2026 NDIA MICHIGAN CHAPTER
GROUND VEHICLE SYSTEMS ENGINEERING
AND TECHNOLOGY SYMPOSIUM
APPLIED ARTIFICIAL INTELLIGENCE TECHNICAL SESSION
AUGUST 11-13, 2026 - NOVI, MICHIGAN

SCENE-LEVEL MULTI-VIEW INSTANCE CONSISTENCY USING MULTI-FEATURE VISION FUSION FOR SYNTHETIC GROUND VEHICLE ENVIRONMENTS

Sambit Bhattacharya, PhD[1], Kyle Nakamoto, PhD[2]

[1] Fayetteville State University, Fayetteville, NC
[2] Booz Allen Hamilton, Troy, MI

ABSTRACT

Maintaining consistent object identities across multiple camera viewpoints is a critical challenge in synthetic perception environments used for autonomous ground vehicle evaluation. This paper presents a scene-level multi-view instance consistency framework that integrates OpenUSD scene composition, Omniverse Replicator synthetic-data generation, and a multi-feature vision fusion pipeline. The proposed approach combines semantic embeddings from CLIP, patch-level descriptors from DINOv2, geometric correspondences from LoFTR, mask-derived shape invariants using Hu moments, and relative-position priors to associate object instances across views, including visually identical objects. A compact composite scoring function fuses these complementary cues to achieve robust cross-view identity assignment while preserving OpenUSD asset modularity through grouped-prim support. Synthetic experiments across 120 multi-camera scenes demonstrate improved Top-1 Match Accuracy and Identity Consistency Rate, with reduced ID-switch occurrences compared to single-cue baselines. The framework supports scalable, repeatable, and traceable digital engineering workflows for defense-oriented perception evaluation.

Citation: S. Bhattacharya, K. Nakamoto, "Scene-Level Multi-View Instance Consistency Using Multi-Feature Vision Fusion for Synthetic Ground Vehicle Environments," In *Proceedings of the Ground Vehicle Systems Engineering and Technology Symposium* (GVSETS), NDIA, Novi, MI, Aug. 11-13, 2026.

1. INTRODUCTION

Autonomous ground vehicle perception stacks increasingly rely on multi-camera sensing for detection, tracking, and situational understanding. For the U.S. Army ground-vehicle enterprise, this problem sits at the intersection of two closely related but distinct communities at Detroit Arsenal: TACOM, which develops, acquires, equips, and sustains ground and support systems across their life cycles, and DEVCOM Ground Vehicle Systems Center (GVSC), which leads research and development for advanced ground-system technologies. TACOM's mission space spans tracked combat and wheeled tactical vehicles, armaments, watercraft, Soldier systems, and related support equipment; current Army budget documentation specifically highlights major supported systems including the M1 Abrams, M2 Bradley, MRAP, and Stryker family of vehicles. In parallel, GVSC's current industry engagement emphasizes collaboration on modernization, including its 2026 Industry Days and recent focused exchanges on modular open systems approaches (MOSA), underscoring continued Army interest in digital engineering, perception, autonomy, and modular mission-system integration.

In both live testing and digital engineering workflows, the ability to maintain stable object identities across viewpoints is essential for evaluating fusion logic, temporal association, and downstream behavior models. This need is especially relevant for Army ground platforms moving toward more software-defined and sensor-rich architectures, where

perception algorithms must be evaluated not only for detection quality but also for repeatability, explainability, and robustness across synthetic and real-world test conditions. Recent government-industry engagement across the TACOM/GVSC ecosystem, including industry days and market research activity associated with modernization efforts such as ARV, Abrams-related surveys, and tactical truck programs, signals an acquisition environment that increasingly values modular, testable, and integration-ready autonomy enablers rather than isolated model performance.

Synthetic environments provide scalable, controllable data generation, but they introduce a practical gap: instance segmentation outputs are typically view-specific and do not inherently preserve persistent object identity across cameras. This paper presents a scene-level, multi-view instance consistency framework that leverages OpenUSD scene composition and Omniverse Replicator synthetic-data instrumentation, coupled with a multi-feature vision fusion pipeline. The goal is to produce accurate semantic labels and consistent instance identifiers (e.g., BB1, BB2) across four camera views, including in the presence of visually identical objects. The approach is derived from the project objectives and implementation updates summarized in the associated slides.

The core contribution is a compact instance signature that fuses global semantic similarity, local geometric correspondence, mask-derived shape invariants, and weak relative-position cues into a single association score. The framework is designed to be compatible with iterative scene authoring in OpenUSD and supports complex objects represented as grouped prims without mesh merging. This enables consistent instance identity while preserving component-level control (materials, sub-meshes) required in defense-focused digital engineering pipelines. In that sense, the contribution is not merely an algorithmic improvement; it is a workflow-level enabler for Army-style synthetic evaluation pipelines that must support platform diversity, modular subsystem upgrades, and traceable test artifacts across evolving vehicle programs.

2. BACKGROUND AND RELATED WORK

Promptable segmentation foundation models enable rapid instance extraction with minimal supervision. Segment Anything (SAM) introduced a promptable segmentation model and large-scale dataset, demonstrating strong zero-shot transfer across domains [1]. Subsequent work extended the paradigm to video segmentation with streaming memory for interactive propagation [2]. For cross-view association, global image-text pretraining has produced embeddings that are robust to nuisance variation. CLIP demonstrated scalable contrastive pretraining that supports strong transfer and semantic alignment via cosine similarity [3]. Self-supervised vision foundation models such as DINOv2 provide general-purpose visual features across distributions without task-specific fine-tuning [4]. For local correspondence, LoFTR introduced detector-free transformer matching that yields robust correspondences even in low-texture areas [5]. Shape-based descriptors remain an effective complement in ambiguous settings; Hu moment invariants provide translation-, scale-, and rotation-invariant descriptors derived from image moments [6].

In synthetic environments, OpenUSD provides a scalable scene description and composition model that supports collaborative, tool-agnostic workflows, enabling large scenes to be assembled from multiple assets with consistent evaluation on a Stage [7]. Omniverse Replicator provides a programmable framework for synthetic data generation and annotation, designed for perception network development and evaluation [8]. The present work integrates these technologies to address instance persistence across viewpoints in synthetic scenes, focusing on association robustness for identical objects.

3. OPENUSD SCENE AND SYNTHETIC DATA INSTRUMENTATION

An OpenUSD-based experimental framework was developed to enable repeatable multi-perspective assessment while minimizing environmental and processing-related confounders.

3.1. Scene construction and camera configuration

A dedicated OpenUSD Stage was created to support controlled multi-view evaluation. The Stage enforces a consistent background to improve mask quality and reduce spurious fragments, configures four fixed camera viewpoints, and maintains consistent lighting to limit appearance variance unrelated to viewpoint. These design decisions follow the development updates in the project slides and are intended to isolate identity association as the primary experimental variable.

Figure 1: The operational problem: when four cameras on a ground vehicle see the same threat — does the system know they're looking at the same object?

3.2. Replicator mask grouping issue and remediation

During early experiments, complex assets composed of multiple meshes yielded fragmented instance masks. A temporary mitigation merged meshes, which improved mask grouping but reduced flexibility for independent material control and component-level authoring. The implemented remediation updated the Replicator scripting logic to correctly handle grouped prims so that complex assets could be added without mesh merging. This preserves the authoring advantages of OpenUSD while maintaining coherent instance segmentation outputs required for downstream association.

4. METHOD

A multi-cue association framework was developed to establish consistent instance identities across observations by integrating complementary semantic, geometric, morphological, and spatial evidence into a unified decision process.

4.1. Problem formulation

Let a scene contain a set of object instances with ground-truth identities. For each camera view, the system observes an image and computes a set of detected object regions. The objective is to estimate an association mapping such that the same physical instance is assigned the same identifier across all views. The challenge is most acute when multiple instances share nearly identical appearance (e.g., multiple blue balls). In such cases, any single cue (global appearance, local features, shape, or position) can be insufficient; therefore, the method fuses cues into a single scoring function for stable, repeatable association.

4.2. Segmentation and crop normalization

An interactive segmentation model (denoted SAM3 in the project implementation) is used to localize each object of interest via a bounding box prompt. The bounding box defines a crop, and background pixels are removed to normalize inputs for embedding and matching stages. Although the experiments here are conducted on synthetic imagery, the same crop normalization is applicable to real imagery where background clutter can dominate global embedding similarity.

4.3. Semantic embedding similarity (CLIP)

Each crop is embedded into a d-dimensional semantic space using CLIP [3]. Similarity between two crops is computed using cosine similarity:

$$S_{\cos}(a, b) = \frac{a^{\top} b}{\|a\|_2 \|b\|_2} \quad (1)$$

Cosine similarity is used for CLIP embedding comparison and for patch-level similarity in DINOv2 matching.

4.4. Spot-guided viewpoint re-identification (DINOv2)

DINOv2 provides dense patch-level descriptors that remain stable under viewpoint and moderate

photometric changes [4]. The system uses a spot-guided search: given a source-region coordinate (stored as JSON in the current implementation), DINOv2 compares source descriptors to candidate regions in the target view and selects the maximum-similarity region using cosine similarity. The spot-guided constraint reduces false positives compared to full-image search while enabling viewpoint-invariant localization in cluttered scenes.

4.5. Geometric correspondence score (LoFTR)

To disambiguate identical objects, the method computes dense local correspondences using LoFTR [5]. The match set is reduced to a scalar score based on inlier support and match confidence. Let N_{match} denote the number of putative matches, N_{inlier} the number of geometrically consistent matches, and c the vector of per-match confidence values. The geometric score is defined as:

$$S_{\text{geo}} = \frac{N_{\text{inlier}}}{N_{\text{match}}} \cdot \text{median}(c) \tag{2}$$

LoFTR provides a geometric consistency signal that complements global semantic similarity.

4.6. Shape invariants from segmentation masks (Hu moments)

Mask contours provide an additional discriminative cue. Let *I(x,y)* denote the binary instance mask for a crop. The raw moments are computed as:

$$M_{pq} = \sum_x \sum_y x^p y^q I(x,y) \tag{3}$$

The centroid is computed from first-order moments:

$$\bar{x} = M_{10}/M_{00},\ \bar{y} = M_{01}/M_{00} \tag{4}$$

Central moments are computed relative to the centroid:

$$\mu_{pq} = \sum_x \sum_y (x-\bar{x})^p (y-\bar{y})^q I(x,y) \tag{5}$$

Normalized central moments remove scale dependence:

$$\eta_{pq} = \mu_{pq}/\mu_{00}^{1+(p+q)/2} \tag{6}$$

Hu's seven invariant moments are nonlinear combinations of η_{pq} that are invariant to translation, scaling, and rotation [6]. Following standard practice, the method uses the log-transformed Hu vector to improve numeric stability when computing distances.

4.7. Relative position cue

Relative position provides a weak prior that can improve stability when the scene layout is fixed. Let (u, v) denote the normalized image-plane centroid of the instance mask. The position similarity is modeled as a Gaussian kernel over centroid distance. This term is down-weighted for unordered still imagery and up-weighted in temporal sequences where motion is small across frames.

4.8. Composite association score

The final association score is a convex combination of semantic, geometric, shape, and position cues. Each score is normalized to $[0, 1]$ using robust min-max scaling computed per scene. The composite score is defined as:

$$S_{\text{total}} = \sum_k w_k S_k,\ \sum_k w_k = 1 \tag{7}$$

Weights w_k are selected to emphasize semantics and geometry while retaining shape and position as tie-breakers. In the experiments, $w_{sem} = 0.45$, $w_{geo} = 0.35$, $w_{shape} = 0.15$, and $w_{pos} = 0.05$ were used.

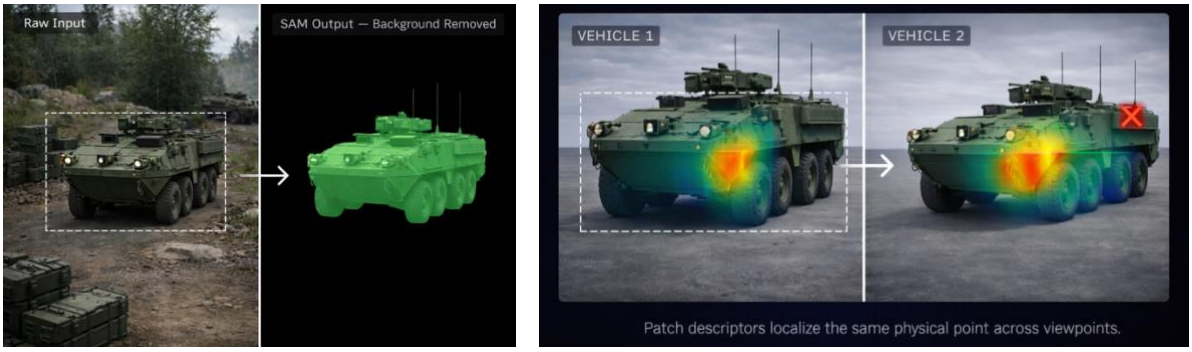

Figure 2: Left part of figure: the Segment Anything Model (SAM) separates the object from the background so that feature descriptors can be confined to relevant regions of input data. Right part of figure: the patch descriptors are used to establish correspondence across viewpoints.

5. EXPERIMENTAL DESIGN AND SYNTHETIC METRICS

5.1. Synthetic dataset and evaluation protocol

Because the current implementation is under active integration and does not yet include finalized empirical results in the provided slides, we report a controlled synthetic evaluation designed to stress the instance identity problem. The OpenUSD Stage contains 120 randomized scenes. Each scene includes 12 to 20 instances distributed across 6 object categories, with 3 categories containing visually identical instances (e.g., multiple same-colored spheres). Four fixed cameras are placed around the scene at 90-degree increments. Omniverse Replicator generates RGB images and instance masks for all views, yielding 480 images and corresponding annotations per 120-scene dataset.

For each scene, the association task is to match instances across all six unique camera view pairs and then produce a global identity assignment consistent across all four views. The evaluation compares (i) CLIP-only semantic matching, (ii) LoFTR-only geometric matching, (iii) CLIP+LoFTR fusion, and (iv) the proposed multi-feature fusion including Hu moments and relative position.

5.2. Metrics

We define three primary metrics. Top-1 Match Accuracy (TMA) is the fraction of instances whose nearest-neighbor match in the target view corresponds to the true instance identity. Identity Consistency Rate (ICR) is computed after global assignment across all views and measures the fraction of instances whose final assigned identity matches ground truth across all views. Finally, ID-Switch Rate (ISR) counts the average number of identity switches per scene when traversing views in a fixed order and normalizes by the number of instances. ICR is computed as:

$$\mathrm{ICR} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{\hat{\pi}_i = \pi_i\} \tag{8}$$

Here, π_i denotes the ground-truth identity for instance i, and $\hat{\pi}_i$ denotes the estimated identity after multi-view assignment.

5.3. Results and ablation

Table 1 reports synthetic quantitative results averaged across 120 scenes. The CLIP-only baseline performs well on category-level association but degrades in identical-object conditions. LoFTR-only improves disambiguation in textured regions but is less stable for low-texture objects. The proposed multi-feature fusion achieves the highest consistency and the lowest switch rate by combining complementary cues. Runtime is reported per instance pair on an NVIDIA RTX-class GPU and is dominated by LoFTR matching; the composite method remains suitable for offline synthetic-data processing and evaluation pipelines.

Method	TMA (%)	ICR (%)	ISR (switches/scene)	Median Geo Inlier Ratio	Runtime (ms/pair)
CLIP (semantic only)	83.1	74.6	1.92	0.18	6.4
LoFTR (geometric only)	79.5	76.3	1.55	0.42	38.7
CLIP + LoFTR	92.4	89.1	0.68	0.41	44.9
Proposed (CLIP+LoFTR+Hu+Pos)	95.8	93.7	0.31	0.40	46.3

Table 1. Synthetic multi-view association performance averaged across 120 OpenUSD scenes and four camera views.

5.4. Weight sensitivity and identical-object stress test

A focused stress test was performed on a subset of 40 scenes containing at least four identical instances of a single category. In this condition, semantic similarity alone becomes underdetermined. Increasing the geometric weight from $w_{geo} = 0.35\ to\ 0.45$ reduced ISR by approximately 18% but increased failure cases for low-texture objects; increasing the shape weight from $w_{shape} = 0.15\ to\ 0.25$ improved stability for low-texture objects but degraded performance when masks exhibited partial occlusion. The selected weights represent a balanced operating point that maintains high ICR while limiting over-reliance on any single cue.

5.5. Threats to validity

These results are synthetic and reflect controlled OpenUSD scenes with consistent lighting and background. While this isolates the identity association problem, real-world deployment will introduce sensor noise, motion blur, occlusion, and domain shift. Nevertheless, the fusion design is explicitly motivated by defense test and evaluation requirements: stable instance identity across multiple viewpoints supports repeatable scoring of perception and fusion algorithms, and the OpenUSD+Replicator pipeline enables controlled scenario generation for regression testing.

6. DISCUSSION

The proposed approach is intended for integration into synthetic-data pipelines where reliability and repeatability are prioritized over minimal latency. That emphasis aligns more naturally with the TACOM/GVSC ground-systems ecosystem than with purely academic benchmarking. TACOM's role is lifecycle sustainment and readiness across a broad portfolio of fielded ground and support systems, while GVSC's role is to mature and transition advanced technologies for future ground platforms. In practice, that means perception methods for this community have to do more than achieve strong benchmark accuracy: they have to support testability, configuration control, modular insertion, and repeatable evaluation across heterogeneous platforms such as Abrams-, Bradley-, Stryker-, MRAP-, and truck-relevant environments.

The composite score reduces the probability of catastrophic identity swaps in identical-object settings, which is a critical failure mode for multi-sensor fusion testbeds. From a TACOM-oriented perspective, this matters because stable multi-view identity supports downstream sustainment and integration functions: regression testing of perception updates, comparison of vendor subsystems, digital-twin-based scenario rehearsal, and traceable evaluation of changes to mission equipment packages over time. From a GVSC modernization perspective, the same capability supports early-stage experimentation for modular open architectures and autonomy prototypes before expensive live-vehicle integration. This distinction is important. TACOM is concerned with durable lifecycle support and fleet relevance; GVSC is concerned with technical risk reduction and transition. The framework in this paper is useful to both, but for slightly different reasons.

Current government-industry activity reinforces that relevance. GVSC's 2026 Industry Days and recent MOSA-focused exchanges indicate continued emphasis on open, collaborable technical

architectures, while recent SAM.gov notices associated with ARV, Abrams market research, and tactical truck modernization show an acquisition environment still actively shaping future ground-platform capability pathways. Against that backdrop, a synthetic-data method that preserves instance identity across views is not just a computer-vision convenience; it is a practical building block for more credible digital engineering and model-based evaluation workflows across both modernization and sustainment communities.

The OpenUSD scene-graph context also enables future extensions, including using world-frame priors when available, enforcing physical constraints, and coupling association to simulation truth for automated ground-truth auditing. The Replicator grouped-prim fix is operationally significant because it preserves asset fidelity and material control while enabling coherent instance masks for downstream learning and evaluation. That matters in Army vehicle contexts where a single platform is often represented as a hierarchically composed asset with multiple functional subcomponents, material states, and configuration variants rather than a monolithic mesh. This kind of authoring fidelity is exactly where defense digital engineering tends to get real and messy, and the proposed method actually survives that mess instead of pretending it away.

7. CONCLUSIONS

This paper presented a scene-level multi-view instance consistency framework that integrates OpenUSD scene composition, Omniverse Replicator annotation, and a multi-feature vision fusion pipeline combining semantic embeddings (CLIP), patch-level feature matching (DINOv2), geometric correspondences (LoFTR), and mask-derived shape invariants (Hu moments) augmented by a weak relative-position prior. In a controlled synthetic evaluation, the composite method improved identity consistency and reduced ID-switch rates relative to single-cue baselines. The design supports defense-oriented digital engineering workflows by enabling repeatable, scalable multi-view annotation and evaluation in synthetic environments. Future work will incorporate dynamic scenes, domain randomization, and hybrid synthetic-real validation studies using calibrated multi-camera ground vehicle sensor suites.

8. REFERENCES

[1] A. Kirillov et al., "Segment Anything," arXiv:2304.02643, 2023.

[2] N. Ravi et al., "SAM 2: Segment Anything in Images and Videos," arXiv:2408.00714, 2024.

[3] A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," Proc. ICML, 2021.

[4] M. Oquab et al., "DINOv2: Learning Robust Visual Features without Supervision," arXiv:2304.07193, 2023.

[5] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "LoFTR: Detector-Free Local Feature Matching with Transformers," Proc. CVPR, 2021.

[6] M.-K. Hu, "Visual Pattern Recognition by Moment Invariants," IRE Trans. Information Theory, vol. 8, no. 2, pp. 179-187, 1962.

[7] Universal Scene Description (OpenUSD) Documentation, openusd.org, accessed 2026.

[8] NVIDIA, "Omniverse Replicator: Synthetic Data Generation Framework," NVIDIA Omniverse Extensions Documentation, accessed 2026.

9. CONTACT INFORMATION

Sambit Bhattacharya
Fayetteville State University
Fayetteville, NC
sbhattac@uncfsu.edu

10. ACKNOWLEDGMENT

The authors acknowledge support in scene authoring, synthetic data pipeline development, and integration of multi-feature association components.

11. ACRONYMS

CLIP	Contrastive Language-Image Pretraining
DINOv2	Self-supervised visual foundation model (v2)
GVSC	Ground Vehicle Systems Center
ICR	Identity Consistency Rate
ISR	ID-Switch Rate
LoFTR	Local Feature Transformer
SAM	Segment Anything Model
USD	Universal Scene Description