Structure-Aware Feature Rectification with Region Adjacency Graphs for Training-Free Open-Vocabulary Semantic Segmentation (Supplementary Materials)

Qiming Huang, Hao Ai, Jianbo Jiao
The MIx Group, School of Computer Science
University of Birmingham
{qxh366, hxa456}@student.bham.ac.uk, j.jiao@bham.ac.uk

S1. More ablation study for hyperparameters

To investigate the impact of key hyperparameters on our model's performance, we conducted a series of ablation studies. The experiments focused on the parameters of the Simple Linear Iterative Clustering (SLIC) algorithm and the selection of features from the Grey-Level Co-occurrence Matrix (GLCM).

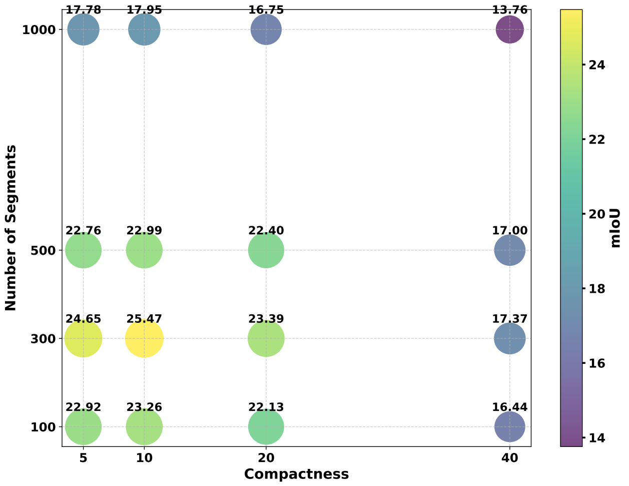


Figure S1. Ablation study on the number of segments (n_segments) and compactness for the SLIC algorithm. Results are reported as mIoU on the COCO-Stuff-171 validation set. The baseline model is NACLIP.

We performed a grid search to optimise the number of segments and the compactness for the SLIC algorithm, with the quantitative results shown in Fig. S1. Different hyperparameters of SLIC produce different region proposal results, as visualised in Fig. S3. The number of segments directly controls the scale of the superpixels; increasing this value results in finer, more numerous regions. The compactness parameter manages the trade-off between spatial proximity and colour similarity; a lower value allows superpixels to conform more closely to image textures and edges, while a higher value produces more uniform, regularly shaped regions. Our analysis indicates that the optimal balance for this task was achieved with 300 segments and a compactness of 10, yielding a peak mIoU of 25.47.

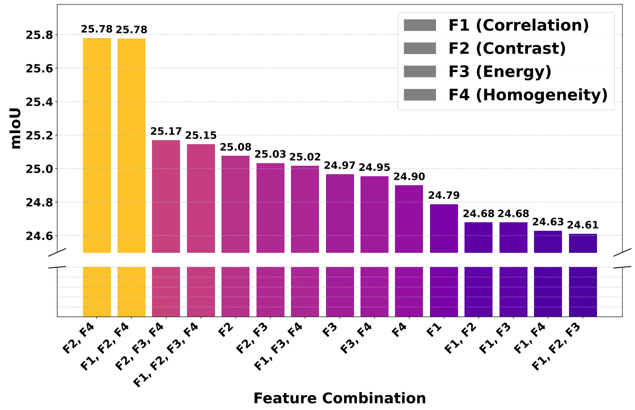


Figure S2. The impact of different SLIC feature combinations on the performance of the NACLIP baseline. The experiment was conducted on the COCO-Stuff-171 validation set with baseline NACLIP.

Furthermore, we analysed the contribution of different combinations of four GLCM texture features: Correlation (F1), Contrast (F2), Energy (F3), and Homogeneity (F4) as suggested in [1]. The results, presented in Fig. S2, reveal that the combination of Contrast (F2) and Homogeneity (F4) yielded the highest mIoU of 25.78. Notably, this two-feature subset outperformed the combination of all four features, highlighting that an appropriate selection of features is more effective than using them all.

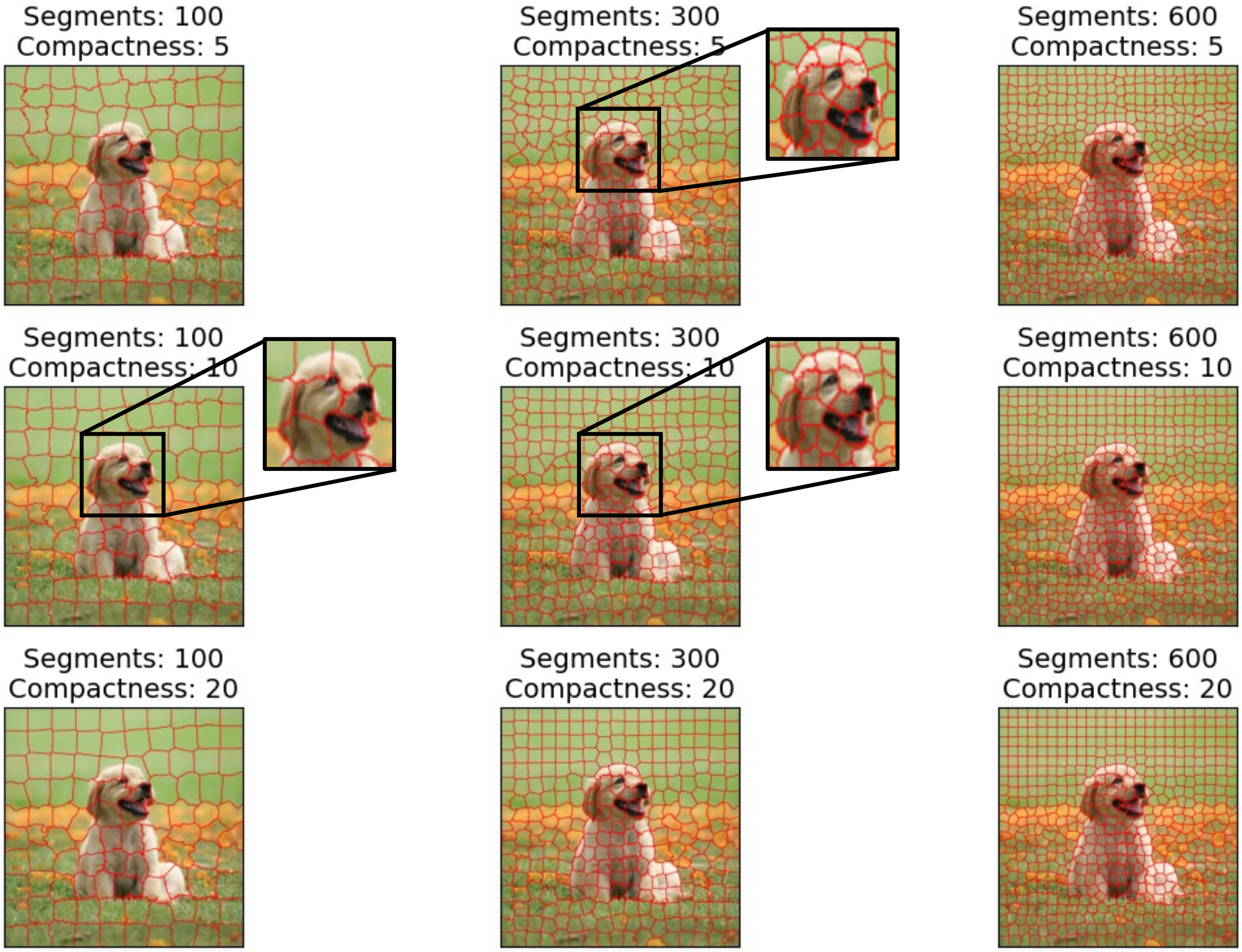


Figure S3. Comparison of SLIC segmentation results across different combinations of Segments (number of superpixels) and Compactness.

S2. Generalisation analysis

To further evaluate the generalisation capabilities of our proposed method, we conduct a series of analyses under various challenging conditions, including common image corruptions, domain shifts, and zero-shot segmentation on extra domain-related remote sensing datasets postdam[1].

Robustness to common image corruptions. We first assess the model's resilience to common visual perturbations that degrade image quality. Table S1 quantitatively measures the performance changes of our method built upon NACLIP baseline under four conditions: overexposure, underexposure, grayscale conversion, and texture destruction (via Gaussian blur). For overexposure, we used a brightness factor of 1.8 to significantly increase the luminosity, causing highlights to become washed out. For underexposure, we used a brightness factor of 0.4 to decrease the luminosity, obscuring details in shadows. For texture destruction, we used a kernel size of $(9, 9)$ pixels and $\sigma = 5$ to create a significant and noticeable blurring effect that effectively destroys surface textures. Our method maintains reasonable performance in most cases, but significantly reduces when facing strong low exposure.

To better understand this, we provide a visualisation of the segmentation results under these conditions. **Effect of Lighting.** As shown in Fig. S4, our model performs well under extreme lighting changes. Despite significant information loss in the bright, washed-out areas of overexposed images or the dark, detail-lacking regions of underexposed images, our model consistently generates reasonable segmentation masks for objects like 'doughnuts', 'zebras'.

[1] Isprs potsdam dataset on kaggle. Available online: https://www.kaggle.com/datasets/jahidhasan66/isprs-potsdam

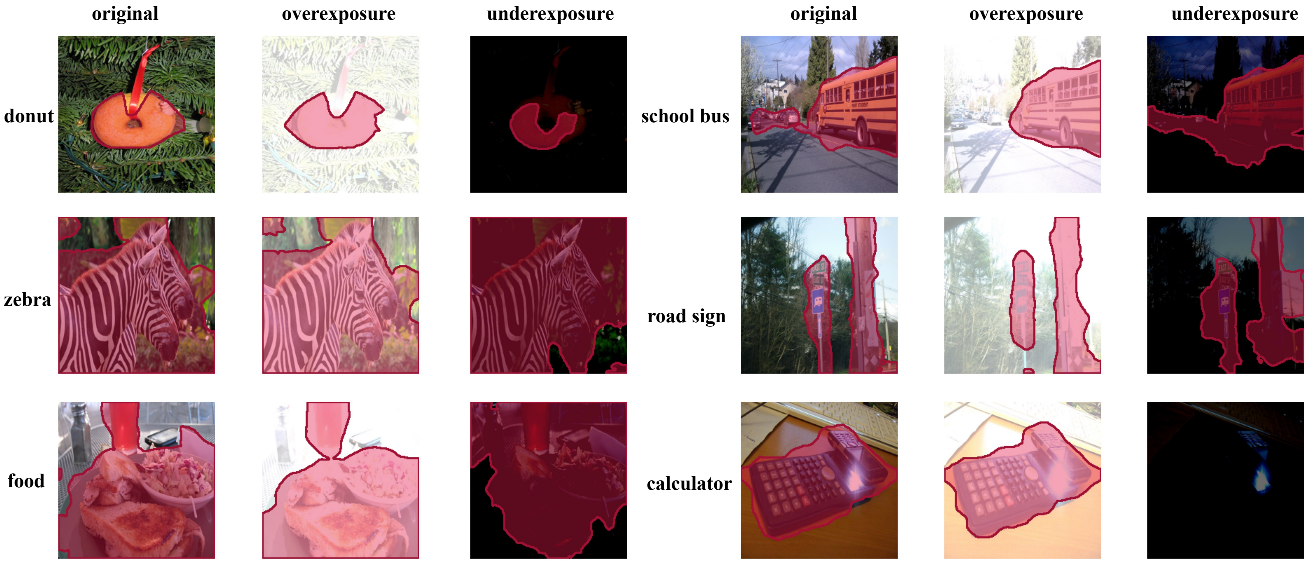


Figure S4. The effect of overexposure and underexposure on segmentation performance.

Table S1. The performance drops of our method under different cases. We use NACLIP as the baseline.

	V20	Stuff	PC59	ADE
overexposure	-2.1	-1.8	-2.5	-1.5
underexposure	-8.5	-10.8	-7.5	-4.5
grayscale	-1.1	-1.5	-2.1	-1.2
texture destruction	-1.8	-2.3	-2.2	-1.8

Table S2. Comparison of Zero-Shot Performance on the Potsdam Remote Sensing Datasets. Results are reported on mIoU

	Potsdam
NACLIP	28.6
NACLIP + Ours	30.4
ClipSurgery	30.2
ClipSurgery + Ours	32.1

Effect of texture destruction. To simulate the loss of fine-grained details and high-frequency textures, we applied a Gaussian blur filter. This process involves convolving the image with a Gaussian kernel. The intensity of the blur is controlled by the kernel size and the standard deviation (σ). In our experiments, we used a kernel size of $(9, 9)$ pixels and $\sigma = 5$ to create a significant and noticeable blurring effect that effectively destroys surface textures. The visualisation results are shown in Fig. S6.

Analysis of domain shift. We further investigate the model's ability to generalise across different visual domains, a critical aspect of real-world applications. Fig. S7 showcases the segmentation performance on images that have undergone significant style and domain shifts. We test on artistic renderings (*e.g.* oil painting), images with altered colour schemes (grayscale vs. coloured), and various other style-transferred examples. The model consistently produces precise segmentations for objects like ’dogs’, ’pineapples’, and ’boats’ across these diverse visual styles.

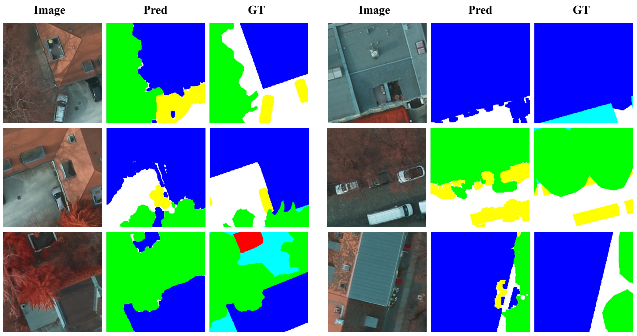


Figure S5. Visualisation results on the Potsdam dataset of our method built upon NACLIP.

Zero-Shot generalisation to Postdam remote sensing dataset. Additionally, we evaluate our method's zero-shot performance on a completely unseen and specialised domain: the Potsdam remote sensing dataset. As reported in Table S2, when our module is integrated with existing baselines (NACLIP and ClipSurgery), it yields substantial improvements in mean Intersection over Union (mIoU). The Visualisation result is shown in Fig. S5.

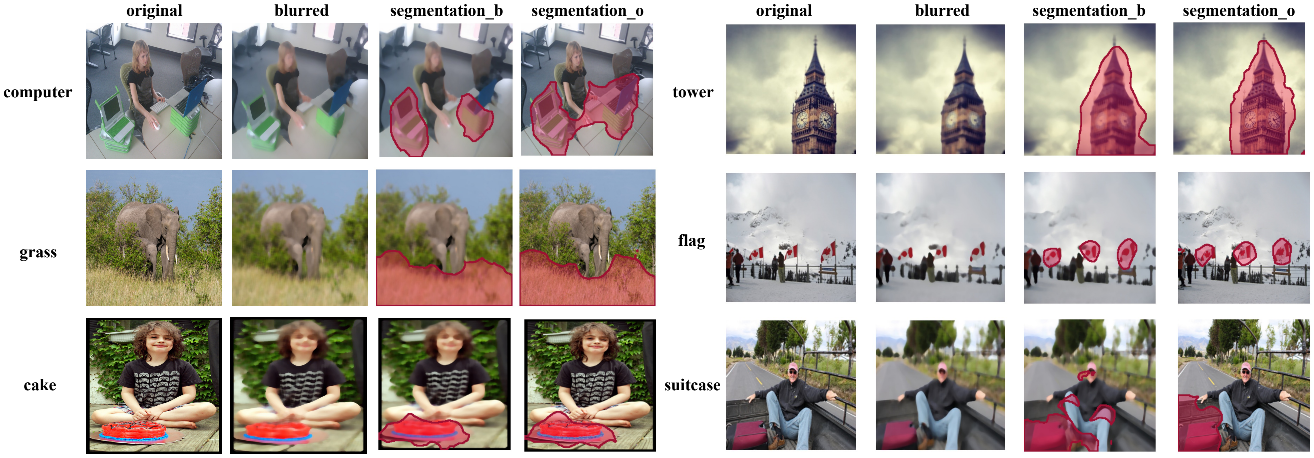


Figure S6. The effect of texture destruction (Gaussian blur) on segmentation performance.

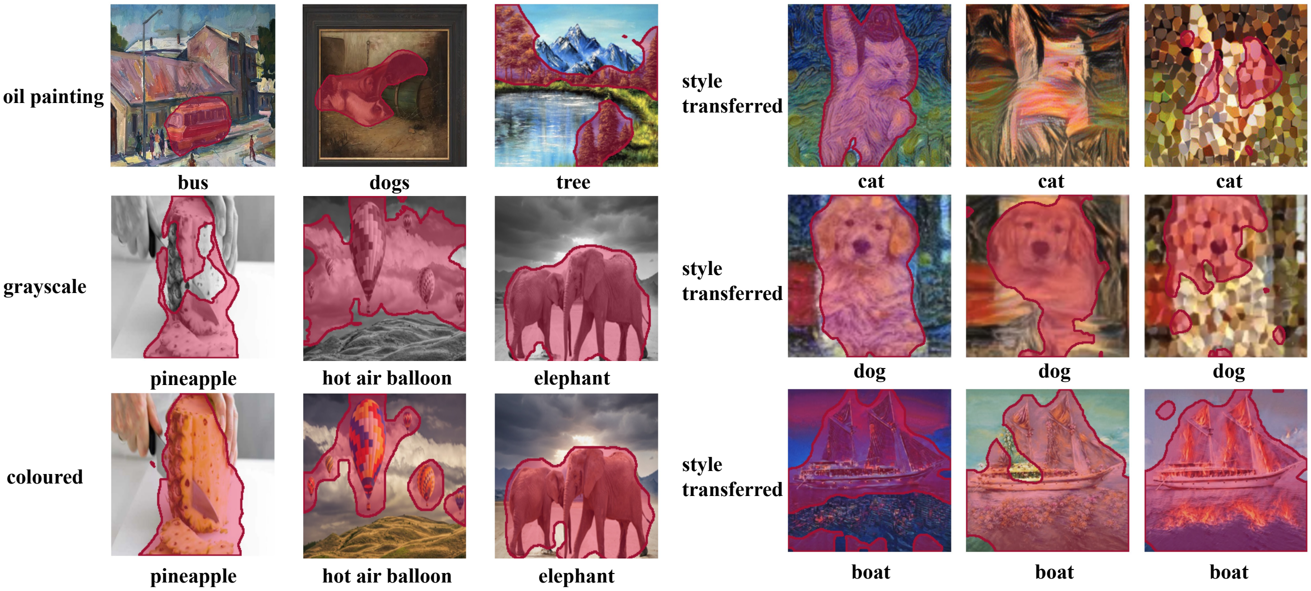
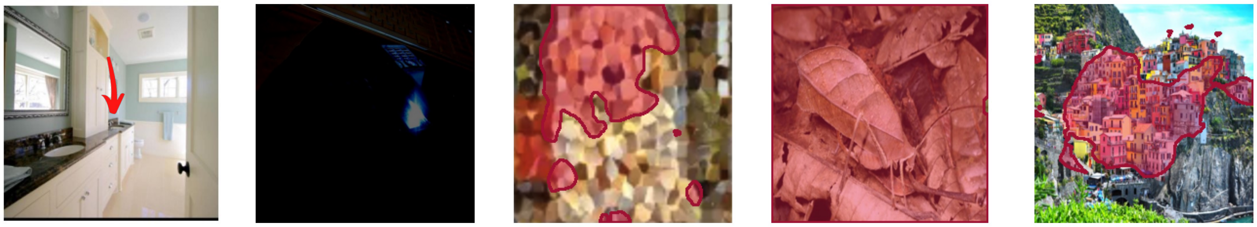


Figure S7. The effect of domain shift on segmentation performance.



small object: faucet **underexposure** **colour texture disorder** **camouflaged objects** **excessive colour complexity**

Figure S8. Illustration of challenging scenarios leading to segmentation failures. The cases, from left to right, include: (a) an object that is too small to be accurately detected (faucet); (b) severe underexposure resulting in loss of detail; (c) disordered color and texture, making boundaries ambiguous; (d) an object camouflaged against a similar background; and (e) a scene with excessive colour complexity and numerous small details.

S3. Analysis of failure cases

We analysed the failure cases of our method to better understand its limitations. Fig. S8 illustrates five representative scenarios where the segmentation performance is compromised: **Small Object Insensitivity.** Our approach relies on an initial superpixel segmentation. Consequently, objects that are exceptionally small, such as the 'faucet', may be smaller than the generated superpixels and are incorrectly absorbed into larger background regions. This prevents them from being represented as distinct nodes in the region adjacency graph. **Extreme Lighting Conditions.** In cases of severe underexposure, the lack of sufficient colour and brightness information cripples the feature extraction process. Both SLIC and GLCM features become unreliable, leading to a near-complete failure to identify any objects. **Ambiguous Boundaries and Camouflage.** The model's performance degrades when there is no clear distinction between foreground and background. This occurs in scenes with chaotic colour and texture, where boundaries are inherently ambiguous, and in cases of camouflage, where the object's texture features are nearly identical to the background's. **Excessive Scene Complexity.** Our method can be challenged by scenes containing an overwhelming density of small, intricate details. The high frequency of colour and texture changes results in an overly fragmented superpixel map and a highly complex region graph, which hinders the effective propagation and feature rectification.

S4. Visualisation comparison with ClipSurgery

Fig. S9 visualises the qualitative impact of our method when applied to ClipSurgery. It is evident that our approach refines the model's attention mechanism. The baseline ClipSurgery model, while effective, often produces coarse and noisy attention maps that fail to precisely localise the target object (*e.g.*'bus', 'grass'). By incorporating our structure-aware feature rectification using region adjacency graphs, the resulting attention becomes more focused and clean. Our method successfully prunes background noise and sharpens the activation to align with true object boundaries.

S5. Combining SAM with our method

The construction of the Region Adjacency Graph (RAG) is critical to our method's success. A natural consideration was to leverage powerful segmentation models like SAM to generate the graph's nodes. However, as Fig. S10 reveals, this approach introduces a significant topological problem. SAM masks, while semantically meaningful, vary dramatically in size. This leads to the formation of a highly centralised RAG, where massive background regions become "hub nodes" that connect to a vast number of smaller regions (column 3). Such a structure is unstable for feature propagation, as these hubs can wash out important local details. To solve this, our method instead employs a superpixel-based tessellation of the image (column 4). This approach guarantees a granular and uniform partitioning, resulting in a balanced and regular RAG. Each node maintains local connectivity with a consistent number of neighbours, providing a stable and unbiased foundation for the structure-aware feature rectification.

References

[1] Florentina Tatrin Kurniati, Daniel HF Manongga, Eko Sediyono, Sri Yulianto Joko Prasetyo, and Roy Rudolf Huizen. Glcm-based feature combination for extraction model optimization in object detection using machine learning. *arXiv preprint arXiv:2404.04578*, 2024. 1

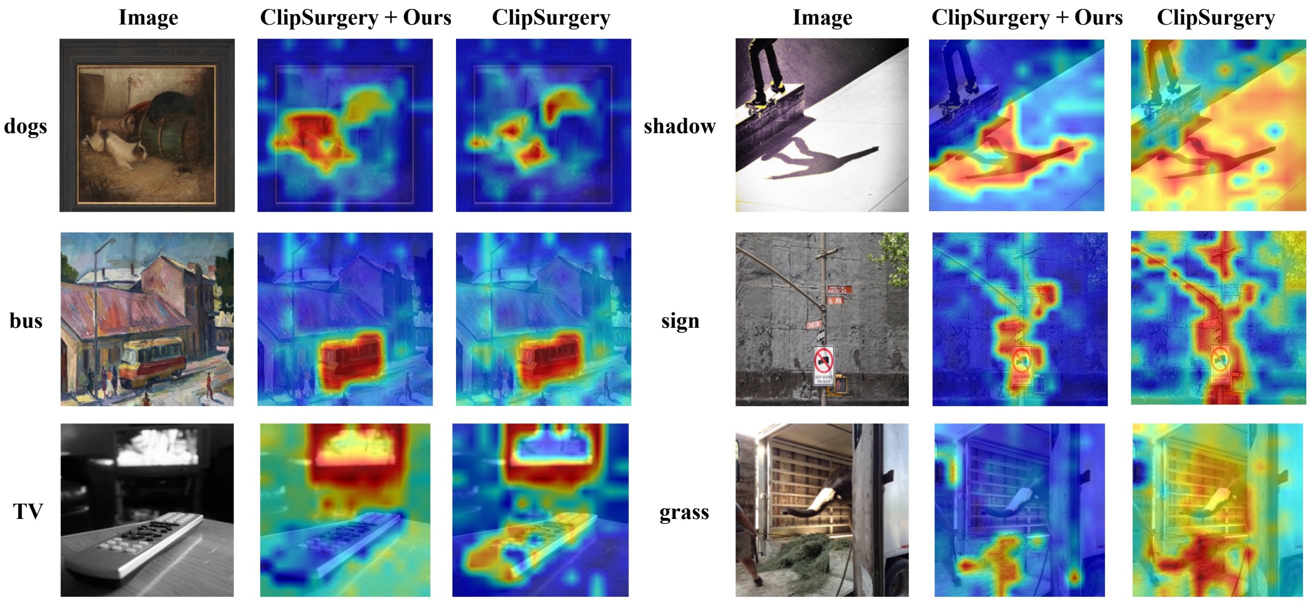


Figure S9. Comparison of model attention visualisation results with ClipSurgery and ClipSurgery + Ours

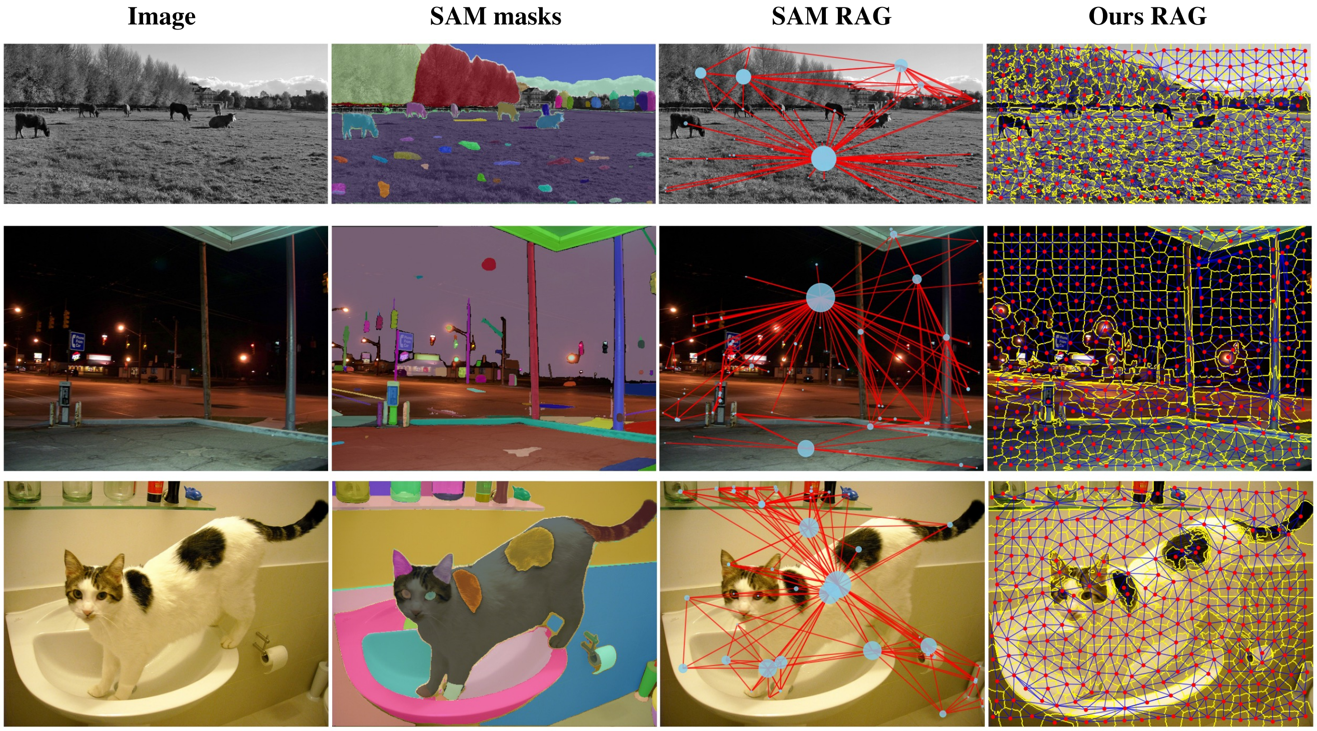


Figure S10. Comparison of RAG construction methods. While SAM masks (column 2) provide semantic regions, their imbalanced sizes create a problematic RAG (column 3). Large masks become hub nodes with numerous connections, dominating inter-region calculations; in column 3, larger circles indicate a greater number of neighbours. Our approach uses superpixels (column 4) to create a RAG with uniformly sized regions and a consistent number of neighbours, ensuring a more stable and reliable graph structure.