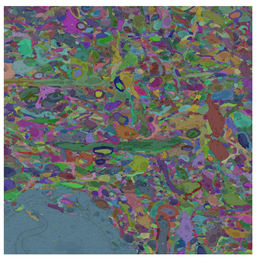
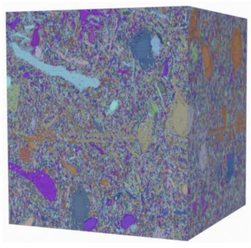
Compresso: Efficient Compression of Segmentation Data For Connectomics

Brian Matejek, Daniel Haehn, Fritz Lekschas, Michael Mitzenmacher, Hanspeter Pfister

John A. Paulson School of Engineering and Applied Sciences

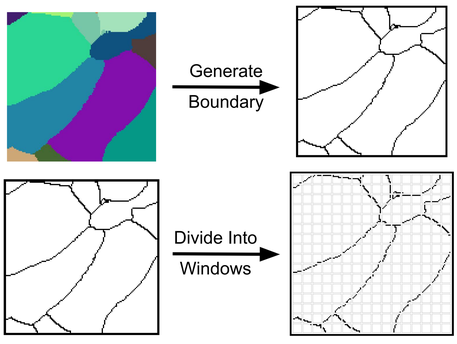
Visual Computing Group

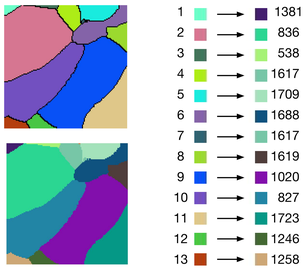




- Connectomics datasets are approaching petabytes in size requiring compression for storage and transmission.
- Automatic reconstruction techniques generate massive quantities of label volumes.
- Compresso reduces this **17.50 terabyte** label volume to **25.94 gigabytes**, a ratio of **675x**.
- Compresso extends to all types of segmentation datasets.

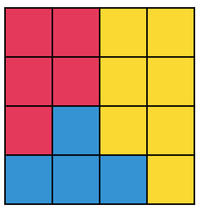
- Compresso works by decoupling the compression of per-segment shapes and per-pixel labels.
- To compress the per-pixel shapes, we generate a boundary map and divide it into 3D congruent windows.

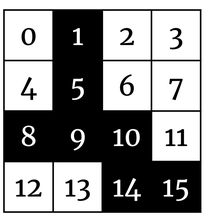




- All of the voxels that are not on a segment boundary have the same label as at least one of their immediate neighbors.
- Rather than store a label for every voxel, we store one label for the entire component enclosed by a contiguous boundary.

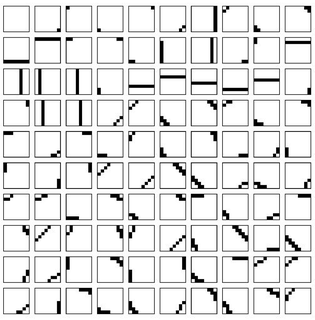
- We encode each window with a single integer value based on the boundary pixels within that window.
- This example window has a value of 50,978.





$$\sum_{i=0}^{n-1} \mathbb{I}(i) 2^i$$

$$2^1 + 2^5 + 2^8 + 2^9 + 2^{10} + 2^{14} + 2^{15} = 50978$$



- Segmentation datasets are highly structured with very few unique boundary patterns over the volume.
- Here are the 100 most common boundary patterns on a typical dataset representing 82% of the volume.
- We use a lookup table to store these identical boundary patterns.

- Compresso followed by a general-purpose compression scheme (e.g. BZ2 or LZMA), outperforms existing methods.
- The principles governing Compresso extend to other types of segmentation datasets, including labeled MRI images.
- Source code is available on Github at **https://github.com/vcg/compresso**.