HiSS: Human-inspired Semantic Segmentation for Vehicle Interior Scene Understanding

Aleksander Kostuch
AGH University of Krakow
Aptiv Technical Center Krakow
kostuch@agh.edu.pl

Joanna Jaworek-Korjakowska
AGH University of Krakow
Center of Excellence in Artificial Intelligence (CEAI)
jaworek@agh.edu.pl

1. Supplementary Materials

1.1. Multi-step Approach Mathematical Notation

Let the input image be defined as:

$$I \in \mathbb{R}^{H \times W \times C} \tag{1}$$

where H and W are the spatial dimensions, and $C = 1$ for IR image or $C = 3$ for RGB channels.

Stage 1: Focus Mask Generation. We define a binary segmentation model or a detection model:

$$S_b : \mathbb{R}^{H \times W \times C} \to [0, 1]^{H \times W} \tag{2}$$

which produces a soft focus mask:

$$M = S_b(I) \tag{3}$$

This mask highlights regions of interest and suppresses irrelevant areas, mimicking human attention.

Stage 2: Focused Semantic Segmentation. We apply the mask M to the input image via channel-wise multiplication:

$$\tilde{I} = I \odot M \tag{4}$$

where $\odot$ denotes element-wise multiplication with broadcasting across channels.

Then, the modified image $\tilde{I}$ is fed into a multiclass semantic segmentation model:

$$S_m : \mathbb{R}^{H \times W \times C} \to \mathbb{R}^{H \times W \times K} \tag{5}$$

producing the final segmentation map:

$$Y = S_m(\tilde{I}) \tag{6}$$

where K is the number of semantic classes.

The focus mask can be constructed using classical methods, such as:

$$M_G = G(M) \quad \text{(Gaussian mask)} \tag{7}$$

$$M_D = D(M) \quad \text{(Distance-transform mask)} \tag{8}$$

Final Formulation. The complete segmentation pipeline can be compactly expressed as:

$$Y = S_m \left(I \odot S_b(I) \right) \tag{9}$$

or, in other variants:

$$Y_G = S_m \left(I \odot G(S_b(I)) \right) \tag{10}$$

$$Y_D = S_m \left(I \odot D(S_b(I)) \right) \tag{11}$$

1.2. Segmentation Metrics

Segmentation tasks require pixel-level evaluation metrics to assess the fine-grained understanding of cabin interiors:

Pixel Accuracy The proportion of correctly classified pixels across all classes:

$$\text{Pixel Accuracy} = \frac{\sum_{i=0}^{k} p_{ii}}{\sum_{i=0}^{k} \sum_{j=0}^{k} p_{ij}} \tag{12}$$

where p_{ij} is the number of pixels of class i predicted as class j, and k is the number of classes.

Mean Intersection over Union (mIoU) The IoU averaged over all classes:

$$\text{mIoU} = \frac{1}{k+1} \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}} \tag{13}$$

1.3. Solution diagram extended with an example

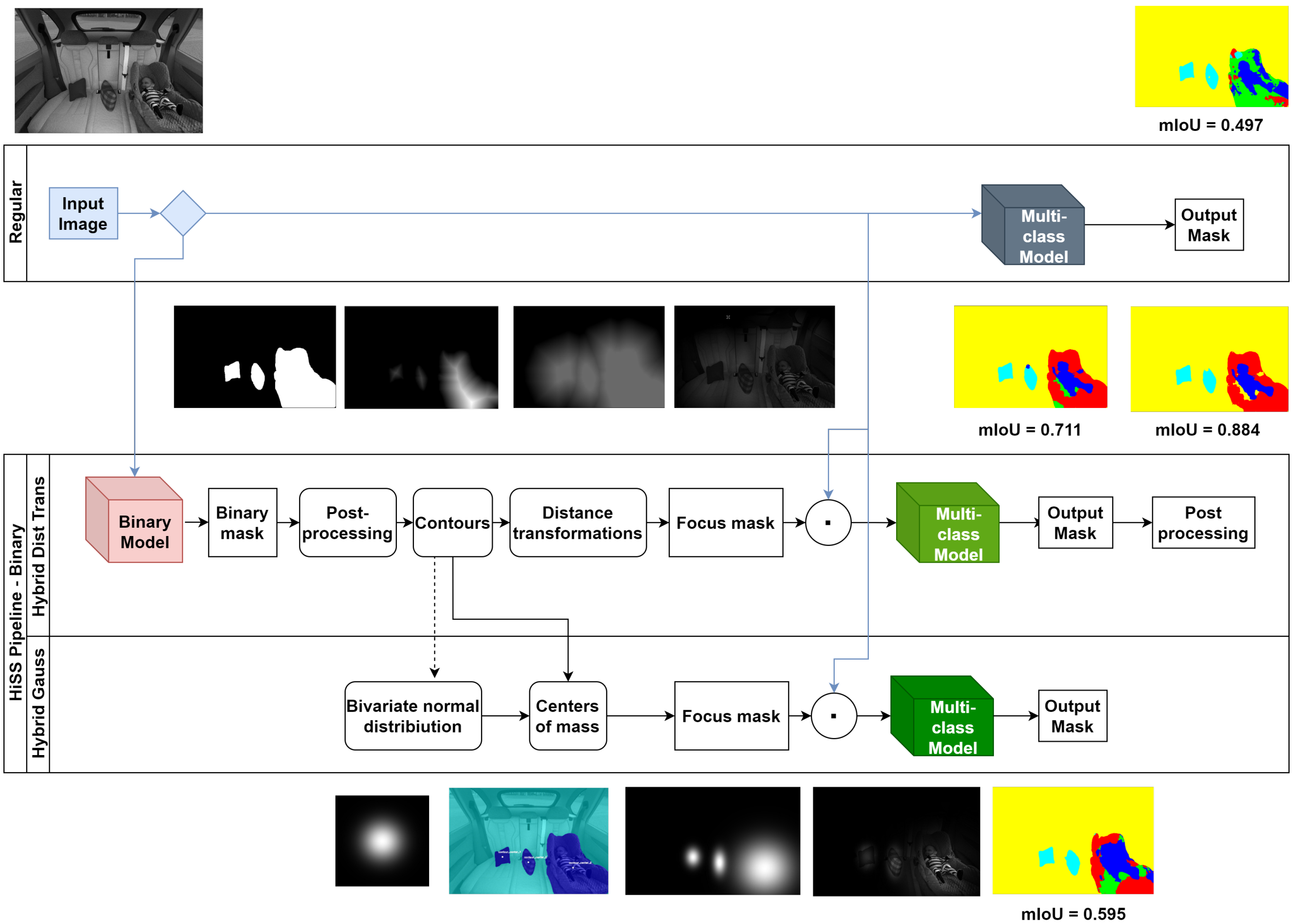

Figure 1. Proposed solution diagram showing regular and two hybrid model approaches: SegFormer + SegFormer with bivariate projection on object centers and SegFormer + SegFormer with distance transform on segmentation masks. This extended diagram includes preview images of the subsequent processing steps in two lanes: a) Distance Transformation approach: binary segmentation, post-processing, contours detection, distance transformation, focus mask creation and multiplication b) Gaussian approach: binary segmentation, post-processing, contours detection, bivariate normal distribution generation and placement, focus mask creation and multiplication. These steps are presented along with the final results – segmentation masks and mIoU metric results for each approach.