WHERE FROM ABOVE, WHAT FROM THE STREET: LEARNING SEMANTIC ROUTING FOR CROSS-VIEW SEGMENTATION UNDER POSE UNCERTAINTY

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-view semantic segmentation aims to integrate satellite and street-view imagery to produce fine-grained urban semantic maps by exploiting complementary observations across perspectives. Satellite imagery provides a stable spatial reference for organizing predictions, while street-view imagery contributes additional fine-grained semantic evidence complementary to overhead observations. Existing approaches typically assume accurate spatial registration between the two views and exploit this correspondence for cross-view feature alignment and fusion. However, such accurate registration is difficult to guarantee in practical scenarios, as camera locations and orientations may contain errors. We argue that exploiting street-view semantics does not require committing semantic transfer to a single precise geometric correspondence; instead, the key is to learn how semantic evidence should be associated with spatial entities in the satellite representation. Based on this perspective, we formulate cross-view fusion as a semantic routing problem and propose CVRoute, which learns to route street-view semantic evidence to relevant satellite regions under uncertain geometry. Specifically, CVRoute models satellite regions and street-view observations as cross-view entities and learns their semantic relationships for robust evidence transfer, with a training-only privileged relationship teacher guiding early optimization through curriculum learning. Experiments on OmniCity and Brooklyn show that CVRoute improves overall segmentation performance and exhibits substantially smaller degradation as pose perturbations increase, reducing the dependence on precise spatial alignment.

1 INTRODUCTION

Fine-grained semantic understanding of urban environments supports applications ranging from urban mapping and land-use analysis to 3D city modeling and autonomous systems (Li et al., 2023; Zheng et al., 2020; Yan & Huang, 2022). Satellite imagery provides broad and spatially consistent coverage and has been widely used for geospatial segmentation and building understanding (Zheng et al., 2020; He et al., 2024; Sikdar et al., 2023), whereas street-view imagery directly observes facades and vertical structures that are weakly visible from overhead views (Yan & Huang, 2022; Fan et al., 2024). These complementary sensing characteristics motivate *cross-view semantic segmentation*, which combines satellite and street-view observations to obtain fine-grained semantic maps beyond what either perspective can provide alone (Zhai et al., 2017; Li et al., 2023; Ye et al., 2024).

Existing cross-view semantic segmentation approaches commonly transfer information between views by establishing an explicit spatial relationship before feature fusion. SG-BEV, for example, converts street-view features into a bird's-eye-view representation and further uses satellite-guided reprojection for fine-grained building segmentation (Ye et al., 2024). More broadly, view-to-BEV and view-to-map perception has been extensively studied through geometry-based lifting, BEV queries, and camera-aware cross-view attention (Philion & Fidler, 2020; Li et al., 2024b; Zhou & Krähenbühl, 2022). These approaches demonstrate the value of organizing information from different perspectives in a shared spatial frame, but their spatial transfer mechanisms naturally depend on the geometric relationship between the input observations and the target representation.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

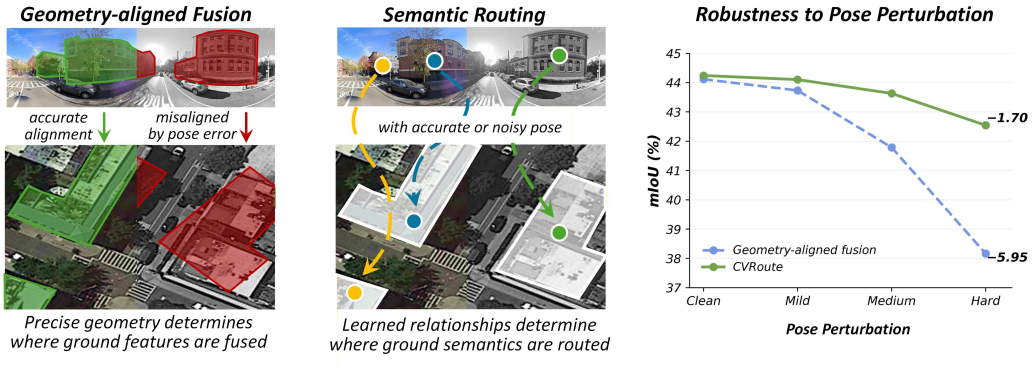

Figure 1: Motivation for semantic routing under pose uncertainty. Pose errors weaken the spatial correspondence used by alignment-based cross-view fusion, motivating CVRoute to associate street-view semantic evidence with satellite regions through learned cross-view relationships.

In practical cross-view settings, however, accurate spatial correspondence cannot always be assumed. A long line of ground-to-aerial geo-localization research has documented the difficulty caused by large viewpoint changes, unknown orientation, and imperfectly aligned cross-view observations (Lin et al., 2015; Workman et al., 2015; Vo & Hays, 2016; Hu et al., 2018; Liu & Li, 2019; Zhu et al., 2021). More recent metric localization methods estimate fine-grained cross-view pose when the geometric relationship is unavailable or only coarsely specified (Xia et al., 2021; 2022; Shi & Li, 2022; Lentsch et al., 2023; Xia et al., 2023). Such imperfect geometric information can weaken the reliability of correspondence-driven feature transfer. More fundamentally, we argue that exploiting street-view semantics for cross-view segmentation does not require committing semantic transfer to a single precise geometric correspondence. Since the spatial organization of the prediction space can be anchored by the satellite representation, street-view imagery can instead provide complementary semantic evidence that is associated with relevant spatial regions.

Based on this observation, we reformulate cross-view fusion as a semantic routing problem, where the key question is how to route street-view semantic evidence to appropriate spatial regions in the satellite representation under uncertain geometry. Accordingly, CVRoute models satellite regions and street-view observations as cross-view entities and learns their semantic relationships for robust evidence routing. Unlike alignment-based fusion that warps ground features into overhead coordinates, CVRoute keeps the satellite-defined spatial support intact and transfers only complementary ground-view semantic evidence through learned entity relationships. To facilitate reliable routing learning, CVRoute introduces a training-only privileged relationship teacher that guides early optimization and gradually enables routing based on predicted relationships through curriculum learning. Figure 1 summarizes the motivation for this formulation under pose uncertainty.

In summary, this work makes three main contributions: 1) We reformulate cross-view fusion from direct geometry-aligned feature transfer into semantic routing under uncertain geometry. 2) We propose CVRoute, which models satellite regions and street-view observations as cross-view entities and learns their semantic relationships to route street-view evidence to relevant satellite regions. 3) We introduce a training-only privileged relationship teacher with curriculum learning to facilitate routing learning. Extensive experiments on OmniCity and Brooklyn demonstrate the effectiveness and robustness of CVRoute under increasing pose perturbations.

2 Related Work

2.1 Single-view Urban Building Understanding

Urban building understanding has been studied from both overhead and street-level imagery. Satellite-based methods extract building footprints and other geospatial semantics at scale (Li et al., 2018; Zheng et al., 2020; Sikdar et al., 2023), and recent work further targets fine-grained building attributes (He et al., 2024). Street-view imagery provides complementary facade and vertical-structure

cues for estimating properties such as building height (Yan & Huang, 2022; Fan et al., 2024), while large vision-language models have also been explored for building-attribute extraction (Pan et al., 2024). These studies motivate combining overhead spatial support with ground-level semantic evidence.

2.2 Cross-view Semantic Segmentation

Cross-view semantic transfer has been explored through learned view transformations and cross-view synthesis (Zhai et al., 2017; Regmi & Borji, 2018; Toker et al., 2021), while BEV perception methods map camera features into canonical top-down representations through geometric lifting, BEV queries, or cross-view attention (Philion & Fidler, 2020; Li et al., 2024b; Zhou & Krähenbühl, 2022). For urban satellite–street understanding, OmniCity provides a large-scale multi-view benchmark (Li et al., 2023), and SG-BEV performs street-view-to-BEV transformation and satellite-guided reprojection for fine-grained building segmentation (Ye et al., 2024). These methods demonstrate the value of cross-view fusion, but largely rely on a reliable geometric relationship for spatial transfer; our work instead focuses on semantic fusion under pose uncertainty.

2.3 Cross-view Correspondence

Cross-view geo-localization provides a broad foundation for reasoning between ground and overhead observations. Early methods learned shared representations for image-level cross-view retrieval (Lin et al., 2015; Workman et al., 2015), followed by approaches that improve matching by modeling orientation, domain gaps, hard negatives, and spatial structure (Vo & Hays, 2016; Hu et al., 2018; Liu & Li, 2019; Cai et al., 2019; Regmi & Shah, 2019; Shi et al., 2019). Transformer-based architectures and improved training strategies further strengthen cross-view retrieval (Yang et al., 2021; Zhu et al., 2022; Deuser et al., 2023; Li et al., 2024a), while VIGOR relaxes the conventional one-to-one correspondence assumption by allowing ground queries to occur at arbitrary positions relative to satellite references (Zhu et al., 2021). Beyond image-level retrieval, more recent work estimates metric position or full 3-DoF pose within a local overhead region using coarse spatial priors, dense matching, or explicit pose reasoning (Xia et al., 2020; 2021; 2022; Shi & Li, 2022; Lentsch et al., 2023; Shi et al., 2023; Xia et al., 2023; Yuan et al., 2024; Wang et al., 2024). Cross-view association has also been studied at finer granularities, including object-level geo-localization and local-feature correspondence (Sun et al., 2023; Zhang & Zhu, 2025; Lv et al., 2026; Xia & Alahi, 2025; Xia et al., 2026). Across these settings, precise cross-view correspondence is not directly given a priori, but must be inferred when geometric information is unavailable, only coarsely specified, or affected by pose error. CVRoute builds on this general perspective, using learned cross-view relationships for semantic evidence transfer.

3 Method

3.1 Problem Formulation and Framework Overview

We consider cross-view building-attribute segmentation under uncertain camera pose. Given a satellite image $\mathbf{I}^s$ and an associated street-view panorama $\mathbf{I}^g$, the model predicts a pixel-level building-attribute map $\hat{\mathbf{Y}}$ in the satellite coordinate system. Under the pose-uncertain protocol, the model is provided with an approximate 3-DoF pose prior $\tilde{\mathbf{p}} = (\tilde{x}, \tilde{y}, \tilde{\theta})$, which may contain translation and orientation errors and therefore does not define a precise cross-view correspondence.

As illustrated in Fig. 2, CVRoute follows an asymmetric but complementary organization of the two views. The satellite branch provides spatial regions that anchor where predictions are placed, while the street-view branch extracts fine-grained semantic evidence describing what is observed from the ground. Instead of relying on a single deterministic geometric mapping, CVRoute learns relationships between entities from the two views and uses them to route street-view semantics to relevant satellite regions. The routed evidence is then combined with satellite-side semantic information and decoded into the final satellite-space segmentation map.

The asymmetry concerns their functional roles rather than exclusive information content: satellite entities also retain semantic cues, while ground entities carry geometric cues. The final spatial

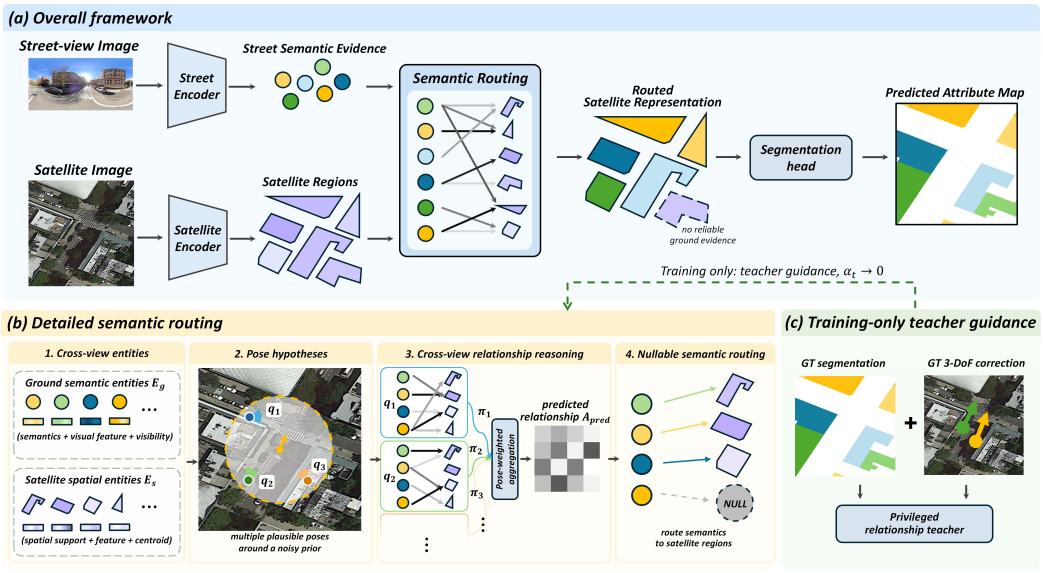

Figure 2: Overview of CVRoute. (a) Street-view semantics are routed to satellite-defined regions for satellite-space prediction. (b) Cross-view entities are matched over multiple pose hypotheses and fused through pose-marginalized nullable routing. (c) A training-only privileged relationship teacher guides early routing and is gradually annealed out; inference uses only predicted relationships.

organization remains anchored by the satellite representation, whereas cross-view semantic transfer is governed by the learned entity relationships.

3.2 Entity-level Semantic Routing

The goal of semantic routing is to determine which street-view semantic evidence should contribute to each satellite region when the geometric relationship between the two views is uncertain. To this end, we first represent both views with a set of entities, then estimate their relationships over multiple plausible pose hypotheses, and finally route ground semantics to satellite entities through a soft assignment.

Cross-view entity representations. For the satellite image $\mathbf{I}^s$, a multi-scale encoder extracts overhead features and organizes them into a set of N_s satellite entities, $\mathcal{S} = \{s_i\}_{i=1}^{N_s}$. Each entity is represented by an appearance token $\mathbf{f}_i^s$, a centroid $\mathbf{c}_i$, an entity confidence o_i, and a soft spatial assignment map $\mathbf{M}_i$. The assignment map indicates the degree to which each satellite pixel belongs to entity s_i, allowing the learned entities to softly organize the satellite foreground into candidate spatial regions. These regions provide the spatial support for subsequent cross-view routing and final semantic prediction.

For the street-view panorama $\mathbf{I}^g$, the ground branch produces a set of N_g semantic entities, $\mathcal{G} = \{g_j\}_{j=1}^{N_g}$. Each ground entity contains a feature token $\mathbf{f}_j^g$, building-attribute logits $\mathbf{l}_j^g$, a visibility score v_j, and a horizontal direction distribution $\mathbf{a}_j$. This distribution describes the probability of the entity appearing along different horizontal directions in the panorama and provides a geometric cue for evaluating its compatibility with satellite entities under different candidate camera poses.

Relationship estimation over pose hypotheses. Because the approximate pose prior can be inaccurate, we do not construct cross-view relationships from a single pose. Instead, we generate a set of candidate pose hypotheses $\mathcal{P} = \{\mathbf{p}_q\}_{q=1}^{Q}$ around the approximate pose prior. Under each hypothesis $\mathbf{p}_q$, the centroid $\mathbf{c}_i$ of satellite entity s_i determines an expected horizontal direction in the panorama. The matcher then evaluates the compatibility between satellite entity s_i and ground entity g_j using geometric, content, cyclic-relation, class-compatibility, and confidence evidence. We

denote the resulting pairwise score under pose hypothesis q by

$$e_{ij}^{q} = G_{ij}^{q} + C_{ij} + 0.5R_{ij}^{q} + K_{ij} + \log o_i + \log v_j, \tag{1}$$

where G_{ij}^{q}, C_{ij}, R_{ij}^{q}, and K_{ij} denote the geometric, content, cyclic-relation, and class-compatibility terms, respectively. Their implementation details are provided in Appendix A.4.

For each pose hypothesis, the pairwise scores are converted into a capacity-constrained soft association $\mathbf{A}^{q}$ using an unbalanced Sinkhorn-style transport with an explicit null state. In parallel, the matcher aggregates the entity-level evidence under each hypothesis into a hypothesis score and normalizes these scores across the candidate set to obtain a pose posterior $\{\pi_q\}_{q=1}^{Q}$. Rather than selecting a single pose, we marginalize the associations over this posterior:

$$\mathbf{A}^{\mathrm{pred}} = \sum_{q=1}^{Q} \pi_q \mathbf{A}^{q}, \qquad \sum_{q=1}^{Q} \pi_q = 1. \tag{2}$$

This procedure allows the predicted relationship to account for pose ambiguity instead of committing to one potentially incorrect geometric configuration.

Semantic routing with null assignment. Each per-hypothesis association includes a null state for satellite entities without sufficiently reliable ground evidence. The capacity constraint discourages many satellite entities from collapsing onto the same small set of ground entities, while the null state allows uncertain entities to abstain from receiving ground-view semantics. Importantly, null represents an uncertain or unmatched cross-view relationship rather than background or the absence of semantic information.

Given the routing matrix $\mathbf{A}^{\mathrm{route}}$, the ground semantic evidence transferred to satellite entity s_i is

$$\mathbf{g}_i = \sum_{j=1}^{N_g} A_{ij}^{\mathrm{route}} \mathbf{l}_j^{g}. \tag{3}$$

The routed evidence is then combined with a satellite semantic prior $\mathbf{p}_i^{s}$:

$$\mathbf{u}_i = \mathbf{g}_i + \mathbf{p}_i^{s}. \tag{4}$$

Therefore, when an entity receives little or no reliable ground evidence, its prediction can still rely on satellite-side semantics rather than being treated as semantically missing. Finally, the entity semantics are written back through the satellite spatial assignments and decoded into the pixel-level building-attribute map. The construction of $\mathbf{A}^{\mathrm{route}}$ during training is described in Sec. 3.3; at inference, it is given solely by $\mathbf{A}^{\mathrm{pred}}$.

3.3 Privileged Relationship Curriculum

Early in training, the predicted cross-view relationships are unreliable, and erroneous associations can route ground semantics to incorrect satellite regions and consequently corrupt the supervision received by downstream prediction layers. We therefore introduce a training-only privileged relationship curriculum that provides reliable routing guidance during early optimization and gradually transfers the routing decision to the predicted cross-view relationships.

During training, the known 3-DoF pose correction and the satellite segmentation annotation are used to construct a soft relationship teacher. We first extract foreground components from the satellite annotation and obtain their spatial centroids. Given the corrected camera pose, each component is associated with ground entities according to its expected horizontal direction in the panorama, together with the predicted ground visibility. This produces a soft teacher relationship that includes both real ground entities and a null option. Because the teacher is initially defined on annotation-derived components whereas semantic routing operates on predicted satellite entities, we use detached mask-IoU bipartite matching to transfer the component-level relationships to the corresponding predicted entities. The resulting teacher association is denoted by $\mathbf{A}^{\mathrm{tea}}$ and is detached from gradient propagation.

Rather than directly distilling $\mathbf{A}^{\mathrm{tea}}$ into the predicted association, we use the teacher only to guide the routing path. At training epoch t, the routing matrix is

$$\mathbf{A}_t^{\mathrm{route}} = \alpha_t \mathbf{A}^{\mathrm{tea}} + (1 - \alpha_t)\mathbf{A}^{\mathrm{pred}}, \tag{5}$$

where α_t is gradually annealed from 1 to 0. In our implementation, α_t decreases from 1.0 to 0.2 over the first five epochs and becomes 0 from the sixth epoch onward. Therefore, early training benefits from privileged routing guidance, whereas later optimization and inference rely entirely on the model-predicted relationships. No explicit loss is applied between $\mathbf{A}^{\mathrm{pred}}$ and $\mathbf{A}^{\mathrm{tea}}$.

3.4 Training Objectives and Inference

CVRoute is optimized with the satellite-space segmentation objective together with auxiliary geometry, semantic, and pose-consistency supervision. The complete objective and effective loss weights are provided in Appendix A.3. At inference, all privileged training information is removed: the model receives only the satellite image, street-view panorama, and approximate pose prior, with $\mathbf{A}^{\mathrm{route}} = \mathbf{A}^{\mathrm{pred}}$, and produces the final segmentation in a single forward pass.

4 Experiments

4.1 Experimental Setup

Datasets and Tasks. We evaluate CVRoute on two urban cross-view building understanding benchmarks, including Brooklyn and OmniCity (Ye et al., 2024; Li et al., 2023). Both benchmarks provide paired satellite and street-view imagery with building-level attributes. We consider two building attribute segmentation tasks: floor-level prediction and land-use classification. Following previous cross-view semantic segmentation studies (Ye et al., 2024), each task is formulated as pixel-level prediction in the satellite coordinate system, where the model generates an attribute map aligned with the overhead image.

Pose Uncertainty Protocol. Unlike conventional cross-view segmentation methods that assume accurate spatial alignment, we evaluate methods under uncertain camera poses. For each sample, we perturb the 3-DoF pose prior by independently sampling yaw and planar translation errors. Clean uses no additional perturbation (0 m translation and 0° yaw). For Mild, yaw is sampled uniformly from $[-10^\circ, 10^\circ]$, while the two planar translation components are independently sampled from $[-4, 4]$ m. Medium uses yaw in $[-25^\circ, 25^\circ]$ and translation components in $[-10, 10]$ m, and Hard uses yaw in $[-45^\circ, 45^\circ]$ and translation components in $[-20, 20]$ m. Thus, each non-clean sample receives translation and yaw perturbations simultaneously rather than from separate perturbation families.

The perturbations are generated once using a fixed random seed and stored in JSON manifests. The same validation manifests are reused for ST, GP, SG-BEV†, and CVRoute, so corresponding samples receive identical pose perturbations across methods. During CVRoute training, fixed pose perturbations are introduced to expose semantic routing to uncertain geometry. The final model is evaluated on the Clean, Mild, Medium, and Hard validation manifests.

During training, the inverse synthetic correction $(-\Delta\theta, -\Delta x, -\Delta y)$ is available only through a detached training context to construct the privileged relationship teacher. At inference, the model receives only the perturbed pose prior together with the satellite image and street-view panorama; the true correction, ground-truth component assignment, and teacher routing are not provided.

Evaluation Metrics. We report mean Intersection-over-Union (mIoU) as the primary evaluation metric. To analyze robustness under different levels of pose uncertainty, each model is evaluated under four perturbation settings: Clean, Mild, Medium, and Hard. We additionally report Avg4, computed as the average mIoU across the four perturbation levels.

Implementation Details. All CVRoute variants and controlled ablations use the same staged training protocol. Each task is initialized from its task-matched geometry checkpoint before optimizing the cross-view relationship and semantic routing modules. All compared methods are evaluated under the same data split and pose-perturbation protocol. During inference, CVRoute performs a single forward pass without test-time pose refinement or access to privileged pose information.

Table 1: Results on Brooklyn under different levels of pose uncertainty.

	Floor					Land-use				
Method	Clean	Mild	Medium	Hard	Avg4	Clean	Mild	Medium	Hard	Avg4
ST	33.13	33.13	33.12	33.12	33.12	36.96	36.96	36.96	36.96	36.96
GP	42.66	42.29	40.71	37.87	40.88	40.21	40.03	39.29	37.98	39.38
SG-BEV†	44.12	43.73	41.79	38.17	41.95	41.64	41.42	40.26	38.19	40.38
CVRoute	**44.24**	**44.10**	**43.63**	**42.54**	**43.63**	**42.39**	**42.16**	**42.04**	**41.67**	**42.06**

Table 2: Results on OmniCity under different levels of pose uncertainty.

	Floor					Land-use				
Method	Clean	Mild	Medium	Hard	Avg4	Clean	Mild	Medium	Hard	Avg4
ST	25.22	25.22	25.21	25.20	25.21	30.76	30.76	30.76	30.76	30.76
GP	27.57	27.28	26.40	25.30	26.64	**32.73**	**32.29**	31.12	29.80	31.49
SG-BEV†	29.72	29.36	28.09	26.40	28.39	32.69	32.20	30.89	29.42	31.30
CVRoute	**30.39**	**29.94**	**29.68**	**29.07**	**29.77**	32.14	32.07	**31.89**	**31.54**	**31.91**

4.2 Comparison with Existing Methods

Following the baseline setting of SG-BEV (Ye et al., 2024), we compare CVRoute with Spherical Transform (ST), Geometric Projection (GP), and SG-BEV itself. The public repository does not provide a usable implementation or pretrained checkpoint of the full SG-BEV model for our pose-perturbation evaluation, so we reproduce its core pipeline following the paper and supplementary material; our reproduction is denoted by †. Tables 1 and 2 report the results on Brooklyn and OmniCity, respectively.

Overall, CVRoute achieves more robust performance under increasing pose perturbations compared with existing alignment-based fusion methods. While existing approaches rely on explicit spatial alignment or view transformation to associate information between the two views, their performance gradually degrades as the nominal pose becomes inaccurate. In contrast, CVRoute learns entity-level semantic relationships to route ground-view evidence to satellite regions, enabling more stable cross-view transfer under uncertain correspondence.

On the Brooklyn benchmark, CVRoute obtains the best performance on both floor and land-use prediction. In particular, for the floor task, CVRoute improves the Avg4 mIoU from 41.95% of SG-BEV† to 43.63%, with a larger improvement under severe pose perturbations. Under the Hard setting, CVRoute achieves 42.54% mIoU, outperforming SG-BEV† by 4.37 percentage points, demonstrating the advantage of semantic routing when geometric alignment becomes unreliable.

Similar trends can be observed on OmniCity. CVRoute maintains a smaller performance drop from Clean to Hard settings compared with cross-view fusion baselines. For example, in the OmniCity floor task, CVRoute decreases by 1.32 percentage points (from 30.39% to 29.07%) as pose uncertainty increases, whereas SG-BEV† decreases by 3.32 percentage points (from 29.72% to 26.40%). These results indicate that CVRoute is less sensitive to pose errors and can better preserve useful cross-view semantic evidence under imperfect alignment.

4.3 Understanding the Semantic Routing Mechanism

Dependence on ground-view semantic evidence. To verify whether CVRoute genuinely benefits from paired street-view semantic information, we evaluate the model with different panorama inputs. Specifically, we compare the original paired panorama with two degraded settings: shuffled panoramas and zeroed panoramas. As shown in Table 3, replacing the paired panorama with a shuffled one reduces the mIoU from 43.63% to 19.94%, while removing the panorama entirely further decreases the performance to 17.75%. These results demonstrate that CVRoute relies on the semantic evidence provided by the matched street-view observation rather than solely depending on satellite-side information.

Table 3: Effect of ground-view semantic evidence under different panorama inputs.

Panorama input	mIoU (%)
Paired panorama	43.63
Shuffled panorama	19.94
Zero panorama	17.75

Table 4: Analysis of satellite-entity coverage under uncertain pose.

Method	Component recall@0.5 (%)	GT-to-slot recall@0.5 (%)
w/o teacher-guided routing	24.55	23.89
CVRoute	**36.87**	**35.24**

Table 5: Ablation study of teacher-guided routing and object contrastive learning.

Method	Teacher-guided routing	Object contrastive loss	Avg4 (%)
w/o teacher-guided routing	×	×	41.73
w/ object contrastive loss	×	✓	41.80
CVRoute	✓	×	**43.63**
w/ additional contrastive loss	✓	✓	43.57

Entity coverage supporting semantic routing. Beyond final segmentation accuracy, we examine whether teacher-guided training improves the spatial coverage of the satellite entities that serve as the routing targets. As reported in Table 4, CVRoute improves foreground component recall@0.5 from 24.55% to 36.87% and GT-to-slot recall@0.5 from 23.89% to 35.24%. These results indicate that teacher-guided training provides better coverage of foreground regions by satellite entities, supplying a stronger spatial substrate for subsequent cross-view semantic routing.

4.4 Ablation Study of Teacher-guided Routing

We conduct ablation experiments to investigate the contribution of training-only teacher supervision and object contrastive learning. The results are summarized in Table 5.

Removing teacher-guided routing leads to a clear performance degradation, reducing the Avg4 mIoU from 43.63% to 41.73%. In contrast, introducing object contrastive learning alone provides only marginal improvement, suggesting that contrastive regularization alone is insufficient to resolve correspondence ambiguity under pose uncertainty.

Interestingly, combining teacher-guided routing with object contrastive learning does not further improve the performance compared with CVRoute. This indicates that privileged relationship supervision already provides effective guidance for learning semantic routing, while additional contrastive optimization is not necessary. Therefore, CVRoute adopts the simpler teacher-guided routing strategy without object contrastive loss.

4.5 Qualitative Results

Figure 3 qualitatively compares the robustness of SG-BEV† and CVRoute under increasing pose uncertainty. As the pose perturbation increases from Clean to Hard, SG-BEV† exhibits progressively stronger degradation in both spatial structure and semantic prediction. The predicted regions become increasingly distorted or displaced, while their semantic labels also become less consistent. In contrast, CVRoute largely preserves the spatial extent and boundaries of the predicted regions across different perturbation levels. Although some semantic variations remain under stronger pose noise, the overall spatial structure is substantially more stable. These examples are consistent with the quantitative robustness results and illustrate the benefit of reducing the dependence of semantic transfer on precise geometric alignment. Figure 4 visualizes the learned semantic routing mechanism. Given a street-view observation, the predicted routing map assigns its semantic evidence to relevant regions in the satellite representation rather than relying on a fixed geometric projection. In both examples, the routing responses concentrate on the satellite regions associated with the observed buildings, and the resulting CVRoute predictions recover semantic regions that are consistent with the ground truth. Importantly, the routing map is not itself a segmentation prediction; instead, it determines where ground-view semantic evidence is transferred before satellite-space decoding. These examples illustrate the central role of learned cross-view relationships in connecting street-level semantics with satellite-defined spatial structure.

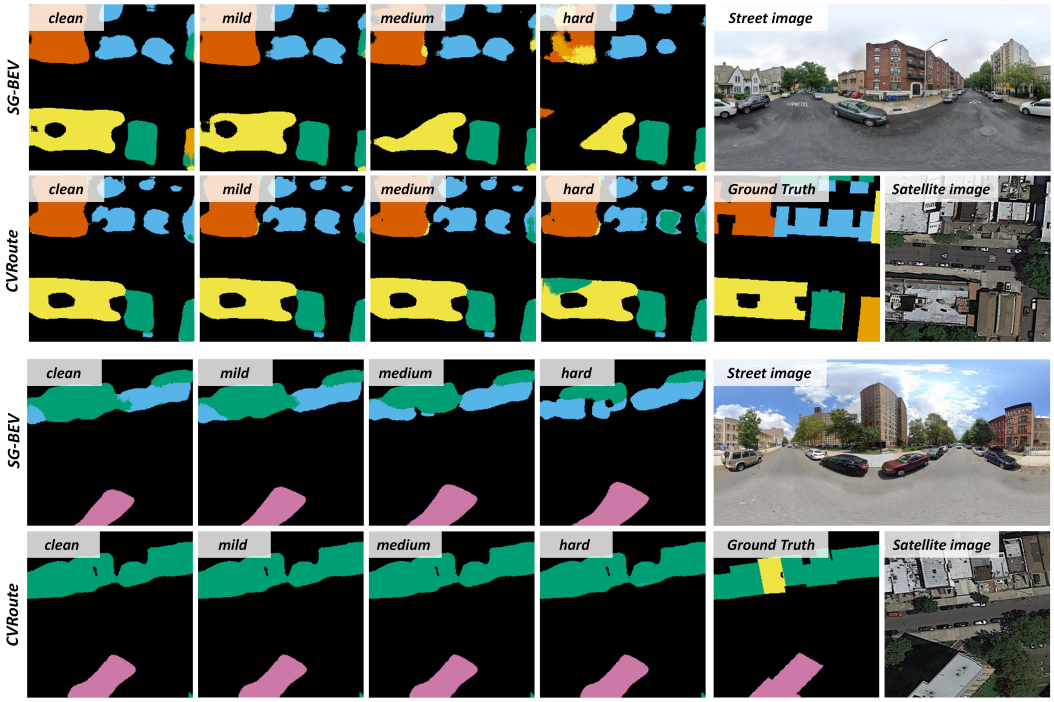

Figure 3: Qualitative comparison between SG-BEV† and CVRoute under increasing pose uncertainty. CVRoute preserves more stable spatial structures and semantic predictions as the pose perturbation increases.

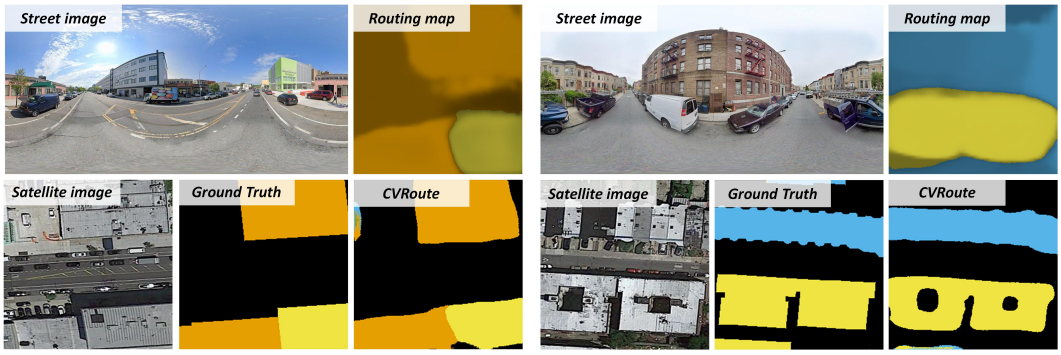

Figure 4: Visualization of the learned semantic routing in CVRoute. The routing maps indicate where semantic evidence from the street-view observation is transferred within the satellite representation before final segmentation.

5 Conclusion

In this work, we propose CVRoute for cross-view semantic segmentation under uncertain camera poses. CVRoute routes ground-view semantic evidence to satellite regions through learned entity-level relationships, guided during training by a privileged relationship teacher. Experiments on Brooklyn and OmniCity demonstrate improved robustness to pose perturbations and confirm the effective use of ground-view semantic evidence.

Limitations and Future Work. CVRoute remains challenged by small, densely arranged regions, where neighboring structures may be merged or missed. It also relies on privileged pose correction to construct the training-time teacher. Future work will explore finer-grained routing and weaker supervision.

AI USE STATEMENT

In this work, we used generative AI tools to assist with research-methodology and experiment-design feedback, implementation and debugging, result interpretation, and manuscript writing and revision. We additionally used generative AI tools to edit scientific figures, search and organize literature, and format citations and tables. All AI-assisted outputs were reviewed by the authors, and AI-generated suggestions were accepted only after checking them against project source code, experimental artifacts, or cited papers, as applicable. We take responsibility for the final content of this work, including text, claims, code, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Section 3 defines CVRoute, including the cross-view entity representations, pose-marginalized semantic routing, privileged relationship curriculum, training objectives, and inference procedure. Section 4.1 specifies the datasets and tasks, fixed pose-uncertainty manifests, evaluation metrics, and staged training protocol. Appendix A documents the training configuration, construction of the privileged teacher, and optimization details, while Appendices B and C provide additional quantitative and qualitative analyses. We plan to include anonymized configuration and evaluation scripts with the supplementary material.

REFERENCES

Sudong Cai, Yulan Guo, Salman Khan, Jiwei Hu, and Gongjian Wen. Ground-to-aerial image geo-localization with a hard exemplar reweighting triplet loss. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8390–8399. IEEE, 2019.

Fabian Deuser, Konrad Habel, and Norbert Oswald. Sample4geo: Hard negative sampling for cross-view geo-localisation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16801–16810. IEEE, 2023.

Kaixuan Fan, Anqi Lin, Hao Wu, and Zhenci Xu. Pano2geo: An efficient and robust building height estimation model using street-view panoramas. *ISPRS Journal of Photogrammetry and Remote Sensing*, 215:177–191, 2024.

Zhiyi He, Wei Yao, Jie Shao, and Puzuo Wang. Ub-finenet: Urban building fine-grained classification network for open-access satellite images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 217:76–90, 2024.

Sixing Hu, Mengdan Feng, Rang MH Nguyen, and Gim Hee Lee. Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7258–7267. IEEE, 2018.

Ted Lentsch, Zimin Xia, Holger Caesar, and Julian FP Kooij. Slicematch: Geometry-guided aggregation for cross-view pose estimation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17225–17234. IEEE, 2023.

Guopeng Li, Ming Qian, and Gui-Song Xia. Unleashing unlabeled data: A paradigm for cross-view geo-localization. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16719–16729. IEEE, 2024a.

Weijia Li, Conghui He, Jiarui Fang, and Haohuan Fu. Semantic segmentation based building extraction method using multi-source gis map datasets and satellite imagery. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 233–2333. IEEE, 2018.

Weijia Li, Yawen Lai, Linning Xu, Yuanbo Xiangli, Jinhua Yu, Conghui He, Gui-Song Xia, and Dahua Lin. Omnicity: Omnipotent city understanding with multi-level and multi-view images. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17397–17407. IEEE, 2023.

Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3):2020–2036, 2024b.

Tsung-Yi Lin, Yin Cui, Serge Belongie, and James Hays. Learning deep representations for ground-to-aerial geolocalization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5007–5015, 2015.

Liu Liu and Hongdong Li. Lending orientation to neural networks for cross-view geo-localization. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5617–5626. IEEE, 2019.

Bo Lv, Qingwang Zhang, Le Wu, Yuanyuan Li, and Yingying Zhu. Mogeo: Beyond one-to-one cross-view object geo-localization. *arXiv preprint arXiv:2603.13843*, 2026.

Fei Pan, Sangryul Jeon, Brian Wang, Frank Mckenna, and Stella X Yu. Zero-shot building attribute extraction from large-scale vision and language models. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 8632–8641. IEEE, 2024.

Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *European conference on computer vision*, pp. 194–210. Springer, 2020.

Krishna Regmi and Ali Borji. Cross-view image synthesis using conditional gans. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3501–3510. IEEE, 2018.

Krishna Regmi and Mubarak Shah. Bridging the domain gap for ground-to-aerial image matching. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 470–479. IEEE, 2019.

Yujiao Shi and Hongdong Li. Beyond cross-view image retrieval: Highly accurate vehicle localization using satellite image. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16989–16999. IEEE, 2022.

Yujiao Shi, Liu Liu, Xin Yu, and Hongdong Li. Spatial-aware feature aggregation for image based cross-view geo-localization. *Advances in Neural Information Processing Systems*, 32, 2019.

Yujiao Shi, Fei Wu, Akhil Perincherry, Ankit Vora, and Hongdong Li. Boosting 3-dof ground-to-satellite camera localization accuracy via geometry-guided cross-view transformer. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21459–21469. IEEE, 2023.

Aniruddh Sikdar, Sumanth Udupa, Prajwal Gurunath, and Suresh Sundaram. Deepmao: Deep multi-scale aware overcomplete network for building segmentation in satellite imagery. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 487–496. IEEE, 2023.

Yuxi Sun, Yunming Ye, Jian Kang, Ruben Fernandez-Beltran, Shanshan Feng, Xutao Li, Chuyao Luo, Puzhao Zhang, and Antonio Plaza. Cross-view object geo-localization in a local region with satellite imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–16, 2023.

Aysim Toker, Qunjie Zhou, Maxim Maximov, and Laura Leal-Taixé. Coming down to earth: Satellite-to-street view synthesis for geo-localization. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6484–6493. IEEE, 2021.

Nam N Vo and James Hays. Localizing and orienting street views using overhead imagery. In *European conference on computer vision*, pp. 494–509. Springer, 2016.

Shan Wang, Chuong Nguyen, Jiawei Liu, Yanhao Zhang, Sundaram Muthu, Fahira Afzal Maken, Kaihao Zhang, and Hongdong Li. View from above: Orthogonal-view aware cross-view localization. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14843–14852. IEEE, 2024.

Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-area image geolocalization with aerial reference imagery. In *Proceedings of the IEEE international conference on computer vision*, pp. 3961–3969, 2015.

Zimin Xia and Alexandre Alahi. Fg2: Fine-grained cross-view localization by fine-grained feature matching. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6362–6372. IEEE, 2025.

Zimin Xia, Olaf Booij, Marco Manfredi, and Julian FP Kooij. Geographically local representation learning with a spatial prior for visual localization. In *European Conference on Computer Vision*, pp. 557–573. Springer, 2020.

Zimin Xia, Olaf Booij, Marco Manfredi, and Julian FP Kooij. Cross-view matching for vehicle localization by learning geographically local representations. *IEEE Robotics and Automation Letters*, 6(3):5921–5928, 2021.

Zimin Xia, Olaf Booij, Marco Manfredi, and Julian FP Kooij. Visual cross-view metric localization with dense uncertainty estimates. In *European Conference on Computer Vision*, pp. 90–106. Springer, 2022.

Zimin Xia, Olaf Booij, and Julian FP Kooij. Convolutional cross-view pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):3813–3831, 2023.

Zimin Xia, Chenghao Xu, and Alexandre Alahi. Loc2: Interpretable cross-view localization via depth-lifted local feature matching. In *International Conference on Learning Representations*, volume 2026, pp. 105738–105764, 2026.

Yizhen Yan and Bo Huang. Estimation of building height using a single street view image via deep neural networks. *ISPRS Journal of Photogrammetry and Remote Sensing*, 192:83–98, 2022.

Hongji Yang, Xiufan Lu, and Yingying Zhu. Cross-view geo-localization with layer-to-layer transformer. *Advances in Neural Information Processing Systems*, 34:29009–29020, 2021.

Junyan Ye, Qiyan Luo, Jinhua Yu, Huaping Zhong, Zhimeng Zheng, Conghui He, and Weijia Li. Sg-bev: Satellite-guided bev fusion for cross-view semantic segmentation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27748–27757. IEEE, 2024.

Dong Yuan, Frederic Maire, and Feras Dayoub. Cross-attention between satellite and ground views for enhanced fine-grained robot geo-localization. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1238–1245. IEEE, 2024.

Menghua Zhai, Zachary Bessinger, Scott Workman, and Nathan Jacobs. Predicting ground-level scene layout from aerial imagery. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4132–4140. IEEE, 2017.

Qingwang Zhang and Yingying Zhu. Breaking rectangular shackles: Cross-view object segmentation for fine-grained object geo-localization. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8197–8206. IEEE, 2025.

Zhuo Zheng, Yanfei Zhong, Junjue Wang, and Ailong Ma. Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4095–4104. IEEE, 2020.

Brady Zhou and Philipp Krähenbühl. Cross-view transformers for real-time map-view semantic segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13750–13759. IEEE, 2022.

Sijie Zhu, Taojiannan Yang, and Chen Chen. Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5316–5325. IEEE, 2021.

Sijie Zhu, Mubarak Shah, and Chen Chen. Transgeo: Transformer is all you need for cross-view image geo-localization. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1152–1161. IEEE, 2022.

A Additional Implementation Details

A.1 Training Configuration

The input satellite image is resized to 256×256, while the street-view panorama is resized to 512×1024. The prediction is produced in the satellite coordinate system at 256×256. The satellite and street-view branches use separate MSCAN-B encoders without weight sharing. The satellite branch uses 24 learnable queries to organize the overhead foreground into latent spatial entities.

All reported CVRoute results follow a staged warm-start training protocol. Each task is initialized from its task-matched geometry checkpoint before optimizing the semantic-routing and cross-view relationship modules. For the Brooklyn floor setting, the geometry initialization is obtained through 60 epochs of base geometry training, followed by 2 epochs of geometry bootstrapping and 10 epochs of joint geometry training; CVRoute is then fine-tuned for 8 epochs. The same initialization protocol is used for the controlled ablation variants.

The final CVRoute stage is optimized with AdamW using a learning rate of 5×10^{-5}, betas $(0.9, 0.999)$, and weight decay 0.01. We use cosine learning-rate decay with one warm-up epoch and a minimum learning-rate ratio of 0.10. Training uses 8 GPUs with a batch size of 4 per GPU (32 in total), without gradient accumulation, mixed precision, or gradient clipping.

At inference, the model performs a single forward pass using only the satellite image, street-view panorama, and noisy nominal pose. The ground-truth pose correction, teacher relationship, and teacher-predicted routing mixture are not used, and no test-time pose refinement or optimization is performed.

A.2 Teacher Signal Construction

Privileged relationship teacher. The relationship teacher is constructed only during training. We first extract foreground semantic components from the ground-truth satellite-space segmentation and use them as class-agnostic spatial targets. For each component, the ground-truth 3-DoF correction is combined with the nominal pose to recover its training-time camera geometry. Together with the component centroid, the predicted horizontal direction distribution of each ground entity, and its visibility, this geometry is used to construct a soft relationship between the satellite component and the ground entities. A null state is retained when no sufficiently reliable ground evidence is available. The resulting teacher relationship is detached from gradient computation and therefore serves only as a transient routing target.

Mapping teacher relationships to predicted entities. The teacher relationship is initially defined on ground-truth spatial components, whereas semantic routing operates on predicted satellite entities. We therefore match ground-truth components to predicted satellite entities using detached bipartite matching based on mask IoU. The component-level teacher relationship is then transferred to the corresponding predicted entity, while unmatched entities are assigned to the null state. Importantly, the ground-truth segmentation is only used to construct this training-time relationship and is never provided to the model during inference.

Routing curriculum. Let $\mathbf{A}^{\mathrm{tea}}$ and $\mathbf{A}^{\mathrm{pred}}$ denote the teacher and predicted relationships, respectively. During training, semantic evidence is routed using

$$\mathbf{A}_t^{\mathrm{route}} = \alpha_t \mathbf{A}^{\mathrm{tea}} + (1 - \alpha_t)\mathbf{A}^{\mathrm{pred}}, \tag{6}$$

where α_t gradually decreases with the training epoch. We use the following schedule:

$$\alpha_t = \begin{cases} 1.0, & e = 1, \\ 0.8, & e = 2, \\ 0.6, & e = 3, \\ 0.4, & e = 4, \\ 0.2, & e = 5, \\ 0, & e \geq 6. \end{cases} \tag{7}$$

Thus, early training relies primarily on privileged relationships to avoid erroneous semantic transfer, while later training progressively hands routing decisions to the predicted associations. The teacher does not directly impose a correspondence loss on $\mathbf{A}^{\mathrm{pred}}$; instead, it affects training through the routed semantic evidence used by the downstream prediction objectives. At inference, $\alpha_t = 0$, and neither the ground-truth pose correction, ground-truth semantic components, nor $\mathbf{A}^{\mathrm{tea}}$ is available. Prediction therefore relies exclusively on the learned cross-view relationships.

A.3 Optimization Details

For completeness, we report the effective training objectives enabled in the final CVRoute configuration. The primary noisy-pose segmentation objective is

$$\mathcal{L}_{\mathrm{seg}} = \mathrm{MaskedCE}(\mathbf{Y}, \mathbf{Z}) + 0.5\,\mathrm{SoftDice}(\mathbf{Y}, \mathbf{Z}). \tag{8}$$

Training additionally uses a clean-pose segmentation term with weight 0.10 and a clean-to-noisy pose-consistency KL term with weight 0.05 and temperature $T = 1.5$. The support, boundary, and ground-entity semantic losses have weights 0.40, 0.20, and 0.40, respectively. The association-class auxiliary loss has weight 0.25.

The satellite entity geometry objective is applied to three objectizer layers with normalized layer weights $(1/7, 2/7, 4/7)$ and an outer coefficient of 0.10. At each layer,

$$\begin{aligned}\mathcal{L}_{\mathrm{geometry}} = \mathcal{L}_{\text{focal-mask}} + \mathcal{L}_{\mathrm{dice}} + \mathcal{L}_{\mathrm{objectness}} + \mathcal{L}_{\mathrm{partition}} \\ + \mathcal{L}_{\mathrm{center}} + \mathcal{L}_{\mathrm{logarea}} + \mathcal{L}_{\text{cannot-link}} + \mathcal{L}_{\text{boundary-cons}} + 0.5\mathcal{L}_{\mathrm{duplicate}}.\end{aligned} \tag{9}$$

A margin-based paired auxiliary constraint is also enabled with weight 0.10 and margin 0.05. The final configuration does not use object-level contrastive supervision, direct pose-posterior supervision, or a direct distillation loss between $\mathbf{A}^{\mathrm{tea}}$ and $\mathbf{A}^{\mathrm{pred}}$; the privileged teacher affects optimization through the routing curriculum.

A.4 Semantic Routing Implementation

Entity and pose-hypothesis configuration. The satellite branch produces $N_s = 24$ entities, while the ground branch produces $N_g = 8$ semantic entities. Each ground entity predicts 7 building-attribute logits, a continuous visibility probability, and a 64-bin horizontal direction distribution. The pose bank contains $Q = 45$ fixed hypotheses formed by the Cartesian product of yaw offsets $\{-25^\circ, -12.5^\circ, 0^\circ, 12.5^\circ, 25^\circ\}$ and nine planar translation offsets: $(0, 0)$, $(\pm 8, 0)$, $(0, \pm 8)$, and $(\pm 8, \pm 8)$ meters. The same hypothesis bank is used during training and inference.

Relationship scoring and pose marginalization. For each hypothesis, G_{ij}^q compares the expected satellite-to-ground bearing with the 64-bin ground direction distribution using a circular kernel with a 12.5° geometry temperature. C_{ij} is the cosine similarity between normalized satellite and ground entity tokens with temperature 0.25. R_{ij}^q compares the normalized cyclic descriptor $[\cos\Delta, \sin\Delta, \cos 2\Delta, \sin 2\Delta]$ and has weight 0.5. K_{ij} measures class compatibility between satellite and ground class probabilities with temperature 0.50 and unit weight. Satellite objectness and ground visibility enter the edge logit as log-confidence terms.

For each pose hypothesis, per-entity evidence is aggregated using log-sum-exp over the transport logits and then combined by an objectness-weighted average. The pose posterior is obtained with a softmax temperature of 0.5, and the final predicted association is the posterior-weighted average of the per-hypothesis associations.

Nullable transport and semantic writeback. The real association logits have shape 24×8 per hypothesis. We append one null column, yielding a 24×9 transport matrix, and use an unbalanced row-normalized Sinkhorn-style procedure with 8 iterations, capacity multiplier 3.25, null-logit initialization -1.5, and numerical epsilon 10^{-8}. The real-column capacity is proportional to predicted ground visibility.

The satellite semantic prior is predicted from the satellite entity tokens. Its class-mean-centered logits are bounded as $0.35\tanh(\cdot)$ before being added to the routed ground semantic logits. The resulting entity logits are written back to the satellite grid by an assignment-weighted sum over

the normalized foreground entity masks. The object-based prediction is finally blended with the dense prediction through a dense-uncertainty gate bounded between 0.01 and 0.35 (threshold 0.54, temperature 0.04).

A.5 Baseline Reproduction Details

ST and GP. Following SG-BEV (Ye et al., 2024), we use Spherical Transform (ST) and Geometric Projection (GP) as geometry-based cross-view projection baselines. Both explicitly map street-view information toward a top-down representation using the supplied camera geometry, and therefore provide direct references for evaluating alignment-based cross-view fusion.

SG-BEV reproduction. Our SG-BEV† implementation reproduces the core SG-BEV pipeline following the architecture and experimental protocol described in the paper and supplementary material. The public repository does not provide a usable implementation or pretrained checkpoint of the full SG-BEV model required for direct evaluation under our fixed pose-perturbation protocol. Moreover, the original work does not report results for our Mild, Medium, and Hard perturbation manifests. We therefore retrain the reproduced model and evaluate the same checkpoint under Clean, Mild, Medium, and Hard conditions; the Clean setting follows the original nominal-pose protocol.

The reproduction uses MSCAN-B initialized from `mscan_b.pth`, 42 training epochs on 8 GPUs with batch size 4 per GPU, and AdamW with learning rate 6×10^{-5} and weight decay 0.01. The learning rate is decayed by a factor of 0.7 every 10 epochs. We use the SG-BEV train/validation split and its standard image preprocessing with ImageNet normalization. This implementation is an in-repository reproduction of the SG-BEV pipeline rather than a byte-for-byte official implementation.

B Additional Quantitative Results

B.1 Multi-seed Evaluation

To assess the sensitivity of CVRoute to training randomness, we conduct three additional independent runs of the Brooklyn floor experiment using different random seeds. Table 6 reports their performance together with the mean and sample standard deviation. The three runs show consistent performance, with Avg4 ranging from 43.09% to 43.70%. Across the three seeds, CVRoute achieves an Avg4 of $43.48 \pm 0.34\%$, with similarly limited variation across the individual pose perturbation levels.

These results indicate that CVRoute maintains stable performance across different random seeds under the evaluated setting, suggesting that the observed gains are not tied to a particular initialization.

B.2 Detailed Ablation Results

Table 7 expands the ablation study in the main paper by reporting performance under each pose-uncertainty level. Removing teacher-guided routing consistently reduces performance across Clean, Mild, Medium, and Hard conditions, with the Avg4 mIoU decreasing from 43.63% to 41.73%. In contrast, introducing object contrastive learning without teacher-guided routing yields only marginal improvements over the routing-free control.

Adding object contrastive learning on top of teacher-guided routing also does not provide further gains. Its Avg4 performance is 43.57%, slightly below the 43.63% obtained by CVRoute, and the same trend is observed across all four perturbation levels. These results further support our design choice of using teacher-guided semantic routing without an additional object contrastive objective.

C Additional Qualitative Results

C.1 Additional Pose Robustness Cases

Figure 5 provides additional qualitative comparisons between SG-BEV† and CVRoute under increasing pose uncertainty. Consistent with the examples presented in the main paper, SG-BEV†

Table 6: Multi-seed evaluation of CVRoute on the Brooklyn floor task. Mean and sample standard deviation are computed over three training seeds.

Run	Clean	Mild	Medium	Hard	Avg4
Seed 1	44.27	44.15	43.65	42.54	43.65
Seed 2	43.64	43.20	43.17	42.33	43.09
Seed 3	44.34	44.21	43.79	42.46	43.70
Mean	44.08	43.85	43.54	42.44	43.48
Std.	0.39	0.57	0.33	0.11	0.34

Table 7: Detailed ablation results on the Brooklyn floor task under different pose-uncertainty levels.

Teacher routing	Contrastive loss	Clean	Mild	Medium	Hard	Avg4
×	×	42.15	42.05	41.73	40.97	41.73
×	✓	42.20	42.12	41.81	41.07	41.80
✓	×	**44.24**	**44.10**	**43.63**	**42.54**	**43.63**
✓	✓	44.18	44.08	43.59	42.45	43.57

shows progressively stronger degradation as the pose perturbation increases, affecting both the spatial extent of predicted regions and their semantic assignments. In contrast, CVRoute preserves substantially more stable segmentation structures across the different perturbation levels, with only limited semantic variation under stronger pose noise. These additional examples further illustrate that semantic routing reduces the sensitivity of cross-view semantic transfer to inaccurate geometric alignment.

C.2 Additional Semantic Routing Visualization

Figure 6 presents an additional example of the learned semantic routing in CVRoute. Consistent with the visualization in the main paper, the predicted routing map associates the semantic evidence observed from the street-view panorama with spatially relevant regions in the satellite representation. The routed evidence is subsequently combined with the satellite-defined spatial structure to produce the final segmentation, which preserves the overall layout of the target regions while assigning their corresponding semantic attributes.

This example further illustrates that the routing map should not be interpreted as a direct segmentation output. Instead, it represents an intermediate cross-view association that determines where street-view semantic evidence is incorporated into the satellite representation. The final prediction is therefore jointly determined by the learned routing relationship and the spatial support provided by the satellite branch.

C.3 Failure Cases

Figure 7 presents representative failure cases of CVRoute. A recurring difficulty arises when buildings or semantic regions are small and densely arranged. In these scenes, neighboring regions may be merged during prediction, while some small semantic regions are missed or absorbed into nearby dominant categories. This suggests that distinguishing and routing semantic evidence among closely spaced spatial entities remains challenging for the current entity-level representation. Improving the granularity of semantic routing, particularly for small and densely distributed urban structures, is an important direction for future work.

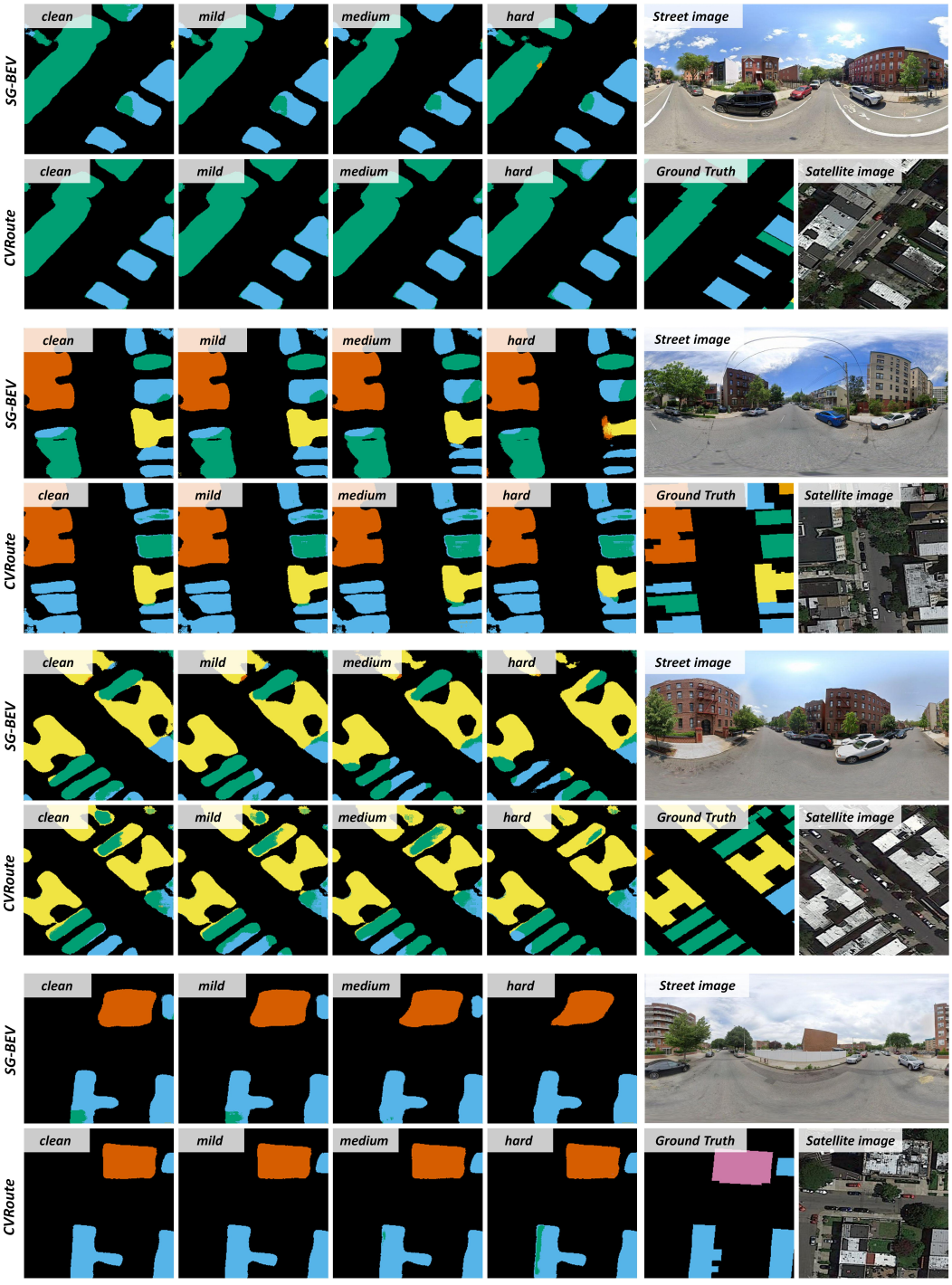

Figure 5: Additional qualitative comparisons between SG-BEV† and CVRoute under increasing pose uncertainty. The four cases complement the examples presented in the main paper and show consistent trends in spatial and semantic robustness.

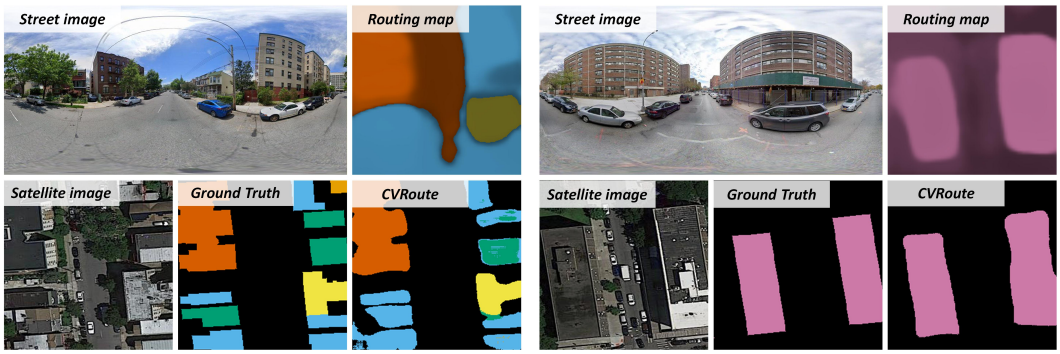

Figure 6: Additional visualization of the learned semantic routing in CVRoute. The routing map indicates the satellite regions to which semantic evidence from the street-view observation is transferred before the final satellite-space segmentation.

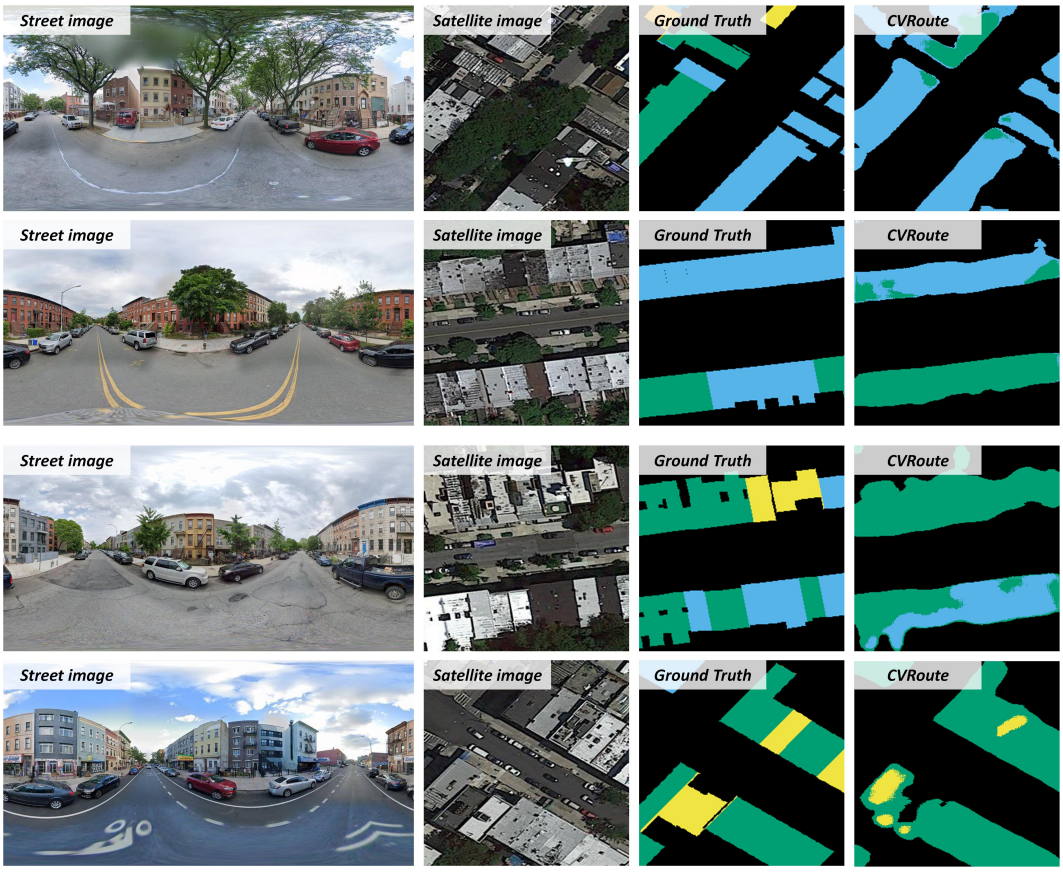

Figure 7: Representative failure cases of CVRoute. Small and densely arranged semantic regions remain challenging, leading to missed regions or merged predictions between neighboring spatial entities.