Decoupling Shape and Texture in SAM-2 via Controlled Texture Replacement Supplemental

Contents

1. Mask Aggregation Evaluation

In order to estimate semantic segmentation performance, disregarding mask fragmentation, we perform predicted mask aggregation before calculating IoU. For each GT region G, we form $U = \bigcup_{i:\, |p_i \cap G|/|p_i| \geq \tau} p_i$ with $\tau = 0.5$ and report $\mathrm{IoU}_{\mathrm{agg}} = \frac{|U \cap G|}{|U \cup G|}$ alongside average IoU which captures well the effect of mask fragmentation characteristic of shape-biased models. Figure 1 illustrates the impact of mask aggregation on segmentation evaluation. The left panel shows the prediction without aggregation, where the model fragments the object (a bear) into numerous small masks. Computing the average IoU over these fragmented predictions yields a low mIoU of 0.0009, heavily penalizing the model despite correct localization. In contrast, the center panel shows the aggregated prediction, where overlapping and redundant masks are merged based on their correspondence to the ground truth. This aggregated result achieves a dramatically higher mIoU of 0.9652, better reflecting the model's ability to segment the object accurately while disregarding fragmentation.

2. CNT visualizations

Figure 2 provides a visualization of the texture transfer method used to create Textured-ADE20K. Figure 3 presents samples from the Textured-ADE20K dataset. Incremental changes in η produce gradual change in texture shift for the resulting

image. For low η values, most of the semantic information in the image is retained. For η = 1, the instances are completely shifted towards the target textures.

3. In The Wild Segmentation Datasets

For this section, we use three datasets:

1. The Potsdam 2D Semantic Labeling Dataset from the ISPRS benchmark [2]. It consists of 38 aerial image tiles (6000×6000 pixels at 5 cm ground resolution), each accompanied by a corresponding digital surface model (DSM) and dense semantic annotations for urban categories including buildings, trees, cars, and impervious surfaces. We use the RGB orthophotos and the available annotated portion of the dataset for evaluation. Potsdam serves as a high-resolution real-world benchmark with detailed structural variation, allowing us to assess how well TextureSAM generalizes to the domain of aerial urban scene segmentation.

2. The UltraHigh Carbon Steel micrograph DataBase (UHCSDB) [1], which contains 961 scanning electron micrographs of ultrahigh carbon steel specimens processed under a wide range of heat treatments. These micro-

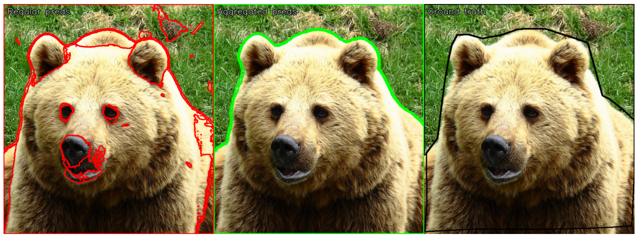


Figure 1. Effect of mask aggregation on segmentation quality. The left panel shows the raw model prediction without aggregation, yielding a very low average IoU of 0.0009 due to excessive fragmentation. The center panel shows the result after aggregating predicted masks, significantly improving alignment with ground truth (right panel) and achieving an mIoU of 0.9652. This highlights the importance of aggregation when evaluating fragmented segmentations.

graphs reflect diverse microstructural constituents such as pearlite, spheroidite, Widmanstätten cementite, and martensite, across multiple magnifications and processing conditions. While UHCSDB was primarily developed as a tool for microstructure informatics and visual exploration, it includes a small manually annotated subset of approximately 23 images with pixel-wise segmentation. We use this subset to test the generalization of TextureSAM to metallographic structures with hierarchical texture complexity.

3. The Dubai Aerial Imagery Dataset, an open-access semantic segmentation dataset published by Humans in the Loop in collaboration with the Mohammed Bin Rashid Space Center. The dataset comprises 72 high-resolution satellite images of Dubai, grouped into six larger tiles, and annotated with pixel-level labels across six semantic classes: building, unpaved land, road, vegetation, water, and unlabeled. The annotations use a standardized colormap, and the images capture the structural and environmental diversity characteristic of urban and desert regions. This dataset provides a compact yet high-quality benchmark for evaluating segmentation performance in satellite imagery under complex real-world conditions involving heterogeneous textures and land cover types.[1]

4. Qualitative segmentation results on the ADE20K validation dataset

Figure 5 shows segmentation results on images from the ADE20K dataset (1st row). It can be seen that TextureSAM (3rd row) produces comparable semantic segmentation to the original SAM-2 (2nd row, 3rd row with modified inference parameters). TextureSAM's predictions align better with the grond-truth (GT), where entire textured regions (e.g. trees, walls.) are marked with the same instance.

5. STMD qualitative results

Figure 4 presents segmentation results on the synthetic STMD dataset. 1st row shows original images, the following rows present segmentation by the different models and the GT annotations. For this semantic-less dataset, TextureSAM segmentation maps better align with GT annotations, while SAM-2 fragments textured regions into individual elements.

6. Number of predicted masks

Figure 6 and Figure 7 show box-plots demonstrating the much lower predicted mask count of Texture-SAM compared with SAM-2, due to SAM-2's over-segmentation of textured regions in RWTD and STMD.

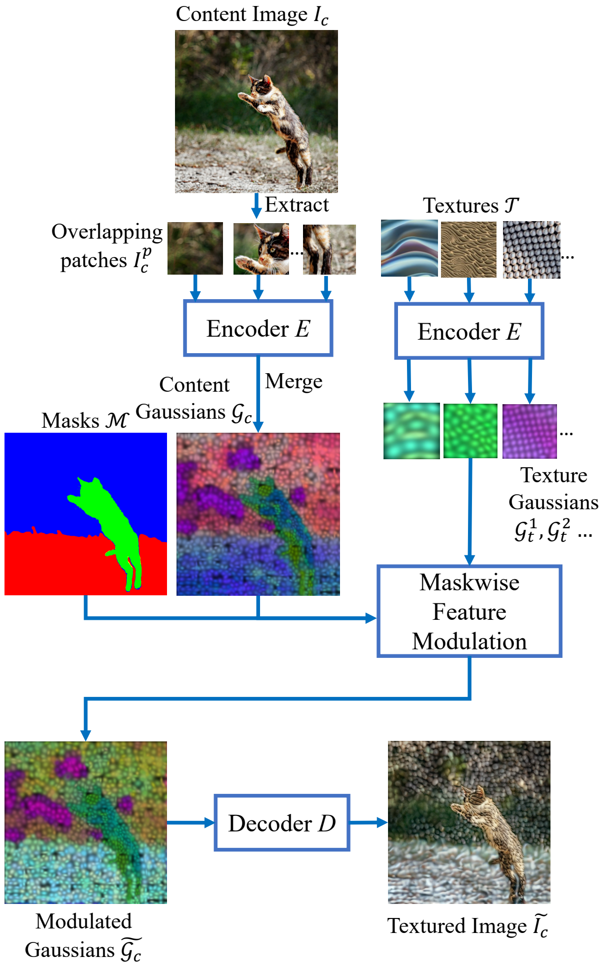

Content Image I_c
Extract
Overlapping patches I_c^p
Textures $\mathcal{T}$
Encoder E
Encoder E
Content Gaussians $\mathcal{G}_c$
Merge
Masks $\mathcal{M}$
Texture Gaussians $\mathcal{G}_t^1, \mathcal{G}_t^2$ ...
Maskwise Feature Modulation
Decoder D
Modulated Gaussians $\widetilde{\mathcal{G}_c}$
Textured Image $\widetilde{I_c}$

Figure 2. Illustration of generating textured image for dataset augmentation using [3].

7. TextureSAM Results on the complete RWTD data

Figure 7 (pages 4-12) compares the segmentation performance of SAM and TextureSAM against ground truth (GT). Each column represents an image alongside its segmentations. As seen, TextureSAM is better adapted to capturing textures, while SAM prioritizes semantic segmentation, leading to more distinct texture-aware regions in TextureSAM's results.

[1] https://www.kaggle.com/datasets/humansintheloop/semantic-segmentation-of-aerial-imagery

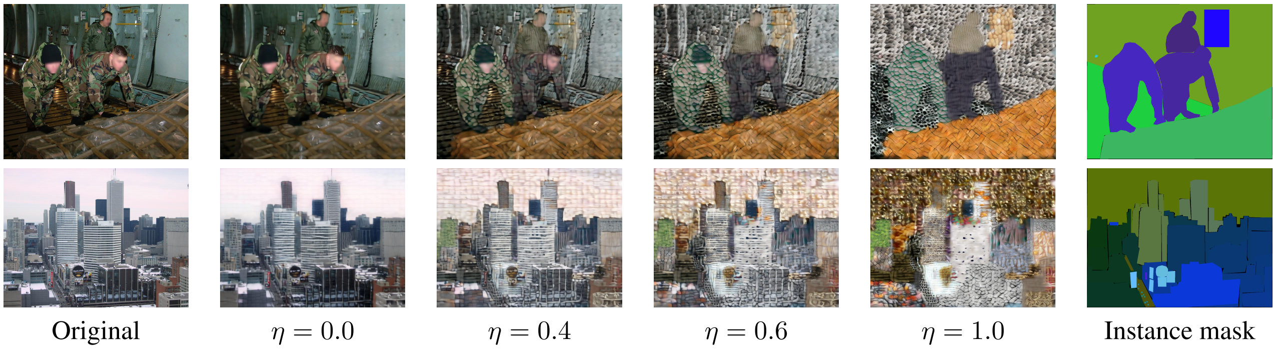


Figure 3. **Samples of Textured-ADE20K dataset.** Incremental changes in η produce gradual change in texture shift for the resulting image. For low η values most of the semantic information in the image is retained. For $\eta = 1$ the instances are completely shifted towards the target textures.

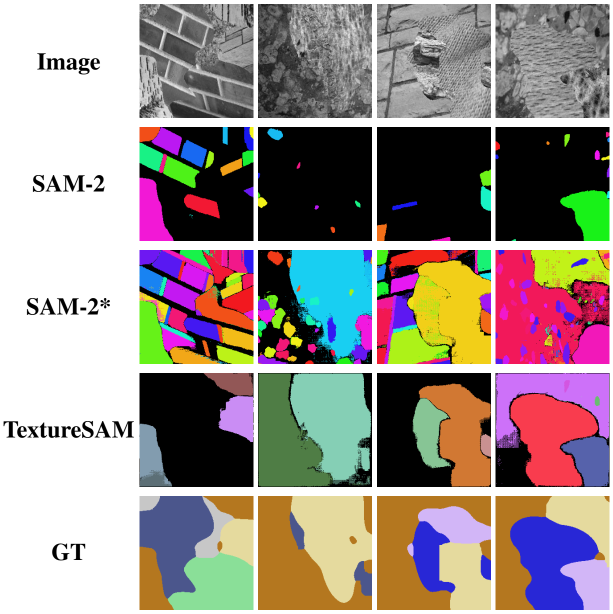


Figure 4. Segmentation results on the synthetic STMD dataset. 1st row shows original images, the following rows present segmentation by the different models and the GT annotations. For this semantic-less dataset, TextureSAM segmentation maps better align with GT annotations, while SAM-2 fragments textured regions into individual elements.

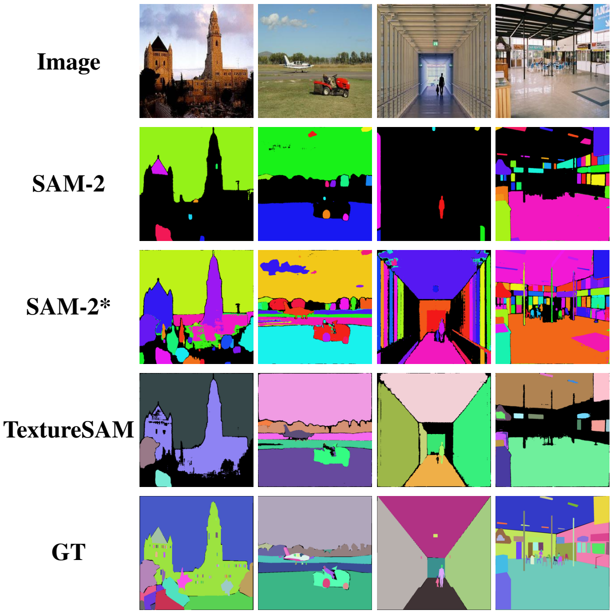


Figure 5. Segmentation results on images from the ADE20K dataset (1st row). It can be seen that TextureSAM (3rd row) produces comparable semantic segmentation to the original SAM-2 (2nd row, 3rd row with modified inference parameters). TextureSAM's predictions align better with the GT, where entire textured regions (e.g. trees, walls.) are marked with the same instance.

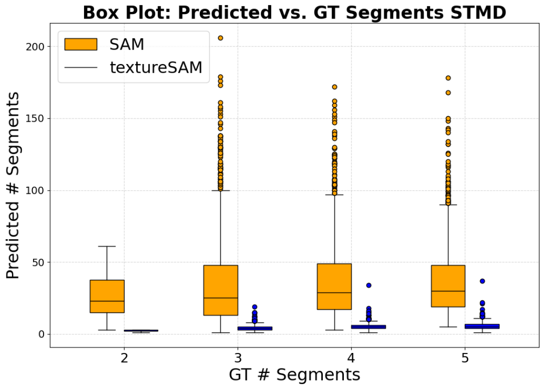


Figure 6. Box plot comparing predicted segments to the ground truth (GT) for the Synthetic Textured Masks Dataset (STMD). We group the results by the number of GT segments per image. SAM-2's fragmentation of textures can be seen in the plot as it generates significantly more masks (segments).

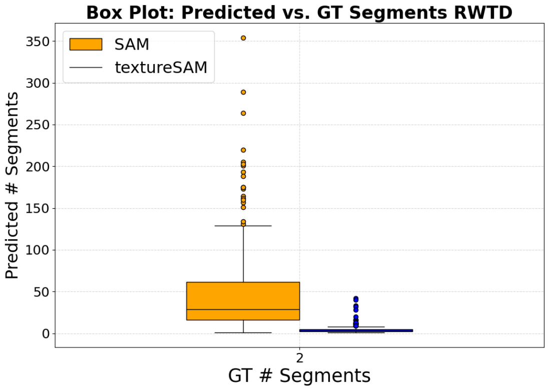


Figure 7. Box plot comparing predicted segments to the ground truth (GT) for the Real World Textured Dataset (RWTD). Each annotation contains two GT segments. SAM2's over-segmentation is evident by it producing significantly more masks compared to TextureSAM.

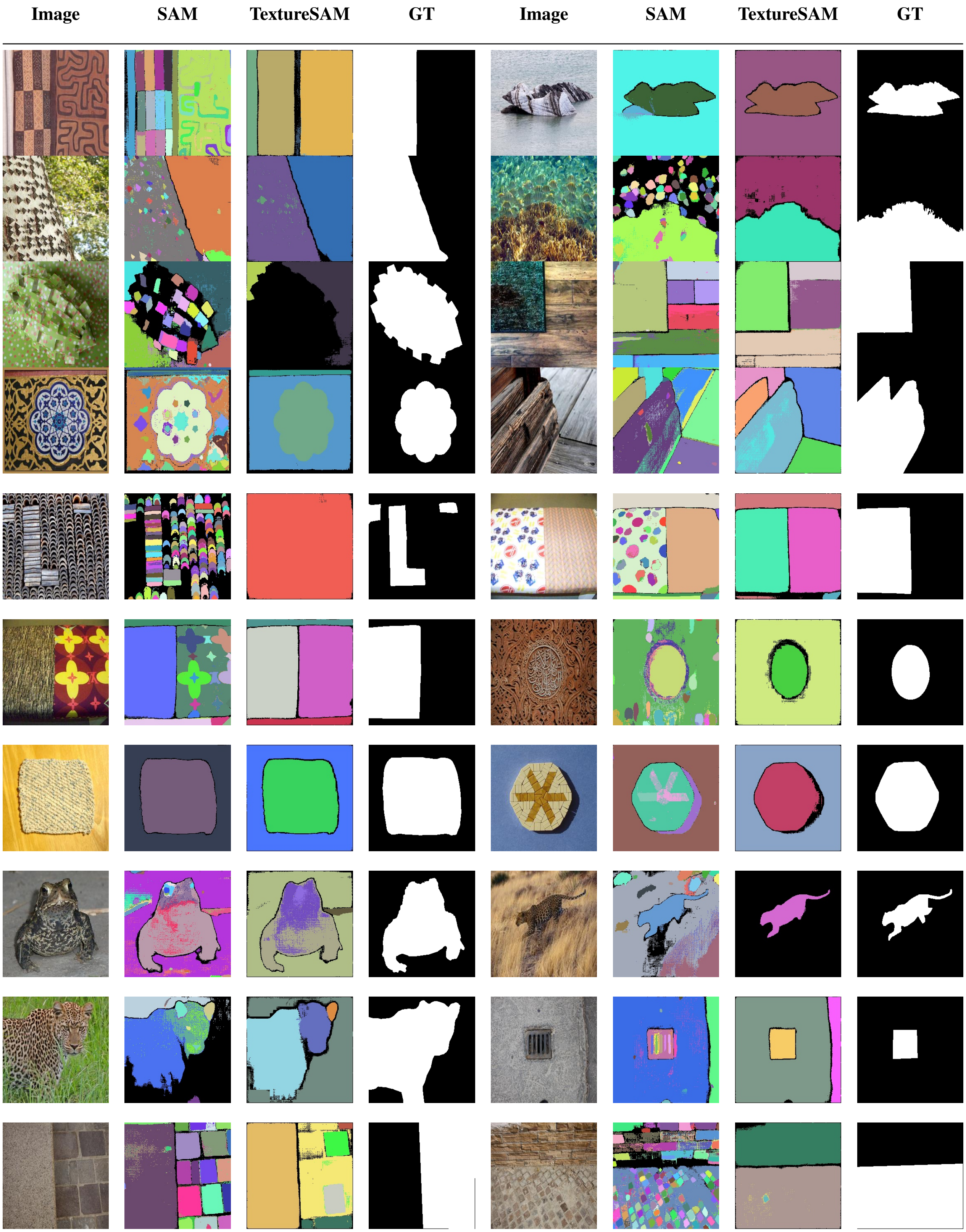


Image
SAM
TextureSAM
GT
Image
SAM
TextureSAM
GT

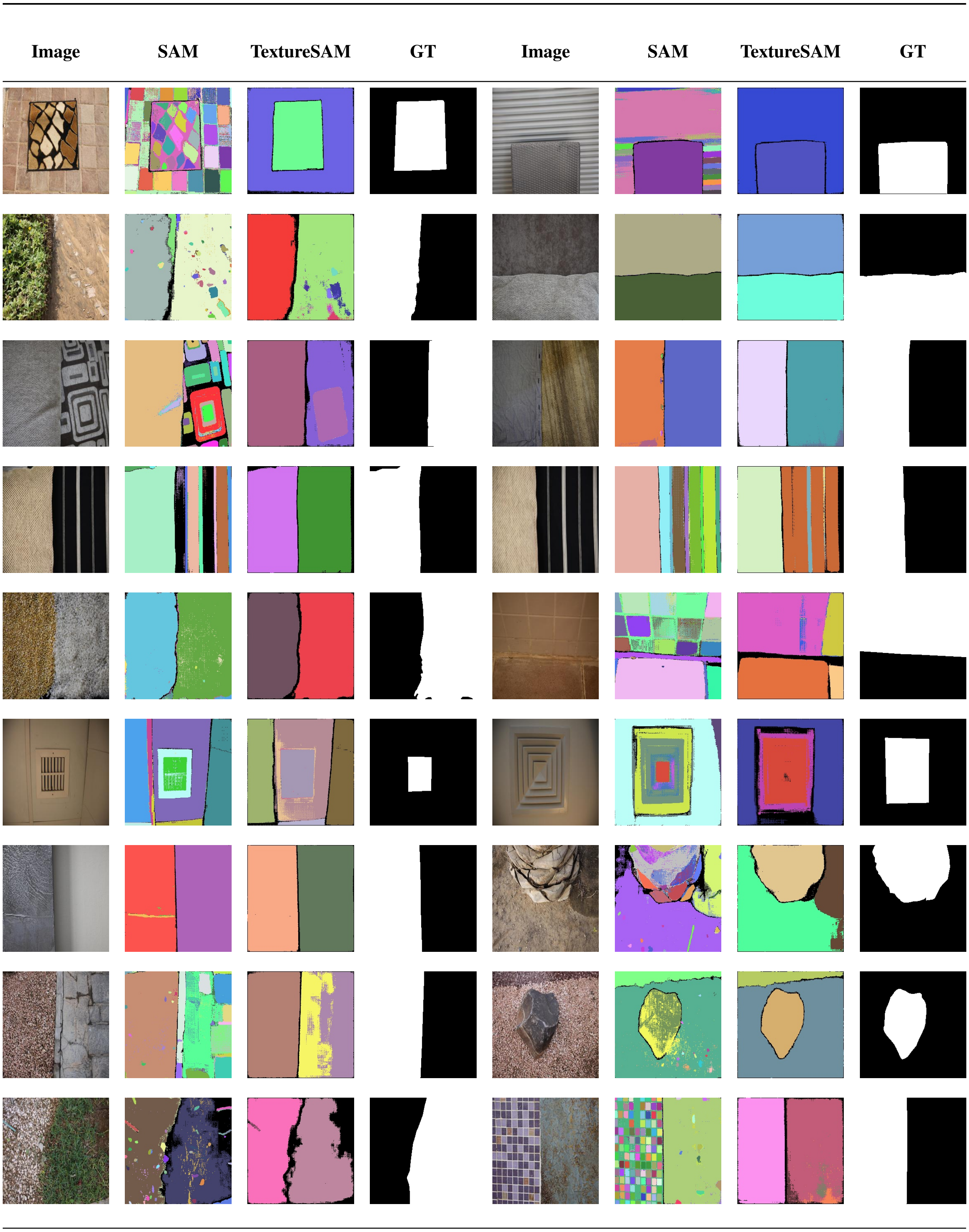


Image
SAM
TextureSAM
GT
Image
SAM
TextureSAM
GT

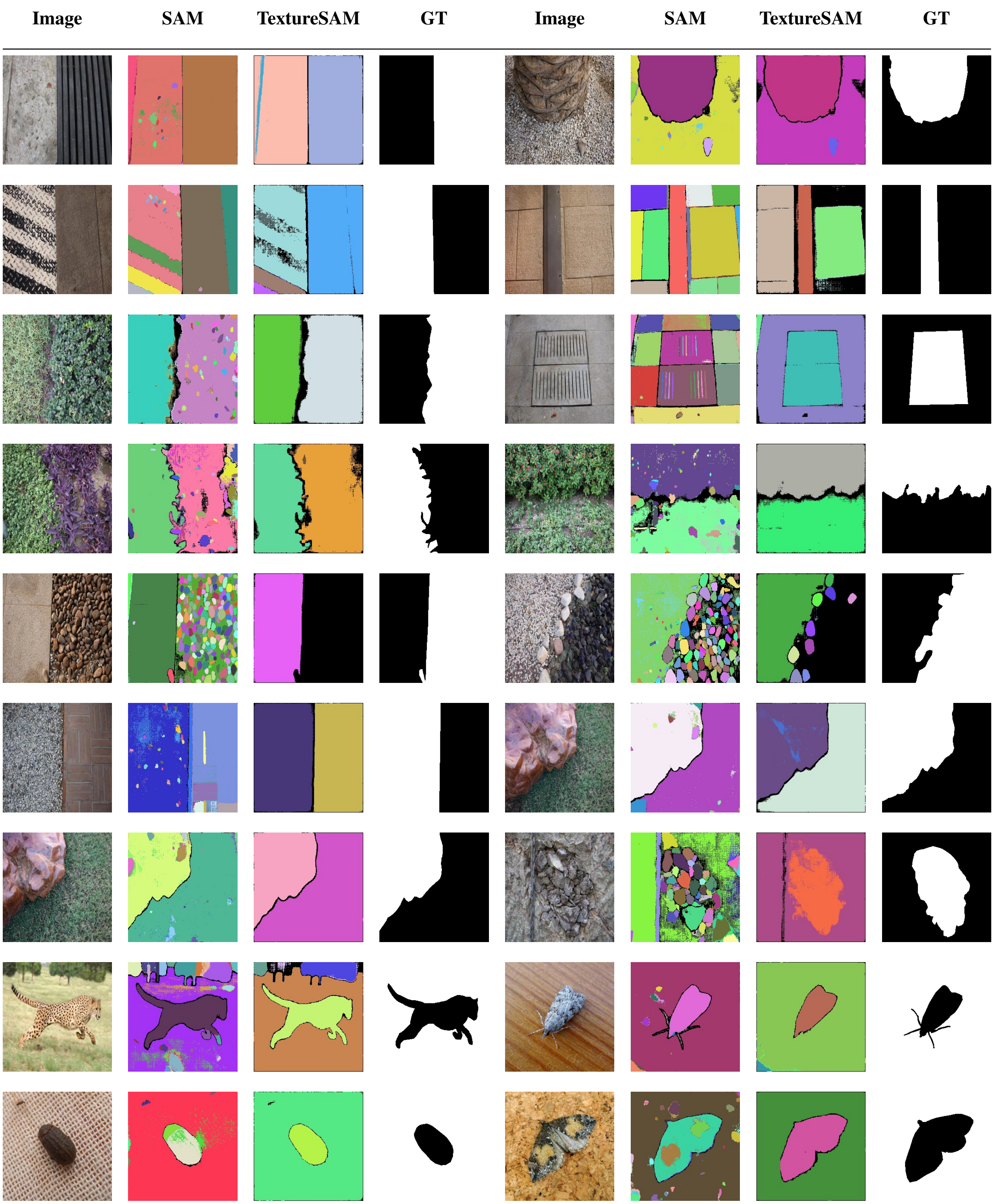


Image
SAM
TextureSAM
GT
Image
SAM
TextureSAM
GT

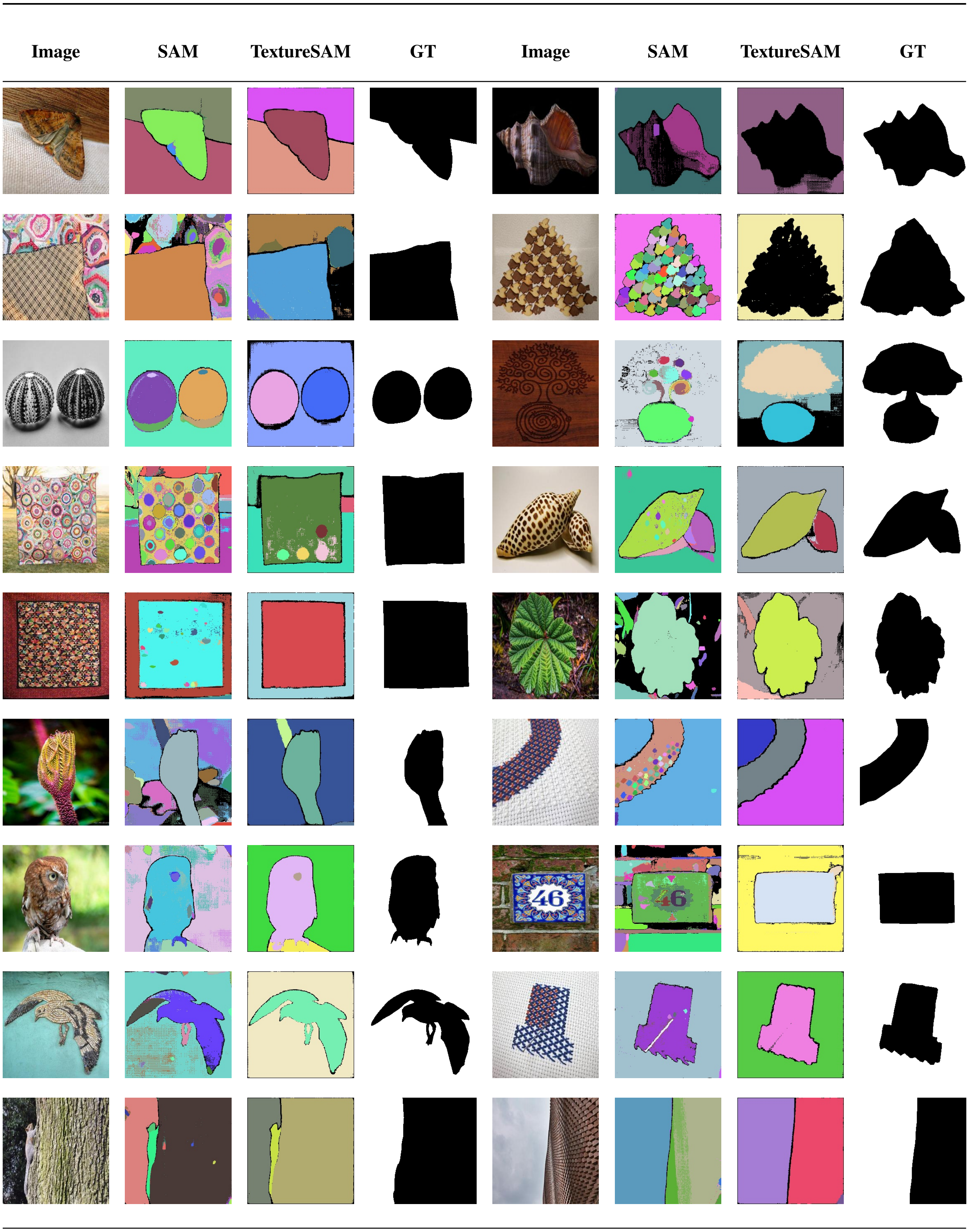


Image
SAM
TextureSAM
GT
Image
SAM
TextureSAM
GT

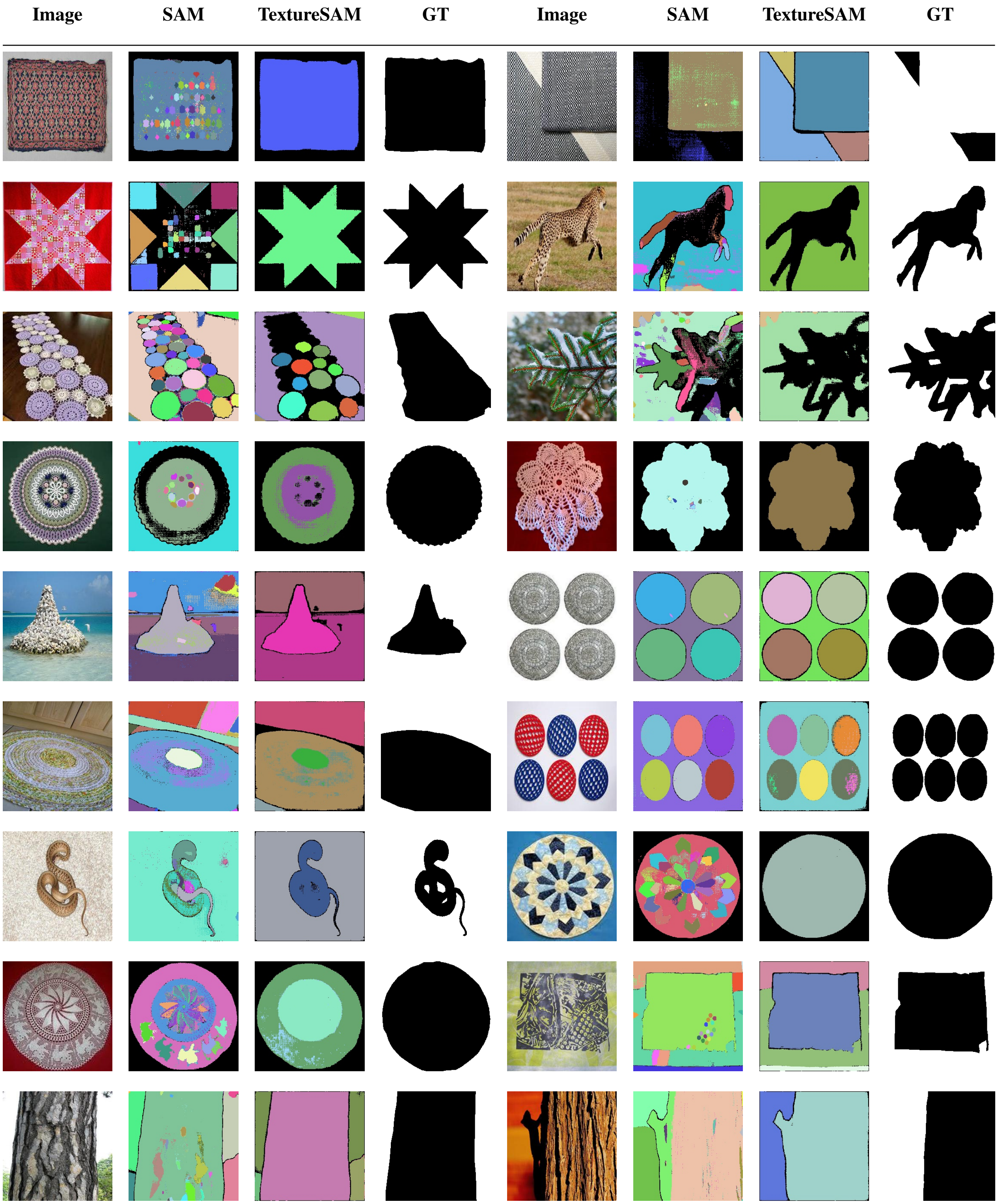


Image
SAM
TextureSAM
GT
Image
SAM
TextureSAM
GT

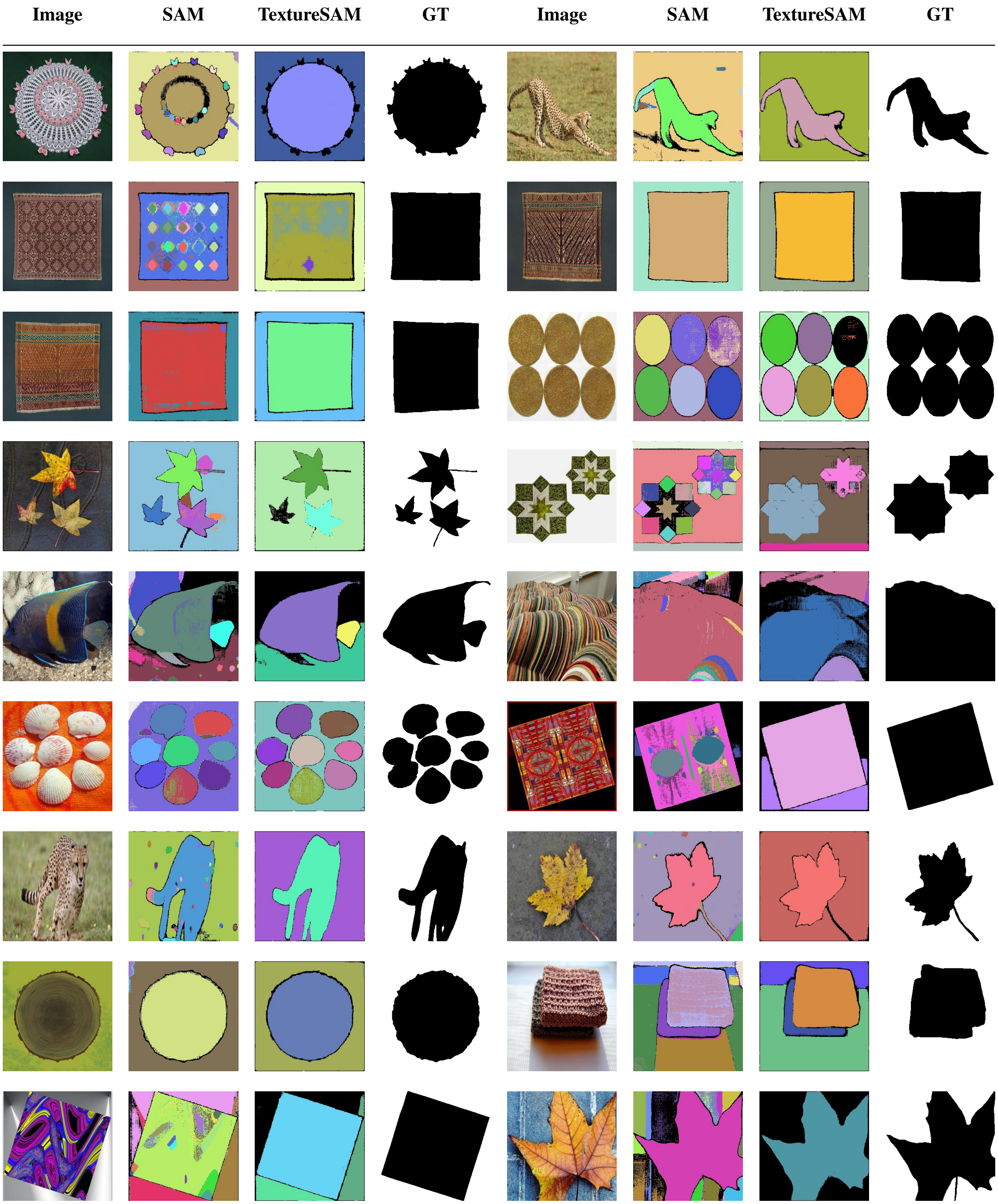


Image
SAM
TextureSAM
GT
Image
SAM
TextureSAM
GT

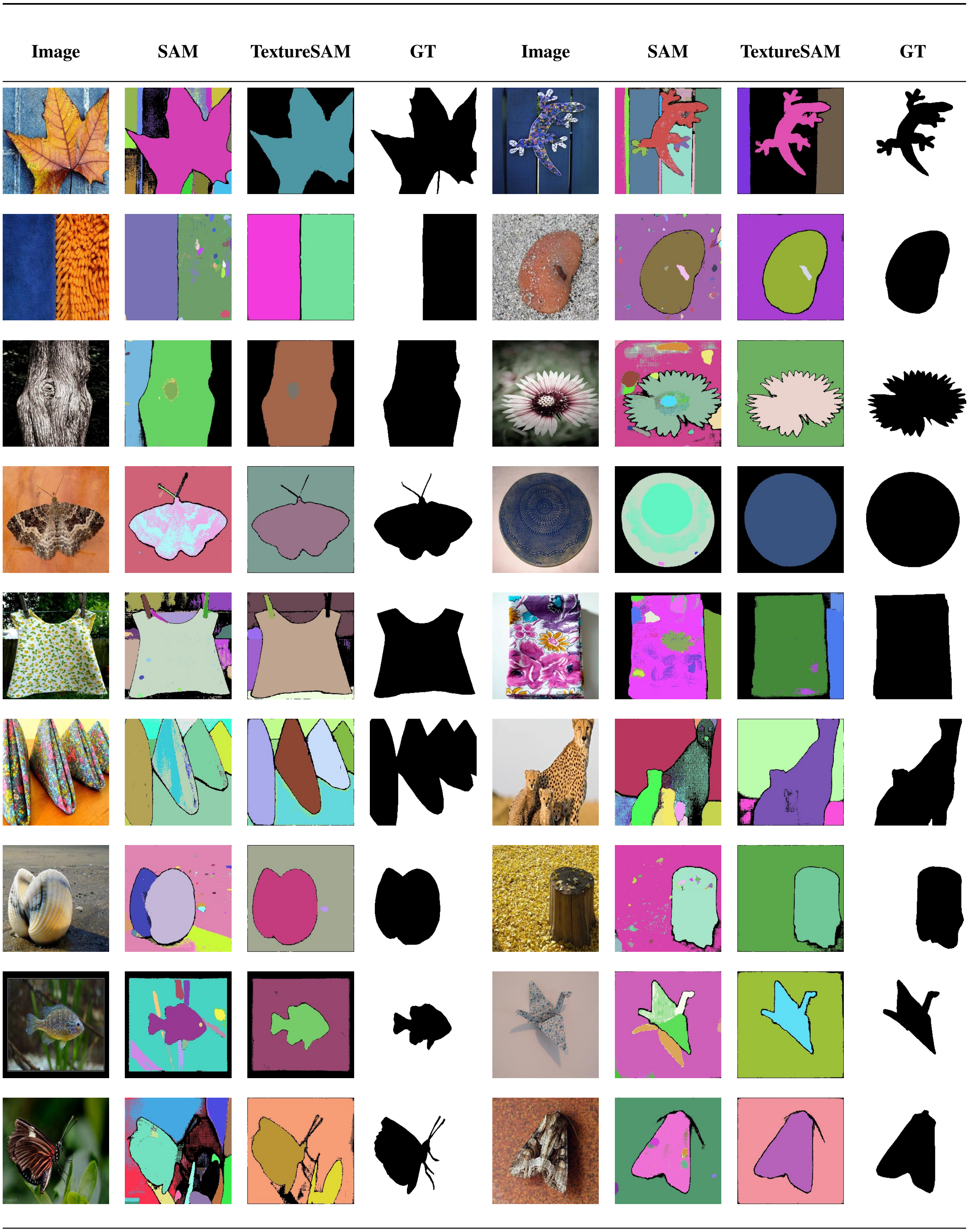


Image
SAM
TextureSAM
GT
Image
SAM
TextureSAM
GT

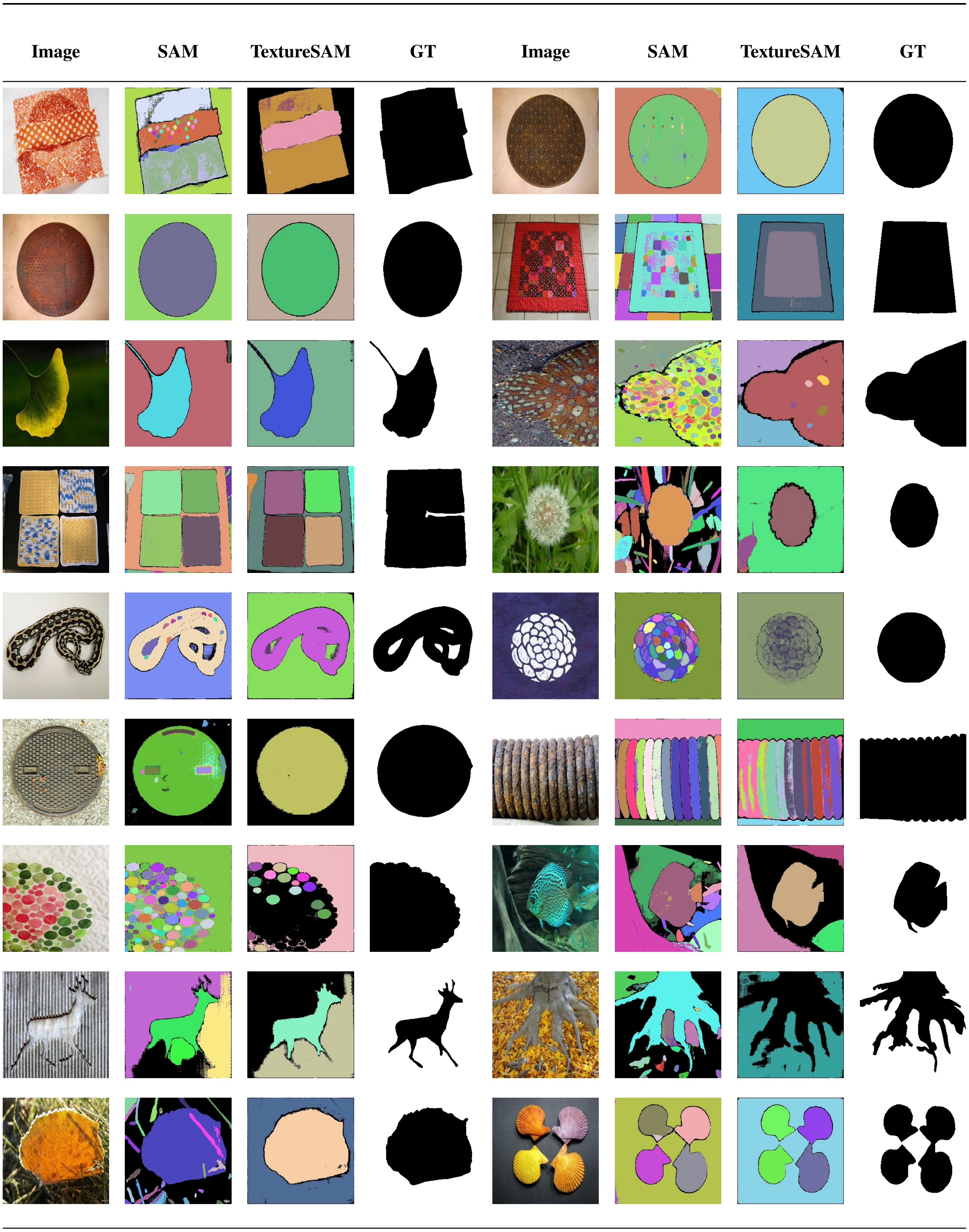
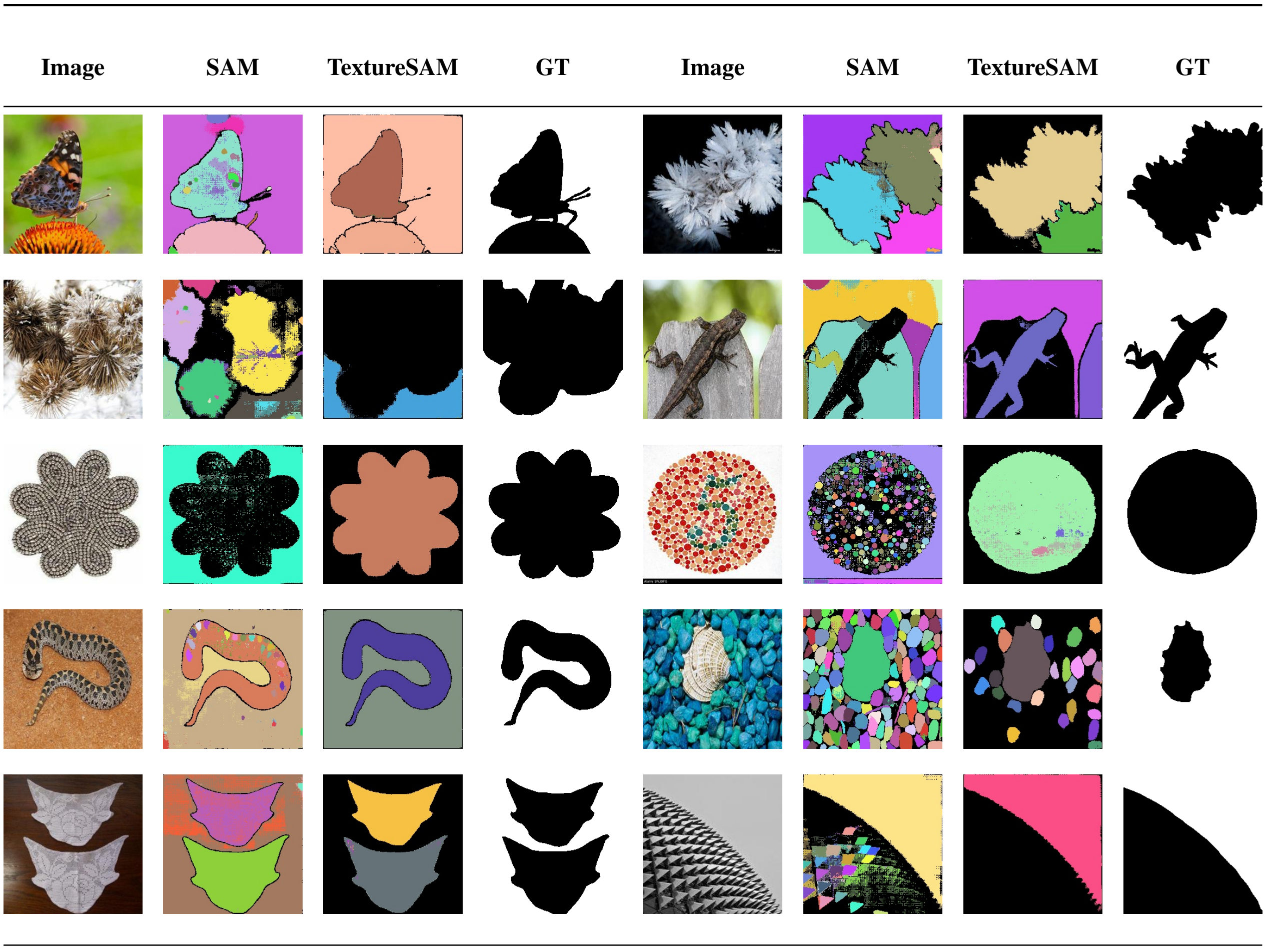


Image
SAM
TextureSAM
GT
Image
SAM
TextureSAM
GT



References

[1] Brian L DeCost, Matthew D Hecht, Toby Francis, Bryan A Webler, Yoosuf N Picard, and Elizabeth A Holm. Uhcsdb: ultrahigh carbon steel micrograph database: tools for exploring large heterogeneous microstructure datasets. *Integrating Materials and Manufacturing Innovation*, 6:197–205, 2017. 1

[2] Franz Rottensteiner, Gunho Sohn, Markus Gerke, and Jan D Wegner. Isprs semantic labeling contest. *ISPRS: Leopoldshöhe, Germany*, 1(4):4, 2014. 1

[3] Peihan Tu, Li-Yi Wei, and Matthias Zwicker. Compositional neural textures. In *SIGGRAPH Asia 2024 Conference Papers*, SA '24, New York, NY, USA, 2024. Association for Computing Machinery. 2