# ClaraVid: A Holistic Scene Reconstruction Benchmark From Aerial Perspective With Delentropy-Based Complexity Profiling

Supplementary Material

## 8. More on Delentropic Scene Profile

This section expands on the design and interpretation of the Delentropic Scene Profile. We justify the use of the Beta distribution and clarify the conditions required for reliable cross-scene comparison. These clarifications aim to guide appropriate usage of DSP as a diagnostic tool in structured 3D scene understanding.

### 8.1. Motivation for Beta Distribution

The choice of distributional model for $\text{DSP}_S$ was guided by the analysis of complexity profiles of diverse datasets—ClaraVid, Skydrone[12], Skyscapes[24], UAVid[25], 3D Matrix City [10], Mill19 [50], and UrbanScene3D [11]—observing varied distributional forms contingent on the reconstruction context. A Gaussian distribution, while apt for uniformly complex scenes, fails to capture skewed or long-tailed distributions prevalent in heterogeneous environments (for example where we have a body of water or sky regions). Alternatives such as the *Pareto* or *Gamma* distributions accommodate skewness but lack flexibility for U-shaped profiles, which emerge in scenarios with pronounced bimodal complexity (e.g., a field with occasional high-detail trees). The Beta distribution, however, adeptly models this spectrum—uniform, skewed, and U-shaped—owing to its shape parameters $\alpha$ and $\beta$, requiring only four parameters (including bounds $a$ and $b$) for a parsimonious yet expressive fit. This simplicity affords intuitive interpretation: $\alpha$ and $\beta$ directly govern the density's form, while $a$ and $b$ anchor it to the empirical delentropy range. Multimodal distributions were eschewed, as their complexity—necessitating mixture models with additional parameters—hampers interpretability, risks overfitting, particularly given the finite sample sizes of $\{H_{\text{del},k}\}$. The Beta distribution emerges as the best choice, balancing flexibility, robustness, and theoretical clarity.

### 8.2. Comparability Across Scenes

The Delentropic Scene Profile provides a statistical summary of structural complexity and is designed to support comparisons of scenes in terms of their representational difficulty. While delentropy provides an absolute measure of complexity, its reliability for cross-scene comparison hinges on controlled acquisition conditions. *Image resolution* plays a critical role: small variations (e.g., $\approx$10%) tend to preserve the stability of the DSP, whereas more aggressive change of scales (e.g., $\geq$25%) can significantly alter high-frequency content, shifting the delentropy distribution (e.g., roof tiles may become indistinct when downsampled). The *spatial extent* of the covered scene is similarly influential. We have observed that DSP remains stable even under 4× scaling in area, but larger deviations or differences in semantic composition introduce shifts in the aggregated profile due to increased structural heterogeneity. Furthermore, the *image collection policy*–particularly the trajectory and sampling layout–can induce biases in the captured content, making structurally similar scenes appear dissimilar under mismatched acquisition patterns. These observations emphasize that to fully leverage DSP as a comparative tool, one should ensure alignment in resolution, spatial coverage, and collection policy—allowing differences in DSP to reflect genuine variations in scene complexity rather than artefacts of sampling.

### 8.3. Implementation Details

The delentropy of an image is obtained by first applying Gaussian blur to the input image using a spatial kernel of $3 \times 3$ pixels with a standard deviation of $1.0$, reducing high-frequency noise and mitigating sensitivity to minor textural variations. Subsequently, spatial gradients along horizontal and vertical directions are computed via Sobel filters with kernel size $3 \times 3$. The resulting gradient field is quantized into a two-dimensional histogram, the *deldensity*, employing $256$ bins per axis. This choice is driven by practical considerations to avoid histogram saturation, an artifact typically encountered when considering both positive and negative gradient values. Finally, the normalized deldensity is used as the joint probability distribution, from which delentropy is calculated.

## 9. Supplementary Evaluation for DSP

In this section, we evaluate the *DSP* on real-world datasets to assess its robustness under varied capture policies and reconstruction conditions.

### 9.1. UAVid

To evaluate the transferability and generalization of the DSP to real-world UAV imagery, we assess its correlation with reconstruction accuracy on the UAVid dataset. Unlike ClaraVid, UAVid features continuous image sequences captured under linear, L-shaped, and U-shaped trajectories, with no explicit grid-based coverage. We select six scenes (seq_13, seq_15, seq_29, seq_31, seq_36, seq_38), prioritizing static intervals. Full experimental setup provided in Section 10.4.

Table 5. **DSP evaluation on UAVid.** Higher DSP $\mu$ correlates with lower PSNR, SSIM and higher LPIPS.

| Scene | DSP | | | Nerfacto[62] | | | Gaussian Splatting [7] | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\mu$ | $\alpha$ | $\beta$ | PSNR ↑ | SSIM ↑ | LPIPS ↓ | PSNR ↑ | SSIM ↑ | LPIPS ↓ |
| seq 38 | *3.984* | *6.147* | *2.216* | 22.83 | 0.637 | 0.536 | 25.18 | 0.857 | 0.234 |
| seq 36 | *3.990* | *1.264* | *0.812* | 22.80 | 0.543 | 0.493 | 25.10 | 0.818 | 0.215 |
| seq 29 | *4.017* | *3.187* | *1.702* | 21.93 | 0.525 | 0.512 | 24.07 | 0.770 | 0.238 |
| seq 13 | *4.018* | *6.911* | *2.159* | 21.41 | 0.497 | 0.578 | 23.70 | 0.730 | 0.326 |
| seq 31 | *4.033* | *1.470* | *1.659* | 21.75 | 0.463 | 0.475 | 23.93 | 0.718 | 0.191 |
| seq 15 | *4.048* | *9.008* | *2.300* | 21.20 | 0.487 | 0.547 | 22.98 | 0.709 | 0.350 |

Despite a reduction in correlation between DSP and reconstruction metrics, the results–summarized in Table 5–remain consistent with prior trends observed in synthetic settings as presented in Figure 6. Notable degradations are attributed to real-world artifacts such as rolling shutter, motion blur, and the absence of structured spatial coverage, which collectively weaken the delentropy signal. We identify two recurring failure modes: (i) blurred input frames that yield reconstructions of unexpectedly high fidelity, and (ii) insufficient coverage in low-altitude flyovers (e.g., single-pass views) that lead to poor reconstructions despite an average measured delentropy. Visual examples provided in Figure 7.

### 9.2. DSP Analysis on Large Reconstructions

To further assess the generalization of *DSP* in real-world conditions, we analyze datasets used in recent large-scale urban reconstruction works, following the experimental setup of CityGaussian [49]. Specifically, we compute DSP values on the training and evaluation splits of UrbanScene3D [11] and Mill19 [50], which capture diverse urban geometries under unconstrained conditions. These datasets include varying levels of structural complexity, from sparse residential layouts to dense industrial sites. We align our computed DSP statistics–presented in Figure 8–with published reconstruction scores reported in CityGaussian, enabling a post-hoc correlation analysis without re-running the experiments.

As illustrated in Figure 9, we observe a general trend whereby scenes with higher DSP values correspond to lower reconstruction quality, particularly in PSNR. For example, the Mill19 industrial scene exhibits elevated delentropy ($\mu > 3.8$) and lower PSNR across multiple reconstruction methods, comparable to Urban High scenes in ClaraVid. This negative correlation persists across architectures–Mega-NeRF [50], CityGaussian [49], GNNeRF [67], Switch-NeRF [68], and Gaussian Splatting [7]–suggesting that DSP offers a dataset-agnostic signal of structural difficulty. SSIM and LPIPS exhibited more variability, and their correlation with DSP was less pronounced than that of PSNR. The consistency of trends across methods and environments reinforces DSP's utility as a lightweight proxy for estimating reconstruction difficulty in real-world benchmarks.

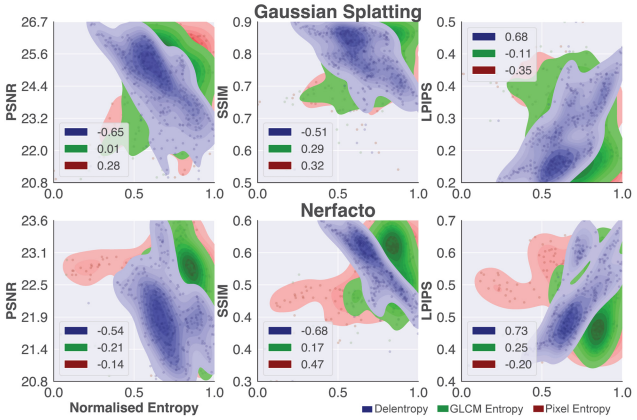


Figure 6. **Correlation between complexity metrics and reconstruction quality on UAVid.** Delentropy (blue) exhibits the strongest and most consistent correlation with PSNR, SSIM, and LPIPS across both evaluated models, outperforming GLCM texture entropy (green) and Shannon pixel entropy (red). Metrics are computed per image and aggregated over the test split.

## 10. Experimental Setup

### 10.1. Reconstruction Performance

Throughout our experiments, we utilize the *Nerfstudio* framework[62] to systematically benchmark various neural reconstruction methods. For **Nerfacto**[62], we adopt the *big* configuration. **TensoRF**[61] is configured with an increased grid resolution of $500^3$. **InstantNGP**[40] employs 16 grid levels, a maximum resolution of 8192, and a hash map size of $2^{21}$. **Zip-NeRF**[63] is re-implemented and used with default configuration. **Gaussian Splatting**[7] is initialized from the dataset's scene-level point cloud with a spatial resolution of *100 cm*, restricting point splitting beyond 3 million points to balance memory efficiency and reconstruction accuracy. Training durations vary across models, ranging from 60 to 720 minutes on an NVIDIA A6000 GPU, depending on architectural complexity. Across all experiments, we leverage FP16 training to optimize memory usage. The reported reconstruction and segmentation results correspond to a half-resolution output relative to the original image dimensions, maintaining a trade-off between computational feasibility and fidelity in large-scale aerial scene reconstruction.

### 10.2. Performance Across Varying Viewpoints

Both Nerfacto and Gaussian Splatting use the configuration described previously in Supplementary 10.1. For semantic learning, Nerfacto incorporates a jointly trained segmentation field, while Gaussian Splatting learns a separate set of 32-width semantic parameters—trained jointly but independent of geometry or positional encodings. Additionally, a 2-layer CNN classifier refines segmentation on the Gaussian

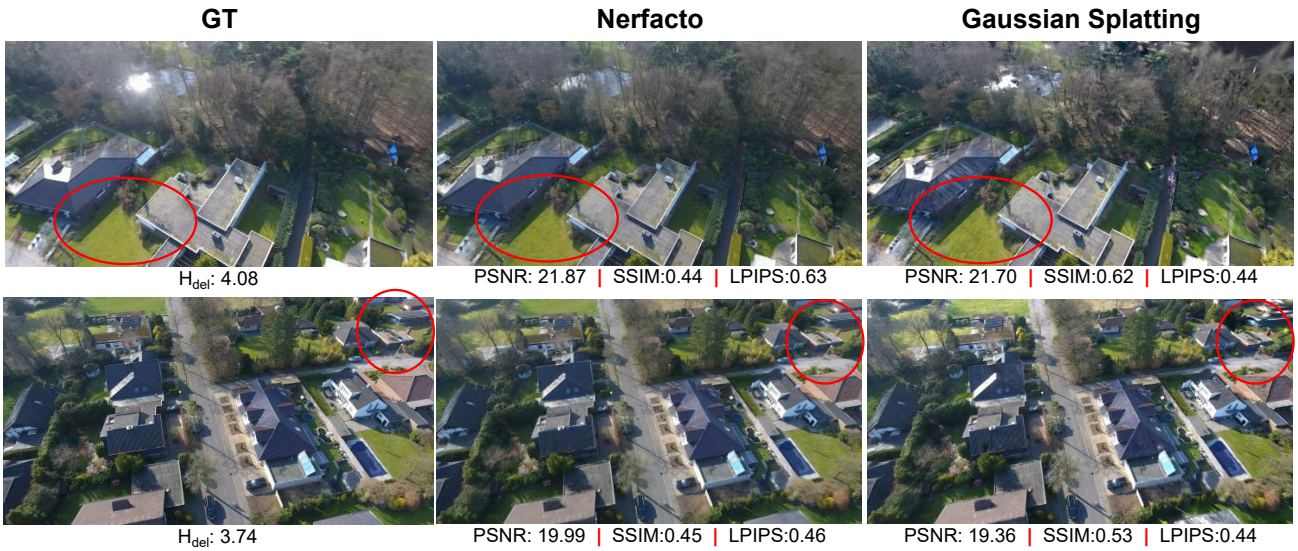


Figure 7. **Failure cases. Top:** Despite low delentropy and detailed Gaussian Splatting reconstruction, large errors arise from poor camera registration in the blurred input. **Bottom:** Incomplete scene coverage degrades performance; additionally, the ground truth image exhibits mild rolling shutter distortion.

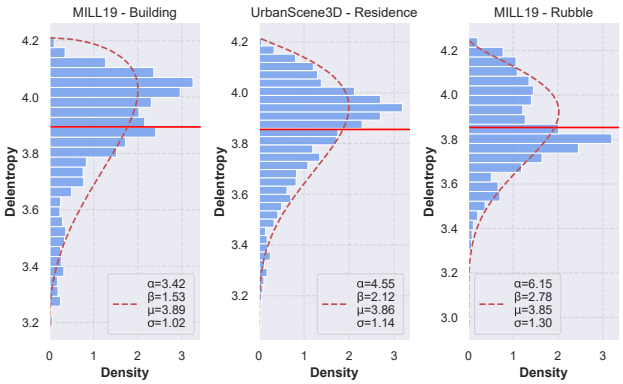


Figure 8. **Delentropy profiles for real-world mapping datasets**. The DSP indicates medium complexity, with variations reflecting structural differences across scenes. The profile is computed across all scene images.

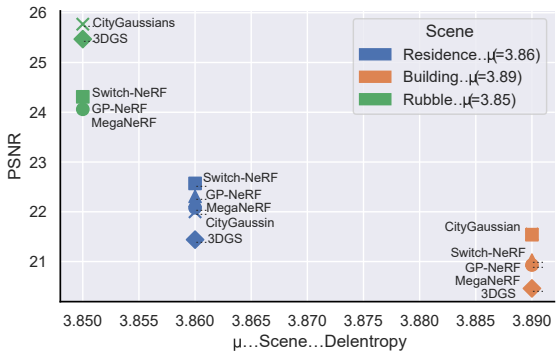


Figure 9. **Reconstruction quality vs. scene complexity**. The relationship between scene complexity, measured by the $\mu$ mean delentropy, and reconstruction fidelity (PSNR) across different neural reconstruction methods. A higher delentropy value correspond to lower PSNR scores, indicating increased reconstruction difficulty.

splats. We compute per scene class weighting to balance semantic segmentation classes. The depth for both methods is obtained from rendering with no additional heads or learned parameters.

### 10.3. Semantic Segmentation

We train a DPT[69] model with a DinoV2-L[70] backbone on both synthetic and real-world UAV datasets. The backbone is kept frozen while only the decoder is optimized. Our training protocol is standardized across datasets, with the number of steps adjusted proportionally to the dataset size—ranging from 5,000 to 30,000—to prevent overfitting. We use each dataset's native resolution and perform random crops of 630×630 pixels. Additionally, we uniformly apply data augmentations such as color adjustments, lighting variations, rotations, and distortions. Finally, the semantic palette is adapted to the real dataset following the [12] protocol, where both training and evaluation are done using reduced pallette.

### 10.4. UAVid Evaluation

For the UAVid evaluation, we focus on mostly static scenes, as dynamic object masks are not available. We select six representative sequences (13, 15, 29, 31, 36, 38), each covering an area of approximately $\approx 0.1$ km$^2$. All sequences consist of 901 frames, training samples are selected at every

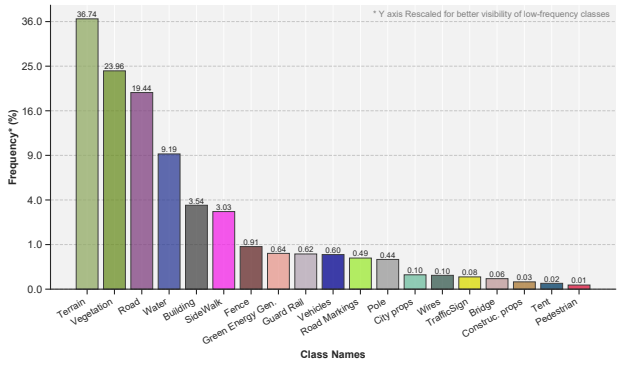


Figure 10. **Semantic Segmentation Pixel Label Distribution** The class distribution exhibits a long tail characteristic for aerial scenarios.

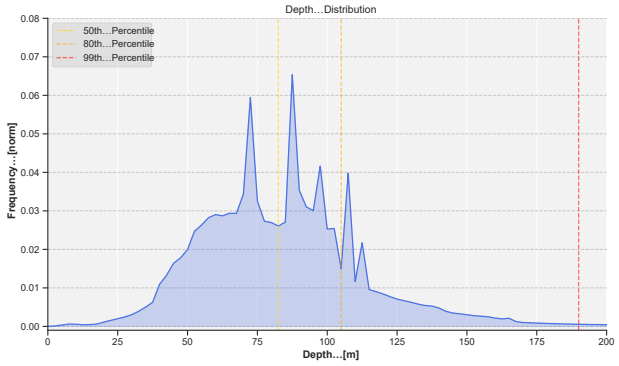


Figure 11. **Dataset Depth Distribution** The dataset's depth values exhibit a broad distribution with prominent spikes from nadir imagery, where ground-level elements dominate. While depths range from 0–1000 m, the 99th percentile is 192[m].

fourth index, while testing ones are chosen at every second offset (i.e., $i \equiv 2 \pmod 4$). Stationary frames with minimal motion are excluded, as they often are poorly registered. Additionally, we consider the first and last 100 frames in train set, to mitigate any potential collection policy errors and minimize the spatial content outside the region of interest. All images are downsampled by a factor of 2, and reconstructions are performed using COLMAP[71, 72]. All other settings for the evaluated models follow the protocol described in Section 10.1.

# 11. Claravid

In this section, we present additional insights into Claravid. Additionally we present a more detailed visual overview in Figure 12.

## 11.1. Semantic Complexity

The semantic complexity of *Claravid* is a direct consequence of its enriched class taxonomy, designed to capture the nuanced structures present in aerial urban and rural environments. By extending the conventional label set to include *wire*, a class that encapsulates thin linear structures, and *green energy*, which aggregates solar panels and renewable energy infrastructure, the dataset reflects the intricate composition of real-world landscapes. Furthermore, the *pole* category has been redefined to encompass slender metallic structures such as communication towers and high-voltage power lines, thus improving its generalization for fine-grained segmentation tasks. Additionally, the introduction of *urban props* and *construction props* provides a finer semantic partitioning of the environment, accounting for human-centric elements in residential areas and industrial material clusters, respectively. The inclusion of a dedicated *tent* class further enhances the granularity of temporary and semi-permanent structures. This expanded semantic schema not only increases the dataset's diversity but also contributes to its delentropy-based complexity profiling, allowing for a more rigorous quantification of scene heterogeneity in both structural and semantic dimensions. We present the label class distribution in Figure 10 and complementary, the depth distribution across the entire dataset in Figure 11.

## 11.2. Rendering Pipeline Configuration

To achieve high-fidelity rendering of large-scale aerial scenes in Unreal Engine 4, we implement targeted modifications to the rendering pipeline, prioritizing fine detail preservation and visibility consistency at extended distances. To mitigate the disappearance of distant objects and enhance detail clarity, we increase the rendering resolution by setting `r.ScreenPercentage=300`, effectively supersampling the scene at 3 times the native resolution. This adjustment minimizes aliasing artifacts—particularly pronounced in oblique aerial perspectives—and ensures that small-scale features remain discernible, albeit at a higher computational cost justified by the resulting visual fidelity. Shadow integrity for distant fine elements, such as foliage and thin structures, is preserved by reducing `r.Shadow.MinRadius` to 0.001, which enables shadow casting for geometrically narrow features that would otherwise be lost. Shadow map resolution is also increased to *2048×2048* for objects considered having fine details, maintaining sharp shadow edges across expansive terrains. Geometric fidelity is ensured by disabling level-of-detail (LOD) transitions, preventing mesh simplification that degrades structural complexity at greater distances or altitudes, a choice prioritizing rendering quality over real-time performance scalability. Similarly, we disable distance-based object culling to maintain persistent visibility of all scene elements, eliminating sudden visibility discontinuities that disrupt spatial and temporal coherence.

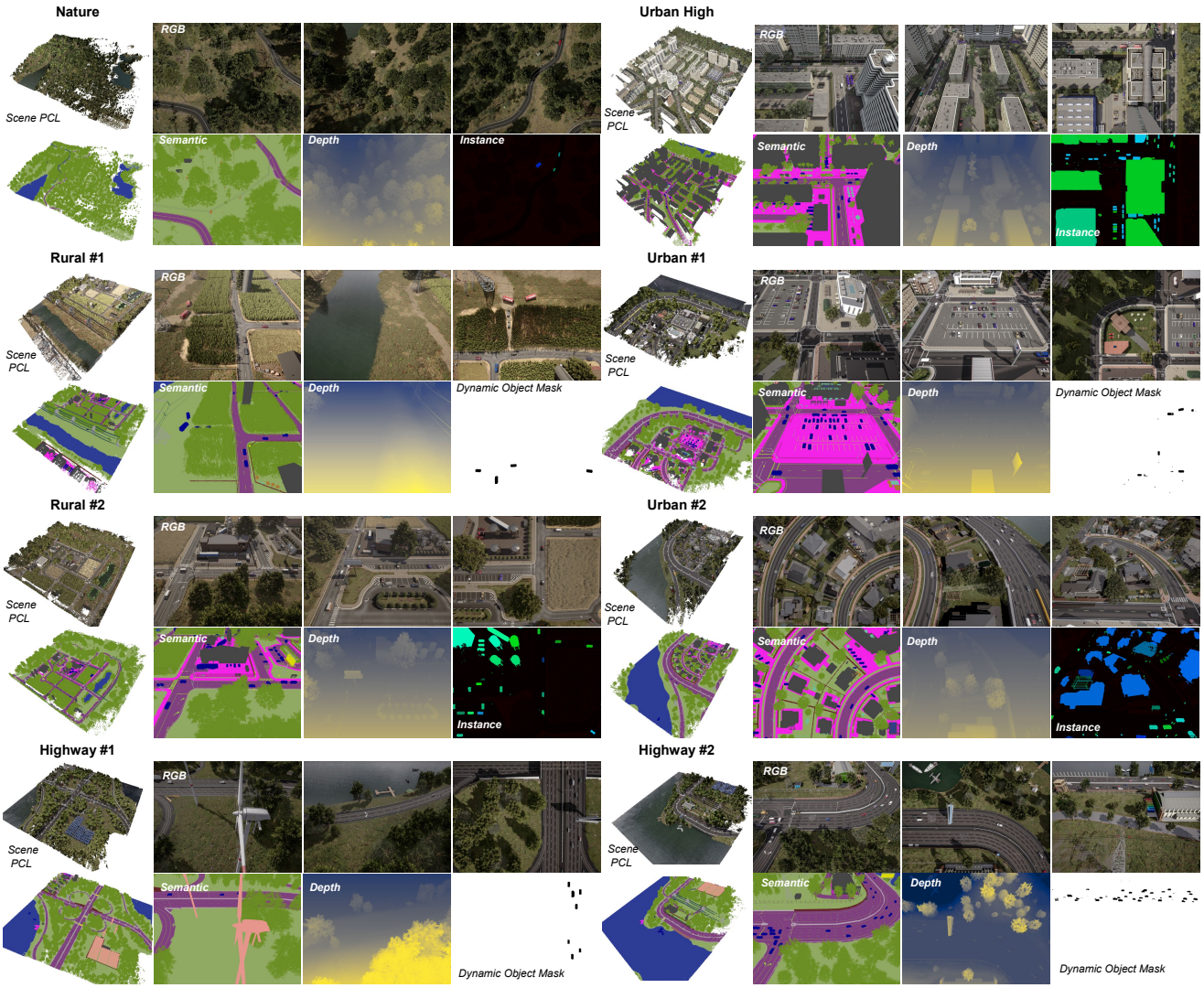


Figure 12. **ClaraVid Modalities.** Overview of the 8 UAV missions, showing representative frames with RGB, depth, semantic, instance, and dynamic annotations, along with scene-level point cloud views.