	512	64	256	512	16	256	8	2	4	512	1	1024	4	16	1	64	1024	64	4	1	1	32	128	32	256	4	2	32	8	128	1	8

Figure 4. Well-performed structures on fold 4 of the BreakFast dataset.



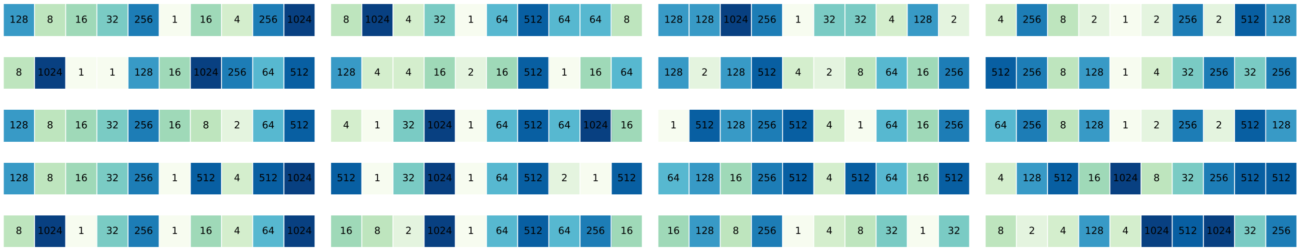
32	32	2	512	512	1024	256	1024	64	16	16	32	64	8	256	1024	4	2	4	512	256	1	1	4	64	32	4	8	1	64	512	16	128	2	32	1	4	64	128	512
4	16	256	32	512	1024	256	1024	64	16	64	8	64	16	2	1	4	4	1	1	2	128	8	512	64	32	4	1	16	256	16	1	8	2	256	4	1024	2	256	128
32	1024	32	1	128	512	4	128	64	16	16	32	64	16	2	1	1	2	4	1024	16	2	4	256	8	16	128	1	16	256	256	64	128	2	1	8	16	8	512	256
32	1024	128	32	512	1024	256	1024	64	16	64	1	64	16	2	16	8	1	8	256	4	4	128	4	64	32	4	8	1	64	16	1	1	2	1	8	16	8	256	512
8	512	8	32	512	1024	256	1024	64	16	64	8	64	16	2	1	4	2	8	1	2	128	128	4	256	1	128	8	1024	512	2	16	1	512	32	1	4	64	128	512

Figure 5. Well-performed structures on fold 1 of the 50Salads dataset.



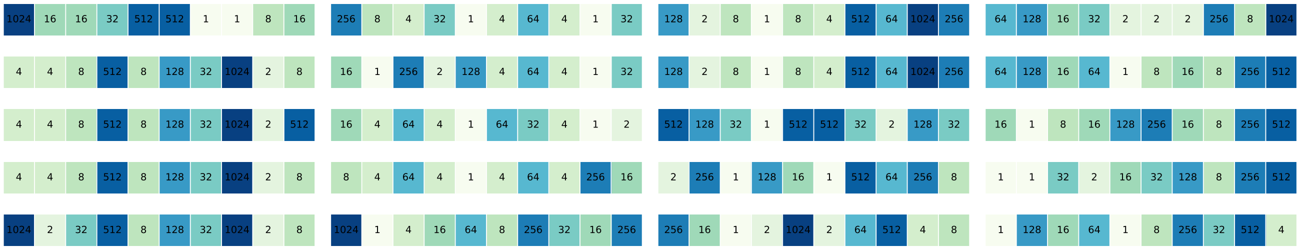
64	128	4	1024	2	32	1	4	256	2	16	8	8	8	16	16	1	16	512	1024	4	8	64	2	512	32	4	1	16	1	128	64	32	1	256	1024	128	512	256	512
2	1	128	1024	2	512	1	64	1	32	16	8	8	512	512	128	32	256	16	4	4	32	64	2	1	4	32	1024	32	128	16	8	4	64	2	1024	128	128	1	512
2	1024	128	16	2	32	1	4	1	32	256	2	64	2	256	2	512	2	16	1024	128	4	128	256	1	32	32	2	4	16	16	64	8	1	256	16	4	1	1	512
4	512	32	1024	2	64	32	128	1024	512	128	2	1	8	16	16	128	2	16	4	4	32	64	2	1	4	32	64	256	128	256	16	32	128	16	512	1	4	1024	512
1	512	32	1024	2	32	1	4	256	2	16	8	8	512	512	2	512	1	32	1	4	8	16	128	64	4	32	2	4	16	16	64	8	1	256	16	128	128	1	512

Figure 6. Well-performed structures on fold 2 of the 50Salads dataset.



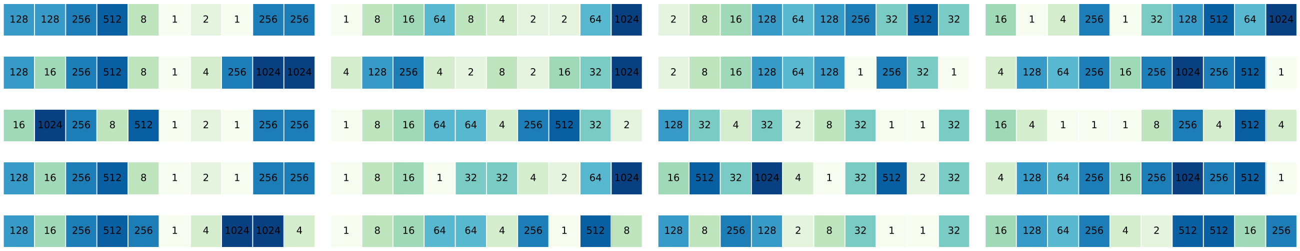
128	8	16	32	256	1	16	4	256	1024	8	1024	4	32	1	64	512	64	64	8	128	128	1024	256	1	32	32	4	128	2	4	256	8	2	1	2	256	2	512	128
8	1024	1	1	128	16	1024	256	64	512	128	4	4	16	2	16	512	1	16	64	128	2	128	512	4	2	8	64	16	256	512	256	8	128	1	4	32	256	32	256
128	8	16	32	256	16	8	2	64	512	4	1	32	1024	1	64	512	64	1024	16	1	512	128	256	512	4	1	64	16	256	64	256	8	128	1	2	256	2	512	128
128	8	16	32	256	1	512	4	512	1024	512	1	32	1024	1	64	512	2	1	512	64	128	16	256	512	4	512	64	16	512	4	128	512	16	1024	8	32	256	512	512
8	1024	1	32	256	1	16	4	64	1024	16	8	2	1024	1	64	512	64	256	16	16	128	8	256	1	4	8	32	1	32	8	2	4	128	4	1024	512	1024	32	256

Figure 7. Well-performed structures on fold 3 of the 50Salads dataset.



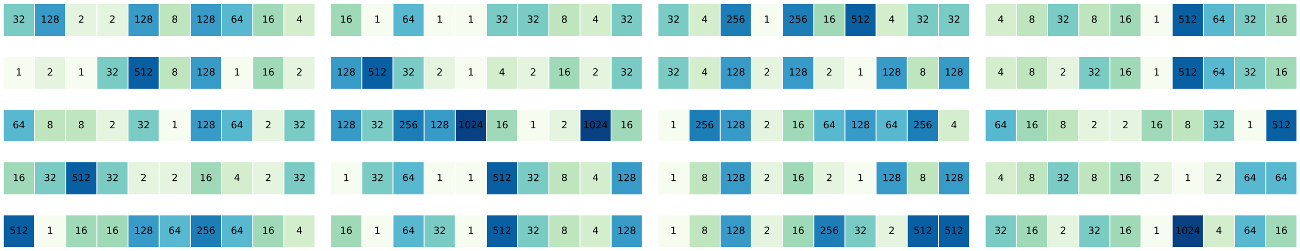
1024	16	16	32	512	512	1	1	8	16	256	8	4	32	1	4	64	4	1	32	128	2	8	1	8	4	512	64	1024	256	64	128	16	32	2	2	2	256	8	1024
4	4	8	512	8	128	32	1024	2	8	16	1	256	2	128	4	64	4	1	32	128	2	8	1	8	4	512	64	1024	256	64	128	16	64	1	8	16	8	256	512
4	4	8	512	8	128	32	1024	2	512	16	4	64	4	1	64	32	4	1	2	512	128	32	1	512	512	32	2	128	32	16	1	8	16	128	256	16	8	256	512
4	4	8	512	8	128	32	1024	2	8	8	4	64	4	1	4	64	4	256	16	2	256	1	128	16	1	512	64	256	8	1	1	32	2	16	32	128	8	256	512
1024	2	32	512	8	128	32	1024	2	8	1024	1	4	16	64	8	256	32	16	256	256	16	1	2	1024	2	64	512	4	8	1	128	16	64	1	8	256	32	512	4

Figure 8. Well-performed structures on fold 4 of the 50Salads dataset.



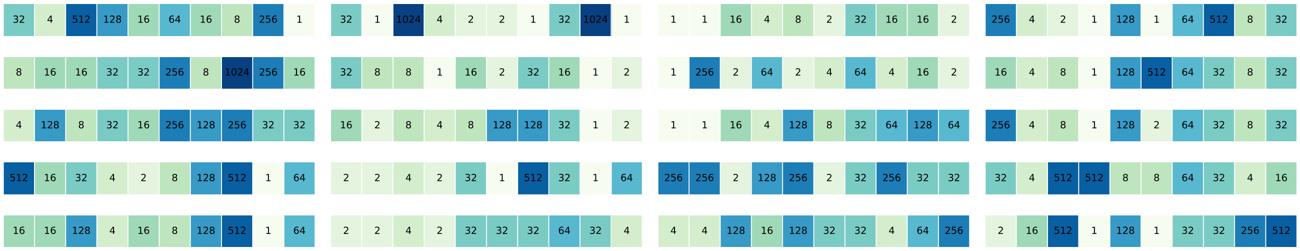
128 128 256 512 8 1 2 1 256 256	1 8 16 64 8 4 2 2 64 1024	2 8 16 128 64 128 256 32 512 32	16 1 4 256 1 32 128 512 64 1024
128 16 256 512 8 1 4 256 1024 1024	4 128 256 4 2 8 2 16 32 1024	2 8 16 128 64 128 1 256 32 1	4 128 64 256 16 256 1024 256 512 1
16 1024 256 8 512 1 2 1 256 256	1 8 16 64 64 4 256 512 32 2	128 32 4 32 2 8 32 1 1 32	16 4 1 1 1 8 256 4 512 4
128 16 256 512 8 1 2 1 256 256	1 8 16 1 32 32 4 2 64 1024	16 512 32 1024 4 1 32 512 2 32	4 128 64 256 16 256 1024 256 512 1
128 16 256 512 256 1 4 1024 1024 4	1 8 16 64 64 4 256 1 512 8	128 8 256 128 2 8 32 1 1 32	16 128 64 256 4 2 512 512 16 256

Figure 9. Well-performed structures on fold 5 of the 50Salads dataset.



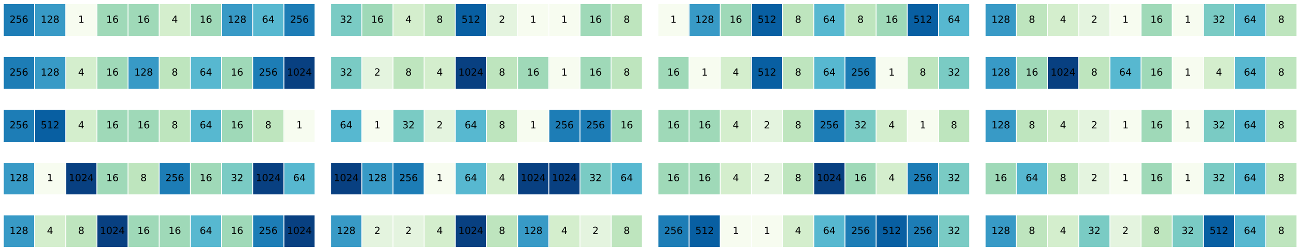
32 128 2 2 128 8 128 64 16 4	16 1 64 1 1 32 32 8 4 32	32 4 256 1 256 16 512 4 32 32	4 8 32 8 16 1 512 64 32 16
1 2 1 32 512 8 128 1 16 2	128 512 32 2 1 4 2 16 2 32	32 4 128 2 128 2 1 128 8 128	4 8 2 32 16 1 512 64 32 16
64 8 8 2 32 1 128 64 2 32	128 32 256 128 1024 16 1 2 1024 16	1 256 128 2 16 64 128 64 256 4	64 16 8 2 2 16 8 32 1 512
16 32 512 32 2 2 16 4 2 32	1 32 64 1 1 512 32 8 4 128	1 8 128 2 16 2 1 128 8 128	4 8 32 8 16 2 1 2 64 64
512 1 16 16 128 64 256 64 16 4	16 1 64 32 1 512 32 8 4 128	1 8 128 2 16 256 32 2 512 512	32 16 2 32 16 1 1024 4 64 16

Figure 10. Well-performed structures on fold 1 of the GTEA dataset.



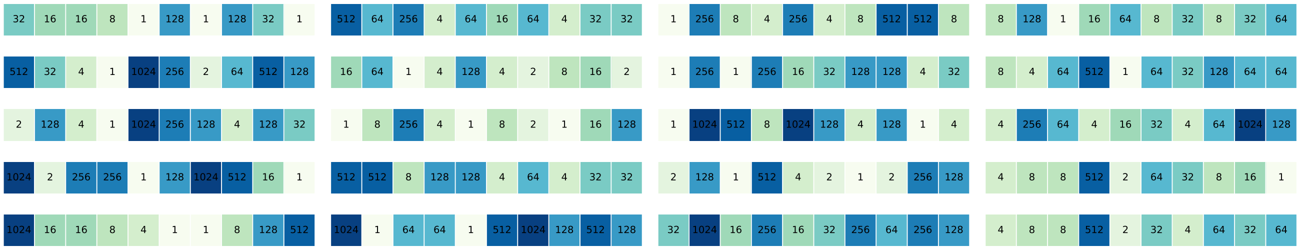
32 4 512 128 16 64 16 8 256 1	32 1 1024 4 2 2 1 32 1024 1	1 1 16 4 8 2 32 16 16 2	256 4 2 1 128 1 64 512 8 32
8 16 16 32 32 256 8 1024 256 16	32 8 8 1 16 2 32 16 1 2	1 256 2 64 2 4 64 4 16 2	16 4 8 1 128 512 64 32 8 32
4 128 8 32 16 256 128 256 32 32	16 2 8 4 8 128 128 32 1 2	1 1 16 4 128 8 32 64 128 64	256 4 8 1 128 2 64 32 8 32
512 16 32 4 2 8 128 512 1 64	2 2 4 2 32 1 512 32 1 64	256 256 2 128 256 2 32 256 32 32	32 4 512 512 8 8 64 32 4 16
16 16 128 4 16 8 128 512 1 64	2 2 4 2 32 32 32 64 32 4	4 4 128 16 128 32 32 4 64 256	2 16 512 1 128 1 32 32 256 512

Figure 11. Well-performed structures on fold 2 of the GTEA dataset.



256 128 1 16 16 4 16 128 64 256	32 16 4 8 512 2 1 1 16 8	1 128 16 512 8 64 8 16 512 64	128 8 4 2 1 16 1 32 64 8
256 128 4 16 128 8 64 16 256 1024	32 2 8 4 1024 8 16 1 16 8	16 1 4 512 8 64 256 1 8 32	128 16 1024 8 64 16 1 4 64 8
256 512 4 16 16 8 64 16 8 1	64 1 32 2 64 8 1 256 256 16	16 16 4 2 8 256 32 4 1 8	128 8 4 2 1 16 1 32 64 8
128 1 1024 16 8 256 16 32 1024 64	1024 128 256 1 64 4 1024 1024 32 64	16 16 4 2 8 1024 16 4 256 32	16 64 8 2 1 16 1 32 64 8
128 4 8 1024 16 16 64 16 256 1024	128 2 2 4 1024 8 128 4 2 8	256 512 1 1 4 64 256 512 256 32	128 8 4 32 2 8 32 512 64 8

Figure 12. Well-performed structures on fold 3 of the GTEA dataset.



32 16 16 8 1 128 1 128 32 1	512 64 256 4 64 16 64 4 32 32	1 256 8 4 256 4 8 512 512 8	8 128 1 16 64 8 32 8 32 64
512 32 4 1 1024 256 2 64 512 128	16 64 1 4 128 4 2 8 16 2	1 256 1 256 16 32 128 128 4 32	8 4 64 512 1 64 32 128 64 64
2 128 4 1 1024 256 128 4 128 32	1 8 256 4 1 8 2 1 16 128	1 1024 512 8 1024 128 4 128 1 4	4 256 64 4 16 32 4 64 1024 128
1024 2 256 256 1 128 1024 512 16 1	512 512 8 128 128 4 64 4 32 32	2 128 1 512 4 2 1 2 256 128	4 8 8 512 2 64 32 8 16 1
1024 16 16 8 4 1 1 8 128 512	1024 1 64 64 1 512 1024 128 512 128	32 1024 16 256 16 32 256 64 256 128	4 8 8 512 2 32 4 64 32 64

Figure 13. Well-performed structures on fold 4 of the GTEA dataset.