Supplementary Material

Supplementary material that may be helpful in the review process.

Appendix A. Multivariate Stratified Sampling Optimization

To preserve the diversity of visual element proportions, we used a multivariate stratified sampling scheme [1]. For each visual element $v \in V$, we compute its pixel-ratio distribution over all images and divide it into quartile groups $g \in G_v$. Let $S_{v,g}$ be the set of images whose proportion of element v falls into quartile g, and let s be the desired sample size. We select a binary vector $\mathbf{x} = (x_1, \ldots, x_n)$, where $x_i = 1$ if image i is kept, by solving:

$$\text{objective:} \quad \max \sum_{i=1}^{n} x_i \tag{A.1}$$

$$\text{subject to:} \quad \left\lfloor \frac{s}{\sum_{v \in V} |G_v|} \right\rfloor \leq \sum_{i \in S_{v,g}} x_i \quad \forall v \in V, \forall g \in G_v, \tag{A.2}$$

which enforces a minimum number of selected images in each quartile across all visual elements, yielding the final 150-image subset.

Appendix B. Randomized Image Assignment Matrix

Given n_p participants $\mathcal{P}$ and n_i selected images $\mathcal{I}$, we require that every image be rated by exactly $K_{\mathcal{P}}$ distinct participants and every participant view exactly $K_{\mathcal{I}}$ images, with $n_i K_{\mathcal{P}} = n_p K_{\mathcal{I}}$. In our study, we targeted $K_{\mathcal{P}} = 10$ ratings per image. We implement a randomized assignment procedure that iteratively fills an $n_p \times n_i$ binary matrix V such that:

$$\sum_{p=1}^{n_p} V_{p,i} = K_{\mathcal{P}} \quad \forall i \in \mathcal{I}, \tag{B.1}$$

$$\sum_{i=1}^{n_i} V_{p,i} = K_{\mathcal{I}} \quad \forall p \in \mathcal{P}, \tag{B.2}$$

where $V_{p,i} = 1$ denotes that participant p views image i.

Appendix C. Extended Experimental Protocol and Hardware Setup

This study was conducted with approval from the Institutional Ethics Committee of our university. The experiment followed a standardized protocol with three main phases, lasting at most 25 minutes to avoid fatigue [2, 3]. Fig. C.1 shows the complete experimental protocol for eye-gaze data collection.

Calibration: The HoloLens 2 device was cleaned between sessions. We performed a nine-point gaze calibration routine, monitoring accuracy in real time. Following established benchmarks, our setup maintained a quantitative spatial accuracy of $0.77° \pm 0.35°$ at the 2.0m focal distance used for stimuli projection.

Task familiarization: Participants viewed two reference images—one clearly perceived as *unsafe*, one as *safe*—for 15 s each to confirm stable tracking.

Safety rating trials: To prevent visual carry-over effects and ensure stable semantic scenes, we implemented a 5-second interval buffer between stimuli. Any trial failing to meet the 15-second viewing duration due to system interruption was automatically discarded.

Appendix D. ICE-T Questionnaire and Open-Ended Questions

The quantitative questions were directly *inspired by* the ICE-T framework [4], covering four dimensions: Insight, Confidence, Essence, and Time. Each dimension was assessed through two statements using a 7-point Likert scale (from "Totally Disagree" to "Totally Agree").

Insight.

QT1: "*In the By-Image view, the ability to switch between attention metrics and segmentation classes helped me identify relevant elements and generate new interpretations of visual attention.*".

QT2: "*The combination of visual representations in the By-Image view with the metrics in the By-Participant view allowed me to discover patterns and better understand participants' exploration behavior.*"..

Confidence.

QT3: "*In the By-Image view, the consistency of colors associated with semantic classes across visualizations, together with tooltips for inspecting exact values, increased my confidence in interpreting the data accurately.*".

QT4: "*The synchronized coordination between visualizations and the overlay of gaze/fixation points on the image increased my confidence in the correctness of spatial and temporal alignment.*".

Essence.

QT5: "*Both image-centered analysis and participant-centered analysis facilitated a quick and overall understanding of visual behavior.*".

QT6: "*The system effectively summarizes complex eye-tracking data by integrating multiple coordinated views, while maintaining clarity.*".

Time.

QT7: "*The use of temporal visualization tools allowed me to quickly relate when participants observed specific elements and identify dominant attention patterns over time.*".

QT8: "*Filtering options and smooth navigation significantly reduced the effort and time required to derive insights from eye-tracking data.*".

The qualitative questions collected open-ended feedback about the usefulness of the system and potential improvements:

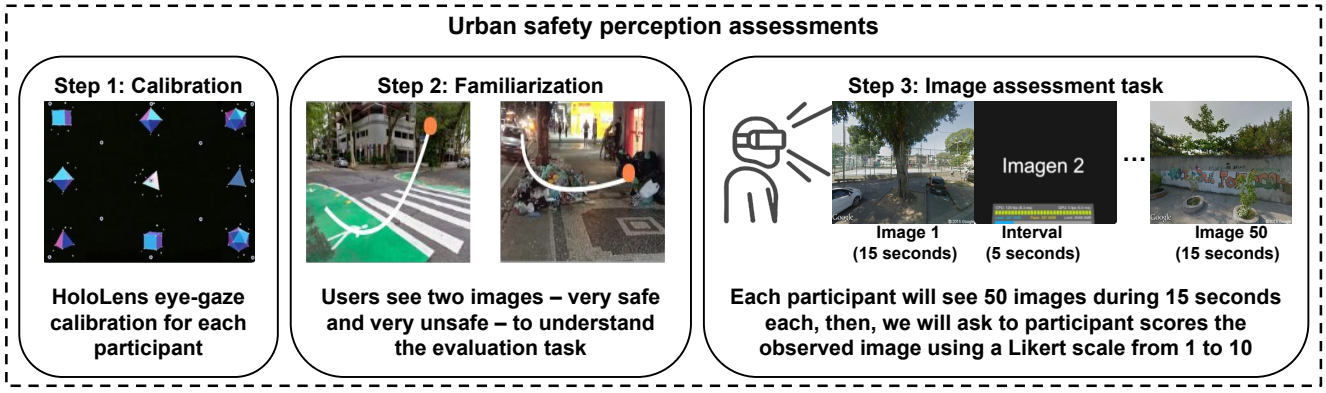

Figure C.1: Urban safety perception protocol. Step 1: gaze calibration with the HoloLens 2 headset. Step 2: task familiarization using one clearly safe and one clearly unsafe example image. Step 3: image assessment, in which each participant views 50 assigned images for 15 s, followed by a 5 s rating phase in which perceived safety is reported on a 1–10 Likert scale.

Table E.1: ICE-T evaluation scores per participant across the four dimensions.

Participant	*Insight*	*Confidence*	*Essence*	*Time*
P1	6.5	7.0	7.0	7.0
P2	6.0	6.0	6.0	6.0
P3	5.5	7.0	6.0	5.5
P4	3.5	4.0	6.0	5.0
P5	5.0	6.5	5.0	6.5
P6	6.5	7.0	6.0	6.5
P7	5.0	6.5	6.5	7.0
P8	5.5	5.5	5.5	6.5
P9	6.0	6.0	6.0	6.0
P10	5.5	7.0	6.5	6.5
P11	6.0	6.0	5.5	5.0
P12	5.5	6.5	6.0	4.5
P13	6.5	6.5	6.0	7.0
P14	7.0	7.0	6.5	7.0
P15	6.0	5.5	6.0	6.5
P16	5.5	5.5	6.0	5.5
P17	7.0	6.5	6.0	6.0
P18	7.0	7.0	7.0	5.0
P19	7.0	6.0	7.0	7.0
P20	6.5	7.0	7.0	7.0
P21	4.5	7.0	7.0	7.0
P22	6.0	5.0	6.5	4.0
Avg.	**5.9**	**6.3**	**6.2**	**6.1**

"*Considering both the By-Image and By-Participant views, which visual component, chart, or interaction did you find most useful for your analysis, and why?*" (QL1);"*What would you change, add, or remove from the interface to make it more efficient? Please feel free to include any additional comments or suggestions about your experience.*" (QL2).

Appendix E. ICE-T Results Per Participants

Table E.1 presents the ICE-T scores per participant, indicating consistently positive evaluations across all four dimensions.

Appendix F. Spatial Uncertainty and Kernel Calculation

To ensure the statistical validity of the *Attention Intensity* metric, it is necessary to calibrate the gaze data to account for the inherent hardware noise of the HoloLens 2. This appendix details the step-by-step derivation of the spatial kernel (σ) and the area exclusion threshold (τ) used to prevent false-positive gaze allocations on minute urban elements (e.g., distant overhead cables or small debris).

Step 1: Angular Precision of the Device

In eye-tracking methodology, spatial *precision* is the standard parameter used to define the dispersion of a Gaussian kernel, as it models the physiological and hardware jitter of a fixation. According to benchmark evaluations [5], the HoloLens 2 exhibits a spatial precision of 0.24° of visual angle at a focal distance of 2.0 meters, which is the exact distance used in our experimental setup.

Step 2: Conversion to Pixels Per Degree (PPD)

To translate this angular precision into the 2D pixel space of our stimuli, we calculated the device's angular resolution. Based on the manufacturer's specifications[1] and hardware evaluations, the HoloLens 2 displays a resolution of 2,048 × 1,080 pixels per eye with a diagonal field of view (FOV) of 52° and a diagonal of 2,315 pixels. Dividing this by the diagonal FOV yields an angular resolution of approximately 44.5 Pixels Per Degree ($2,315\ \text{px}/52^\circ \approx 44.5$ PPD).

Step 3: Deriving the Gaussian Kernel (σ)

Multiplying the device's angular precision by the calculated PPD directly provides the operational standard deviation (σ) for our heatmap generation. Thus, we applied a Gaussian kernel with $\sigma \approx 10.68$ pixels ($0.24^\circ \times 44.5$ PPD). This ensures that the spatial spread of each fixation accurately reflects the empirical uncertainty of the device.

Step 4: Defining the Area Exclusion Threshold (τ)

Finally, to establish a robust noise threshold (τ) for semantic segmentation, we considered an effective noise radius of 2σ ($2 \times 10.68 = 21.36$ pixels). In a 2D Gaussian distribution, a 2σ radius encompasses over 95.4% of the spatial uncertainty. Geometrically, this margin of error forms a circular area of roughly

[1] https://learn.microsoft.com/en-us/hololens/hololens2-hardware

1,433 squared pixels ($A = \pi \times 21.36^2$).

When evaluated against our 800×600 stimulus resolution (480,000 total pixels), this hardware noise floor occupies exactly 0.3% of the visual field ($1,433/480,000 \approx 0.003$). To ensure an even more robust and conservative analysis, we padded this noise floor with a safety margin, establishing a strict area threshold of $\tau = 0.005$ (0.5%). Any semantic mask or disorder cue occupying less than this threshold was considered statistically indistinguishable from hardware noise and was excluded from the analysis.

Appendix G. Inter-Rater Reliability (IRR) and Rater Calibration

To validate the consistency of urban safety perceptions and ensure that our findings reflect environmental features rather than individual biases, we performed a reliability analysis tailored for sparse rating matrices. Given our *Ill-Structured Measurement Design* (ISMD)—where 30 participants evaluated overlapping but incomplete subsets of 150 images—traditional Intraclass Correlation Coefficients (ICC) based on balanced ANOVA are mathematically biased.

Following the framework proposed by Putka et al. (2008), we utilized a Linear Mixed-Effects Model (LMM) with crossed random effects to isolate the variance components of our dataset ($N = 1,500$ observations). This approach allows for explicit rater calibration by partitioning the total variance into three distinct sources: the urban scene (signal), the participant (subjective bias), and residual noise. The variance components estimated via Restricted Maximum Likelihood (REML) are detailed in Table G.2.

Table G.2: Variance Components for Safety Perception Scores.

Source of Variation	Variance (σ^2)	Interpretation
Urban Scene (Image)	0.876	True variance in streetscape quality (Signal)
Participant (Rater)	0.394	Subjective severity/leniency bias (Calibration)
Residual (Noise)	1.766	Unexplained variance and interaction

To calculate the reliability of the aggregated safety scores, we applied the $G(q, k)$ estimator, which accounts for the partial overlap between raters in ISMD designs:

$$G(q,k) = \frac{\sigma^2_{img}}{\sigma^2_{img} + \frac{q \cdot \sigma^2_{part} + \sigma^2_{res}}{k}} \tag{G.1}$$

Where $k = 10$ represents the number of raters per image, and $q \approx 0.67$ is the coefficient of non-overlap for our design. This coefficient is mathematically derived from the ratio of raters per image to the total participant pool ($q = 1 - k/N_{total}$), reflecting that each image was evaluated by exactly one-third of our 30-participant panel ($1 - 10/30$).

The analysis yielded a high inter-rater reliability of $G(q, k) = 0.812$. In the context of behavioral research, a value exceeding 0.80 indicates a robust and high level of consensus. This result confirms that the safety metrics used in UrbanGazeVis are highly stable and that the sample of 10 raters per image provides sufficient statistical power to capture the collective perception of the urban environment.

Appendix H. Detailed Statistical Results for LMM

This section provides the comprehensive statistical outputs for the four semantic data evaluated in this study: *ADE20K Base*, *Grouped-ADE*, *ADE20K + UrbanPD4k*, and *Grouped-ADE + UrbanPD4k*. While the main text focuses on the most predictive features, the following tables provide the full set of fixed effects, including non-significant predictors, to ensure transparency and allow for meta-analytical comparisons.

To focus the analysis on strictly structural and urban elements, broad contextual categories (such as indoor, outdoor, nature, and miscellaneous) have been excluded from these summary tables. Each table reports the estimated coefficients (β), standard errors (S.E.), z-values, Variance Inflation Factors (VIF, where applicable), p-values, and 95% confidence intervals (CI) for the visual elements. Predictors are categorized into elements that positively correlate with perceived safety (Safe) and those that correlate negatively (Unsafe).

Appendix H.1. ADE20K Base Model Results

Table H.1 presents the results for the baseline ADE20K classes. This model serves as the primary benchmark for evaluating the impact of semantic granularity and disorder cues.

Appendix H.2. Grouped-ADE Model Results

Table H.2 details the results for the macro-level Grouped-ADE where elements are clustered into broad infrastructure categories. As discussed in the Discussion section of the main article, this model exhibits significant multicollinearity in certain categories.

Appendix H.3. ADE20K + UrbanPD4k Model Results

Table H.3 provides the complete output for the hybrid model integrating specific physical disorder cues. This model incorporates predictors such as "Broken/Damaged Wall" and "Graffiti", which emerged as critical indicators of perceived insecurity.

Appendix H.4. Grouped-ADE + UrbanPD4k Model Results

Finally, Table H.4 presents the results for the grouped classes augmented with disorder categories. This model evaluates whether the predictive power of disorder cues persists when environmental infrastructure is analyzed at a macro level.

Table H.1: Report: ADE20K (baseline)

Predictor	**Coef. (β)**	**S.E.**	**z-val**	**VIF**	**p-val**	**95% CI**
(Intercept)	5.091***	0.155	32.790	—	< .001	[4.790, 5.400]
Positive (Safe)						
Tree	0.195	0.084	2.310	3.400	0.056	[0.030, 0.360]
Door	0.182**	0.051	3.570	1.300	0.010	[0.080, 0.280]
Palm	0.130	0.056	2.320	1.560	0.056	[0.020, 0.240]
Ashcan	0.113	0.045	2.500	1.020	0.055	[0.020, 0.200]
Fence	0.099	0.093	1.060	4.290	0.398	[-0.080, 0.280]
Plant	0.097	0.055	1.760	1.520	0.162	[-0.010, 0.200]
Bus	0.096	0.050	1.910	1.250	0.135	[-0.000, 0.190]
Minibike	0.078	0.047	1.650	1.130	0.190	[-0.010, 0.170]
Streetlight	0.063	0.046	1.390	1.040	0.262	[-0.030, 0.150]
Car	0.036	0.076	0.470	2.910	0.719	[-0.110, 0.190]
Sidewalk	0.004	0.098	0.040	4.790	0.970	[-0.190, 0.200]
Negative (Unsafe)						
House	-0.003	0.054	-0.060	1.470	0.970	[-0.110, 0.100]
Van	-0.016	0.048	-0.330	1.170	0.803	[-0.110, 0.080]
Person	-0.030	0.055	-0.550	1.500	0.687	[-0.140, 0.080]
Truck	-0.062	0.066	-0.940	2.180	0.448	[-0.190, 0.070]
Stairs	-0.064	0.046	-1.390	1.050	0.262	[-0.150, 0.030]
Grass	-0.072	0.060	-1.190	1.820	0.352	[-0.190, 0.050]
Earth	-0.102	0.097	-1.050	4.740	0.398	[-0.290, 0.090]
Signboard	-0.105	0.056	-1.880	1.560	0.135	[-0.220, 0.000]
Field	-0.130*	0.046	-2.800	1.070	0.027	[-0.220, -0.040]
Mountain	-0.153*	0.047	-3.260	1.110	0.013	[-0.250, -0.060]
Pole	-0.154*	0.049	-3.180	1.170	0.013	[-0.250, -0.060]
Bridge	-0.159*	0.052	-3.060	1.360	0.015	[-0.260, -0.060]
Building	-0.179	0.219	-0.820	24.020	0.509	[-0.610, 0.250]
Sky	-0.188	0.077	-2.430	2.450	0.055	[-0.340, -0.040]
Road	-0.198	0.122	-1.620	7.520	0.190	[-0.440, 0.040]
Wall	-0.370	0.154	-2.400	11.820	0.055	[-0.670, -0.070]
Random Effects						
	Var.	**SD**	**ICC**			
Participant	0.665	0.815	0.184			
Residual	2.941	1.715				

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table H.2: Report: Grouped-ADE

Predictor	**Coef. (β)**	**S.E.**	**z-val**	**VIF**	**p-val**	**95% CI**
(Intercept)	5.091***	0.159	32.080	—	< .001	[4.780, 5.400]
Positive (Safe)						
Vegetation	0.262	0.149	1.760	10.210	0.288	[-0.030, 0.550]
Terrain Vehicle	0.097	0.136	0.710	8.550	0.781	[-0.170, 0.360]
Human	0.029	0.065	0.450	1.930	0.851	[-0.100, 0.160]
Negative (Unsafe)						
Construction	-0.057	0.306	-0.190	43.290	0.851	[-0.660, 0.540]
Floor	-0.073	0.254	-0.290	29.780	0.851	[-0.570, 0.420]
City Elements	-0.107	0.070	-1.530	2.260	0.346	[-0.240, 0.030]
Sky	-0.178	0.099	-1.790	4.110	0.288	[-0.370, 0.020]
Random Effects						
	Var.	**SD**	**ICC**			
Participant	0.692	0.832	0.179			
Residual	3.180	1.783				

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table H.3: Report: ADE20K + UrbanPD4k

Predictor	**Coef. (β)**	**S.E.**	**z-val**	**VIF**	**p-val**	**95% CI**
(Intercept)	5.091***	0.159	31.940	—	< .001	[4.780, 5.400]
Positive (Safe)						
Car / Taxi	0.197***	0.047	4.230	1.160	< .001	[0.110, 0.290]
Door	0.155*	0.051	3.060	1.360	0.010	[0.060, 0.250]
Tree	0.100	0.083	1.200	3.510	0.302	[-0.060, 0.260]
Plant	0.078	0.053	1.460	1.520	0.238	[-0.030, 0.180]
Trashcan	0.076	0.044	1.720	1.040	0.164	[-0.010, 0.160]
Palm	0.076	0.055	1.390	1.580	0.245	[-0.030, 0.180]
Streetlight	0.052	0.045	1.170	1.050	0.308	[-0.040, 0.140]
Motorcycle	0.041	0.046	0.890	1.140	0.412	[-0.050, 0.130]
Bus	0.041	0.049	0.840	1.270	0.429	[-0.060, 0.140]
Fence	0.013	0.089	0.150	4.180	0.884	[-0.160, 0.190]
Negative (Unsafe)						
Van	-0.047	0.047	-1.010	1.170	0.370	[-0.140, 0.040]
House	-0.050	0.053	-0.940	1.500	0.395	[-0.150, 0.050]
Sidewalk	-0.061	0.094	-0.650	4.600	0.532	[-0.240, 0.120]
Stairs	-0.068	0.045	-1.520	1.060	0.224	[-0.150, 0.020]
Car	-0.083	0.074	-1.130	2.870	0.315	[-0.230, 0.060]
Overhead Cable	-0.088	0.054	-1.640	1.500	0.184	[-0.190, 0.020]
Truck	-0.089	0.065	-1.370	2.250	0.245	[-0.220, 0.040]
Person	-0.095	0.054	-1.770	1.530	0.159	[-0.200, 0.010]
Grass	-0.116	0.060	-1.940	1.880	0.116	[-0.230, 0.000]
Ground	-0.117	0.090	-1.300	4.330	0.268	[-0.290, 0.060]
Field	-0.138*	0.045	-3.070	1.070	0.010	[-0.230, -0.050]
Garbage Bag	-0.142*	0.049	-2.900	1.280	0.012	[-0.240, -0.050]
Pole	-0.142*	0.048	-2.990	1.210	0.010	[-0.240, -0.050]
Signboard	-0.152*	0.055	-2.770	1.590	0.015	[-0.260, -0.040]
Mountain	-0.184***	0.046	-4.020	1.120	< .001	[-0.270, -0.090]
Bridge	-0.188**	0.051	-3.690	1.370	0.001	[-0.290, -0.090]
Sky	-0.200*	0.071	-2.830	2.170	0.014	[-0.340, -0.060]
Graffiti	-0.225*	0.075	-2.990	2.990	0.010	[-0.370, -0.080]
Damaged Pavement	-0.229***	0.056	-4.100	1.650	< .001	[-0.340, -0.120]
Road	-0.277*	0.117	-2.360	7.340	0.043	[-0.510, -0.050]
Building	-0.293	0.204	-1.430	22.120	0.238	[-0.690, 0.110]
Wall	-0.326*	0.121	-2.690	7.780	0.018	[-0.560, -0.090]
Broken/Dam. Wall	-0.442***	0.090	-4.930	4.270	< .001	[-0.620, -0.270]
Random Effects						
	Var.	**SD**	**ICC**			
Participant	0.707	0.841	0.203			
Residual	2.774	1.666				

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table H.4: Report: Grouped-ADE + UrbanPD4k

Predictor	**Coef. (β)**	**S.E.**	**z-val**	**VIF**	**p-val**	**95% CI**
(Intercept)	5.091***	0.161	31.700	—	< .001	[4.780, 5.410]
Positive (Safe)						
Car / Taxi	0.231***	0.050	4.590	1.270	< .001	[0.130, 0.330]
Vegetation	0.156	0.138	1.130	9.500	0.425	[-0.110, 0.430]
Trashcan	0.085	0.046	1.850	1.050	0.128	[-0.000, 0.170]
Negative (Unsafe)						
Terrain Vehicle	-0.024	0.124	-0.190	7.710	0.896	[-0.270, 0.220]
Human	-0.038	0.061	-0.620	1.870	0.690	[-0.160, 0.080]
Overhead Cable	-0.082	0.059	-1.380	1.740	0.299	[-0.200, 0.030]
Construction	-0.104	0.275	-0.380	37.730	0.848	[-0.640, 0.440]
City Elements	-0.128	0.066	-1.930	2.190	0.120	[-0.260, 0.000]
Garbage Bag	-0.142*	0.054	-2.630	1.460	0.031	[-0.250, -0.040]
Floor	-0.158	0.226	-0.700	25.460	0.690	[-0.600, 0.290]
Sky	-0.180	0.086	-2.100	3.260	0.107	[-0.350, -0.010]
Graffiti	-0.183	0.095	-1.930	4.450	0.120	[-0.370, 0.000]
Damaged Pavement	-0.204*	0.067	-3.040	2.240	0.011	[-0.330, -0.070]
Broken/Dam. Wall	-0.382*	0.122	-3.130	7.440	0.011	[-0.620, -0.140]
Random Effects						
	Var.	**SD**	**ICC**			
Participant	0.715	0.846	0.196			
Residual	2.943	1.716				

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Appendix I. Segmentation Model Comparison

Fig. I.2 shows the segmentation mask obtained from these three models. We note that SegFormer, PSPNet, and DeepLab do not segment the light pole; they fail to differentiate among walls, sidewalks, and gates, and they fail to segment cars correctly. However, while it requires more computational resources, the superior accuracy and versatility of OneFormer often justify the additional effort.

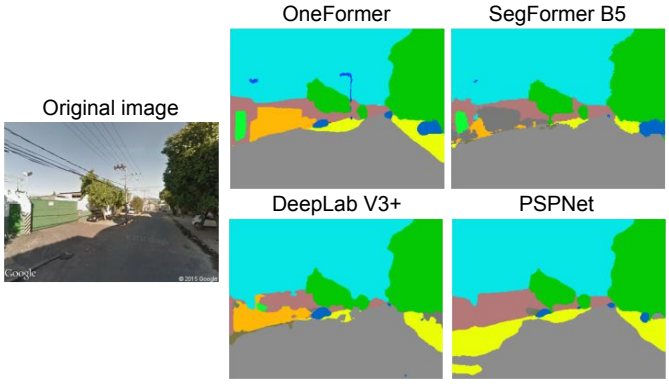

Figure I.2: Comparison of segmentation outputs produced by OneFormer, SegFormer B5, PSPNet, and DeepLabV3+. Some visual elements, such as light poles, cars, gates, and walls, are incorrectly classified, illustrating the differences in spatial accuracy across models.

References

[1] Khowaja, S, Ghufran, S, Ahsan, M. Estimation of population means in multivariate stratified random sampling. Communications in Statistics—Simulation and Computation® 2011;40(5). doi:10.1080/03610918.2010.551014.

[2] Tatler, BW, Hayhoe, MM, Land, MF, Ballard, DH. Eye guidance in natural vision: Reinterpreting salience. Journal of vision 2011;11(5):5–5. doi:10.1167/11.5.5.

[3] Clay, V, König, P, König, S. Eye tracking in virtual reality. Journal of eye movement research 2019;12(1):10–16910. doi:10.16910/jemr.12.1.3.

[4] Wall, E, Agnihotri, M, Matzen, L, Divis, K, Haass, M, Endert, A, et al. A heuristic approach to value-driven evaluation of visualizations. IEEE Transactions on Visualization and Computer Graphics 2018;25(1):491–500. doi:10.1109/TVCG.2018.2865146.

[5] Kapp, S, Barz, M, Mukhametov, S, Sonntag, D, Kuhn, J. Arett: Augmented reality eye tracking toolkit for head mounted displays. Sensors 2021;21(6):2234. doi:10.3390/s21062234.