Doppelgangers++: Improved Visual Disambiguation with Geometric 3D Features

Yuanbo Xiangli[1] Ruojin Cai[1] Hanyu Chen[1] Jeffrey Byrne[2] Noah Snavely[1]
[1]Cornell University, [2]Visym Labs

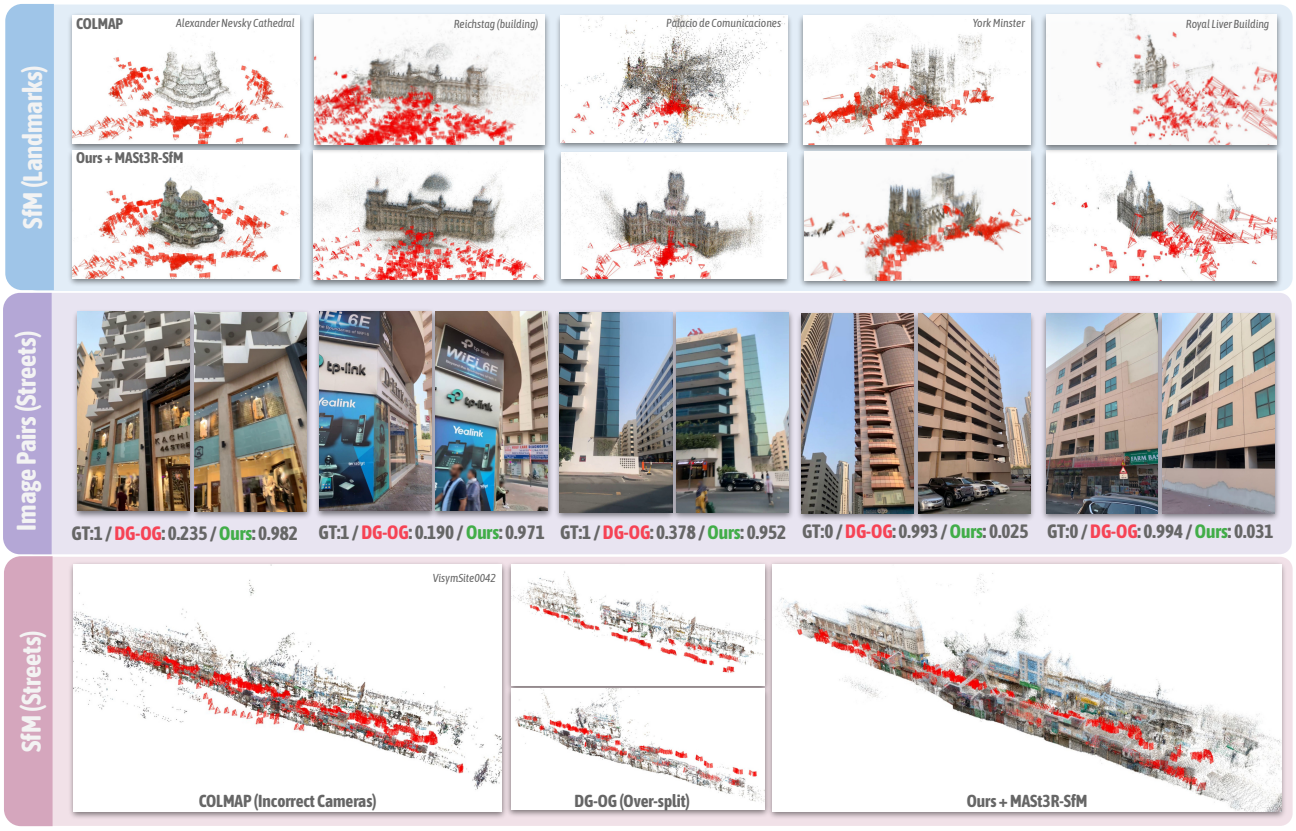

Figure 1. Visual aliasing, or doppelgangers, poses severe challenges to 3D reconstruction. We propose *Doppelganger++*, an enhanced pairwise image classifier that excels in visual disambiguation across diverse and challenging scenes. **(Top)** We seamlessly integrate Doppelganger++ into SfM, successfully disambiguating each scene. **(Middle)** Compared to prior work (which we refer to as DG-OG [3]), Doppelgangers++ is more robust for everyday scenes, showing improved accuracy and robustness. We show pairs that DG-OG classifies incorrectly and ours gets correct. **(Bottom)** Our new *VisymScenes* dataset, featuring complex daily scenes, is particularly challenging for COLMAP and DG-OG, but our method can achieve correct and complete reconstructions.

Abstract

Accurate 3D reconstruction is frequently hindered by visual aliasing, where visually similar but distinct surfaces (aka, doppelgangers*), are incorrectly matched. These spurious matches distort the structure-from-motion (SfM) process, leading to misplaced model elements and reduced accuracy. Prior efforts addressed this with CNN classifiers trained on curated datasets, but these approaches struggle to generalize across diverse real-world scenes and can require extensive parameter tuning. In this work, we present* Doppelgangers++*, a method to enhance doppelganger detection and improve 3D reconstruction accuracy. Our contributions include a diversified training dataset that incorporates geo-tagged images from everyday scenes to expand robustness beyond landmark-based datasets. We further propose a Transformer-based classifier that leverages 3D-aware features from the MASt3R model, achieving superior precision and recall across both in-domain and out-of-domain tests. Doppelgangers++ integrates seamlessly into standard SfM and MASt3R-SfM pipelines, offering efficiency and adaptability across varied scenes. To evaluate SfM accuracy, we introduce an automated, geotag-based method for validating reconstructed models, eliminating the need for manual inspection. Through extensive experiments, we demonstrate that Doppelgangers++ significantly enhances pairwise vi-*

sual disambiguation and improves 3D reconstruction quality in complex and diverse scenarios.

1. Introduction

Visual aliasing—confusing two surfaces that look the same but are nonetheless distinct—is an pernicious problem in 3D reconstruction and SLAM systems. Pairs of images that depict visually similar yet distinct 3D surfaces (called *doppelgangers* by Cai *et al*. [3]) can generate spurious correspondence at the feature matching stage of 3D reconstruction, leading to erroneous downstream reconstructions that feature distorted geometry or incorrectly fused elements. Therefore, to ensure the accuracy of 3D reconstruction, it is critical to distinguish truly matching images from illusory matches arising from doppelganger pairs.

In recent work, this visual disambiguation problem was formulated as a binary classification task on image pairs [3]. The authors collected a Internet dataset of visually similar doppelganger pairs (as well as truly matching images) from Wikimedia Commons, then trained a CNN to classify image pairs as correct or incorrect matches. This classifier can be used to take a feature match graph computed from a set of photos, remove incorrect edges between doppelganger image pairs, then reconstruct a correct model using a structure from motion (SfM) pipeline like COLMAP [12]. While that work shows promising improvements on reconstruction problems, we find that it can still be brittle: First, the 3D reconstruction task demands that the classifier have has precision—even a few bad edges remaining in the image match graph can lead to an incorrect 3D model. Second, the reconstruction task is quite sensitive to the threshold on the classifier score used to prune edges from the match graph. Finally, their method was trained solely on landmark Internet photos and does not reliably generalize to new scenarios, such as more structured captures of everyday scenes, like streets and office buildings.

In this paper, we aim to address these issues and improve doppelganger classification in several key ways:

1. We identify ways to expand and diversify doppelgangers training data. In particular, we leverage semi-structured image data with geographic annotations (rough GPS position and orientation), captured from everyday scenes with the Visym Collector platform [2].
2. We switch from a CNN-based classifier to leveraging features from MASt3R [5], a recent transformer-based geometric model that computes point clouds from two input views. Specifically, we feed an image pair into a pre-trained MASt3R model, collect the intermediate features decoded by MASt3R, and train a classification head to map these features to a doppelganger score.

We call our method *Doppelgangers++*. Our enhanced model achieves higher precision and recall in pairwise classification, and generalizes better across a broader range of scenes and capture scenarios. Doppelganger++ integrates seamlessly into SfM pipelines for 3D reconstruction disambiguation. We further propose to leverage geo-tagged map images to quantitatively evaluate the correctness and completeness of reconstructed models, in comparison to manual inspection as required in previous approaches. Through extensive experiments we show that our model leads to more accurate and complete reconstruction results, and is less sensitive to the threshold used for pruning doppelganger matches.

2. Related Work

Local feature matching and 3D learning. Local feature matching methods have proven effective at establishing correspondences between pairs of images in SfM pipelines. Classic methods like COLMAP [12] rely on the tried-and-true SIFT features [10] to find correspondence. Modern learning-based, data-driven approaches have improved the quality of local feature matching [4, 11, 13, 21]. More recently, the DUSt3R [16] framework has proven to excel in a variety of 3D reconstruction tasks by estimating dense 3D point clouds from pairs of images. MASt3R [9] was proposed as an extension to DUSt3R that specifically targets the task of predicting dense feature matches between image pairs. All of these local feature matching methods are traditionally optimized to maximize the number of correspondences they find between similar image regions. While their ability to identify feature matches has greatly improved, they generally lack the ability to incorporate *negative evidence* into their predictions, and so they often find matches between regions that do not actually correspond to the same 3D surface, particularly within doppelganger image pairs. However, we show that features internal to MASt3R can be repurposed for doppelganger detection.

Disambiguation in SfM and image matching. Disambiguating similar structures and repeated patterns in SfM is a long-standing challenge. Prior work has mainly relied on heuristics-based analysis to detect conflicting relations and ambiguities in the structure of the underlying scene graph [19, 20, 23] or among image-level correspondence (or the lack thereof) [6, 7, 22]. While such methods have shown some success in detecting and resolving ambiguities in SfM pipelines, one fundamental limitation that they do not consider the rich information contained in the images themselves. In contrast, Doppelgangers [3] avoids handcrafted heuristics and takes a data-driven approach to the basic problem of identifying doppelganger image pairs. They train a CNN that classifies a pair of images as either positive or negative, with LoFTR [13]-extracted keypoints and match masks passed as auxiliary input. During inference, the binary classification results are used to pre-process the scene graph obtained from COLMAP to filter out spurious matches prior to running SfM. However, the method's re-

liance on COLMAP feature extraction and matching and the need for auxiliary mask information from LoFTR introduces significant overhead and complexity to the overall pipeline, making it difficult to scale to larger scenes. In addition, we find that this prior method is brittle: it can fail to generalize to domains beyond landmarks (e.g., to office buildings), and it can require parameter tuning to get good results.

Differentiable SfM. To improve upon traditional SfM, several differentiable SfM pipelines have been proposed to optimize the entire 3D reconstruction process. These pipelines often include one or more components that are trained from large datasets. MASt3R-SfM [5] proposes an SfM pipeline that uses MASt3R features for scene graph construction and coarse-to-fine alignment. VGGSfM [15] decomposes the problem into four stages: point tracking, initial camera estimation, triangulation, and bundle adjustment, each of which is differentiable and can be optimized in an end-to-end manner. ACE0 [1] uses a scene coordinate regression network to progressively register new images to a global scene, while the network itself is iteratively trained and refined from the registered images. These methods have shown promising results in optimizing the SfM pipeline, but are not specifically designed to address the doppelganger problem, and are still prone to producing incorrect reconstructions when faced with scenes with repeated patterns or similar structures.

3. Method

We refer to a "doppelganger" as a case where distinct objects or surfaces look almost identical, leading algorithms to confuse one for the other [3]. In this work, we aim to improve the doppelganger classifier via expanded and diversified training data (Sec. 3.1), and leveraging geometric 3D features learned from pairwise reconstruction [9] (Sec. 3.2). We further propose an approach to quantitatively evaluate the correctness of SfM results in term of doppelgangers (Sec. 3.3), instead of manual inspection adopted in previous work.

3.1. The VisymScenes dataset

Cai *et al*. [3] introduced the Doppelganger dataset, built on global landmarks photos sourced from Wikimedia Commons, with viewing direction (*e.g*., North, South) to identify doppelganger image pairs. While the classifier trained on this dataset proved effective, we find that it struggles to generalize well to scenes beyond the dataset's domain.

We enhance and expand their training set to improve the robustness of doppelganger classification, by incorporating casually captured images from a wider range of scenes. To ensure sufficient diversity in the dataset, we introduce the *VisymScenes* dataset. VisymScenes consists of 258K ground images with GPS/IMU metadata, recorded at 149 sites in 42 cities and 15 countries, collected with the Visym Collector Platform [2] (details in supplementary). Each image is accompanied by metadata, such as GPS coordinates, device compass direction, and intrinsic camera calibration. While this metadata can be noisy, it still provides valuable information for identifying potential doppelganger pairs. For example, if two images exhibit numerous geometrically consistent local feature matches yet were taken from distant locations, this serves as strong evidence that the pair is likely to be a doppelganger. Examples are shown in Fig. 2.

We develop a series of filtering rules applied to all matched image pairs identified by the COLMAP feature matching module [12]. Note that these rules are designed according to the capture style of the Visym Collector, where cameras consistently focus on the scene of interest and maintain a reasonable distance from it. For a given pair of images, we use the (metric) distance between the camera centers r, angle between their viewing directions θ, and camera intrinsics $\mathcal{K}$, all derived from camera metadata, to identify confident negative (doppelganger) and positive (correct match) pairs. To identify confident negative pairs, we first identify spatially distant matching pairs, and classify them as doppelgangers. Then, we classify the remaining (nearby) pairs into three cases: the intersection point of the viewing directions is 1) in front of both cameras, 2) behind both cameras; or 3) in front of one camera and behind the other. For case 1), we label pairs with very large view angles (*e.g*., $> 160°$) as negative, since they likely capture different 3D surfaces or mostly non-overlapping content. For case 2), if their view angle exceeds the diagonal field of view, then their view frustums are unlikely to overlap, so we label them as doppelgangers. For case 3), we check for frustum overlap using camera intrinsics; if no intersection is detected, the pair is considered a doppelganger. A similar series of rules are designed for mining positive pairs (*i.e*., true matches). Details of computing view intersection and filtering algorithms for both negative and positive pairs can be found in the supplementary. With these rules, we mined in total 53K positive and negative pairs across 33 sites for our doppelganger task.

3.2. Improved Doppelganger Classifier

Recent advances in data-driven models [9, 16] have demonstrated impressive results in geometric vision tasks. In our work, we leverage the multi-level, 3D-aware features extracted from the MASt3R model [9] to train a doppelganger classifier on labeled image pairs, as demonstrated in Fig. 3.

Equipped with large-scale training data and a ViT backbone, the MASt3R model captures an image representation that encodes rich 3D geometric information between paired images. However, as MASt3R was originally trained to detect correspondences and similarities, its point maps often conflate doppelganger pairs, yielding incorrect point clouds and poses for these challenging pairs (examples in Fig. 7). Despite this limitation, we find that the internal features learned by MASt3R contain sufficient information

Figure 2. **VisymScenes examples**. This new dataset includes residential areas, landmarks, historical sites, business districts, and more. Here, we present four example sites. The top row shows subsets of images captured within each site. The bottom row displays pairs of visually similar but geographically distinct images from each site along with their recorded geolocations on a map. These examples demonstrate that doppelganger issues are prevalent in everyday scenes, presenting significant challenges for reliable 3D reconstruction and image matching.

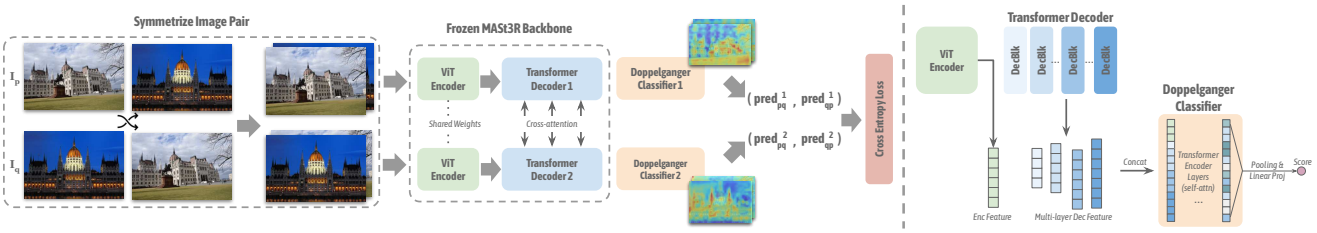

Figure 3. **Model design. (Left)** Given an image pair, we first create a symmetrized version of the pair and feed it into the frozen MASt3R model. Multi-layer features are extracted from each decoder branch, concatenated, and fed into two learnable doppelganger classification heads. Each head generates predictions $(\text{pred}^v_{pq}, \text{pred}^v_{qp}), v \in \{1, 2\}$ (where pq and qp denote the symmetrized image pair), supervised by cross-entropy loss. **(Right)** We use multi-layer decoder features and a Transformer-based classifier head for doppelganger prediction.

for doppelganger classification without the need to fine-tune MASt3R's weights. Moreover, given that the amount of available training data for the doppelganger task is dwarfed by the amount of data used to train MASt3R, we opt *not* to fine-tune the entire model, but instead repurpose these internal features by training an additional output head to predict a doppelganger classification score. In Sec. 4.5, we empirically demonstrate that this choice achieves comparable or better visual disambiguation performance, especially on out-of-domain scenes.

Multi-level decoder features. For an image pair (I_p, I_q), we extract multi-level decoder features from MASt3R's decoder branches. MASt3R takes two images as input, and uses two intertwined decoders DecBlk^1 and DecBlk^2 to decode encoded features $\mathcal{H}^1 = \text{Enc}^1(I_p)$ and $\mathcal{H}^2 = \text{Enc}^2(I_q)$. Each decoder has B blocks, and attends to tokens from the other branch:

$$f^1_i = \text{DecBlk}^1_i\left(f^1_{i-1}, f^2_{i-1}\right), f^2_i = \text{DecBlk}^2_i\left(f^2_{i-1}, f^1_{i-1}\right), \quad (1)$$

where f^v_i denotes the output tokens from the i-th block of the v-th branch. These two branches exchange information to capture the spatial relationships between viewpoints and the global 3D structure of the scene. For each branch v, we concatenate the encoder feature $\mathcal{H}^v$ and the tokens from the decoder blocks into $\mathcal{F}^v = \text{concat}(\mathcal{H}^v, \{f^v_i\}^{B-1}_{i=0})$, which captures rich, multi-level spatial correspondence information between image pairs.

Transformer-based classification heads. MASt3R treats one image as the reference frame and projects the other image into that reference frame. A consequence of this design is that the reconstructed 3D scene for input pair (I_p, I_q) is distinct (or at least, must be in a different coordinate frame) from that of (I_q, I_p). Inspired by the asymmetric design, we propose to 1) use separate Transformer heads: We introduce two independent, Transformer-based classification heads Head^1_{dopp} and Head^2_{dopp} to process $\mathcal{F}^1$ and $\mathcal{F}^2$ respectively; and 2) symmetrize image pairs such that the model evaluates both (I_p, I_q) and (I_q, I_p) to decide whether the given pair of images are true match or not. Thus, we end up with four scores for each image pair:

$$\text{pred}^v_u = \text{Head}^v_{\text{dopp}}(\mathcal{F}^v_u), u \in \{pq, qp\}, v \in \{1, 2\}. \quad (2)$$

Essentially, the two heads serve as expert evaluators, each examining how likely the image pair is to be a true match (or doppelganger) from different aspects, and switching the input order allows the model to analyze the spatial relationships from both directions. We empirically show the effectiveness of this design in Sec. 4.5. Both Head^1_{dopp} and Head^2_{dopp} are supervised by the cross-entropy loss, encouraging $\mathcal{S} = \{\text{pred}^v_u\}$ to all give high probabilities for positive matches and low probabilities for negative ones.

Test-time voting. At test time, we combine the four scores $\mathcal{S}$ via a voting mechanism to make a final decision. Specifically, if the majority of the heads predict that the pair is a positive

match, we take $\max(\mathcal{S})$ as the final score for the image pair. Conversely, if the majority of the heads vote for a negative match, $\min(\mathcal{S})$ is used as the final score. Otherwise, we average the scores from the four heads. This approach ensures that the final decision reflects the strongest evidence supporting the consensus among the heads and thereby improves the reliability of the classifier. Eq. 3 elaborates on this voting mechanism:

$$S_{\text{final}} = \begin{cases} \max(\mathcal{S}) & \text{if } \sum_{s\in\mathcal{S}} \mathbf{1}(s > 0.5) > \sum_{s\in\mathcal{S}} \mathbf{1}(s < 0.5), \\ \min(\mathcal{S}) & \text{if } \sum_{s\in\mathcal{S}} \mathbf{1}(s > 0.5) < \sum_{s\in\mathcal{S}} \mathbf{1}(s < 0.5), \\ \text{mean}(\mathcal{S}) & \text{if } \sum_{s\in\mathcal{S}} \mathbf{1}(s > 0.5) = \sum_{s\in\mathcal{S}} \mathbf{1}(s < 0.5). \end{cases} \tag{3}$$

3.3. Evaluating Doppelganger correction in SfM

There is currently no reliable benchmarking method for evaluating the accuracy and correctness of SfM reconstructions specifically in terms of how well they address the doppelganger issue. Prior work [3] relied on manual inspection of each out model to assess the effectiveness of their approach—a time-consuming, unquantifiable, and potentially error-prone process. In our work, we propose a benchmarking method for qualitatively evaluating reconstructed models with respect to doppelgangers. We leverage mapping sites like Mapillary, which provide images with location metadata that can serve as probes for validating a 3D model. Note that our method targets common scenarios where geotags are unavailable, or so noisy as to not be useful. Therefore, we do not consider the use of geotags for the reconstruction task itself—we explore pure visual disambiguation—but instead gather them from specialized sources for ground truth evaluation.

Specifically, for a given scene with known geolocation, we acquire from Mapillary a set of nearby geo-tagged images and register them to the reconstructed model with COLMAP [12]. We then apply RANSAC to robustly estimate a similarity transformation between the registered camera positions and their corresponding image geolocation metadata (converted to ECEF coordinates). We use the resulting RANSAC *inlier ratio* as an indicator of the model's correctness. An example is shown in Fig. 4. We can observe that before correction, the registered cameras and their geolocations show significant misalignment, with the cameras collapsing to one side due to similar-looking surfaces. After correction, the registered cameras align closely to their true geolocations, indicating that doppelganger pairs are correctly separated in the model. If the model is split into multiple components, we calculate the weighted average of inliers ratios among the components:

$$\text{IR} = \sum_{i=1}^{N} \left(\frac{I_i}{T_i} \cdot \frac{T_i}{\sum_{j=1}^{N} T_j} \right) = \frac{\sum_{i=1}^{N} I_i}{\sum_{i=1}^{N} T_i}, \tag{4}$$

where I_i is the number of RANSAC inliers and T_i is the number of registered Mapillary images in the i-th component. While alignment error could also serve as a metric, we opt to use the inlier ratio to reduce sensitivity to geolocation inaccuracies, as image geolocations may not always be precise. This strategy can also be used to detect broken reconstructions from datasets like MegaScenes [14], which contains over 100K SfM results from Internet photos around the world.

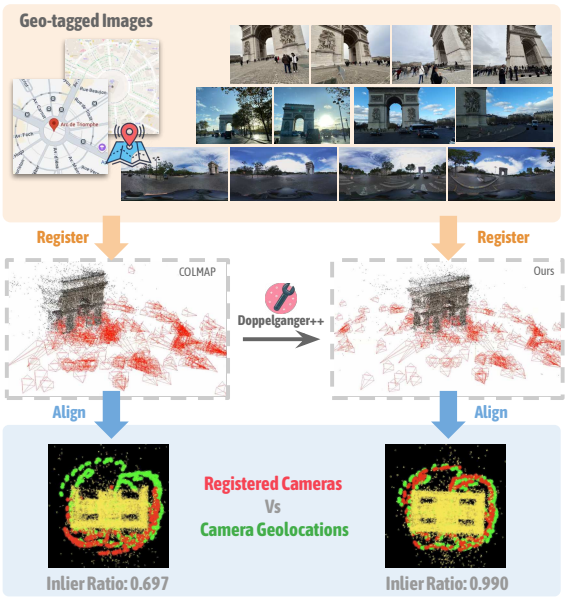

Figure 4. **Evaluation of doppelganger correction in SfM. (Top)** We first collect sequences of geo-tagged Mapillary images around the target location and register them to the SfM model. Then, we use RANSAC to align the registered cameras and their geolocations. The inlier ratio is computed as an indicator of model accuracy. **(Bottom)** In the model corrupted by doppelganger pairs, the registered cameras all collapse to one side. We see that the camera poses estimated with COLMAP (right, in red) do not align well with the geotags (green), leading to a low inlier ratio, but our method leads to a much closer alignment.

4. Experiments

4.1. Experimental setup

Dataset. Our training set is comprised of the Doppelganger dataset [3] (dubbed DG) and the new VisymScenes dataset. Specifically, the DG training set includes 73K image pairs, nearly evenly split between positive and negative pairs. From VisymScenes, we use 47K image pairs across 26 sites as training data, evenly split into positive and negative pairs.

We evaluate on two tasks, 1) pairwise image visual disambiguation and 2) scene reconstruction by integrating our classifier into COLMAP [12]. For pairwise image visual disambiguation, we utilize the DG test set, and mined an additional 3,180 pairs from 7 Visym sites (distinct from the 26 training sites), equally divided into positive and negative pairs. Additionally, we created a test set from the Mapillary

Street-Level Sequences dataset [17]. This Mapillary dataset spans diverse urban and suburban environments and captures a wide range of seasons, weather conditions, cameras, lighting, and structural settings, with each image geo-tagged by GPS and compass angle. Using a similar filtering approach to the VisymScenes dataset, we mined 1,500 positive and 1,500 negative pairs as an out-of-domain test set.

For the scene reconstruction task, we evaluate on 16 scenes sampled from Heinly *et al*. [6], Wilson *et al*. [18], MegaScenes [14] and 5 VisymScenes test scenes. These scenes are challenging for conventional SfM pipelines as well as prior disambiguation methods due to subtle differences between distinct surfaces and repetitive patterns.

Metrics. For pairwise visual disambiguation evaluation, we report Average Precision (AP) and ROC AUC scores, following the protocol in [3]. Additionally, we report precision at a fixed recall and recall at a fixed precision. These statistics help characterize a model's trade-off between achieving high precision and potentially missing true positives.

To comprehensively evaluate SfM reconstructions, we report the number of images in the SfM results, and the geo-alignment inlier ratio described in Sec. 3.3. Respectively, more images in the SfM model imply less false pair pruning, *i.e*., the model has better class separation; and a higher geo-alignment inlier ratio suggests that the reconstructed model is more likely to be accurate and complete. Note that the inlier ratio is also influenced by the accuracy of the geo-tagged images collected from Mapillary. Thus, if one test scene has more accurate geo-tagged images than another, it is likely to yield a higher inlier ratio. Therefore, rather than comparing inlier ratios across different scenes, this metric is more useful for evaluating different reconstruction methods on the same scene, as the same set of geo-tagged images is used.

4.2. Implementation

Model details. Our doppelganger classifier head $\text{Head}_{\text{dopp}}$ is implemented with a Transformer encoder comprised of 3 layers, each with input and output dimension of 768. Each layer has 8 attention heads and a two-layer feed-forward network with a hidden dimension of 2048. Pre-layer normalization and residual connections are applied within each layer. The transformed tokens are aggregated via max pooling and linearly projected into pred^v. We freeze the weights of the public MASt3R [9] model and only train the two doppelganger classification heads, supervised by a cross-entropy loss. We train for 5 epochs with a batch size of 8 using Adam [8] with a learning rate of 1×10^{-4}. More details can be found in the supplementary.

Integration with SfM. Following [3], we integrate Doppelgangers++ into SfM to enhance its disambiguation ability. SfM takes a collection of images $\mathcal{I} = \{I_i\}_{i=0}^{n}$ and generates geometrically verified image pairs $\mathcal{P} = \{(I_p, I_q)\}$, forming a scene graph $\mathcal{G} = (\mathcal{I}, \mathcal{P})$ with images as nodes and pairs as edges. Using $\mathcal{G}$, SfM computes camera poses and reconstructs a 3D point cloud. Our doppelganger classifier acts as a filter on edges in $\mathcal{G}$, removing pairs below a probability threshold τ to eliminate spurious matches due to repeated or symmetric structures. We integrate Doppelgangers++ into both COLMAP [12] and MASt3R-SfM [5] to showcase the effectiveness of the method. Notably, with COLMAP, SIFT features are computed for scene graph construction and MASt3R features for pruning; whereas with MASt3R-SfM, MASt3R features serve for both scene graph construction and pruning, forming a more efficient pipeline.

4.3. Pairwise Visual Disambiguation

We compare our approach with the original Doppelganger work [3] (dubbed *DG-OG*) under two experimental settings: 1) we use the same training data as in [3], *i.e*. the DG training set, to train our model. Both models are evaluated on three benchmark test sets: DG, VisymScenes, and Mapillary. Under this configuration, the DG test set is in-domain, while VisymScenes and Mapillary are out-of-domain scenarios; 2) we expand the training data by including the new VisymScenes training set and retrain both DG-OG and our model. Evaluation is conducted on the same three test sets, with only the Mapillary test set being out-of-domain this time.

Quantitative results are in Tab. 1. Under the first setting, where both methods are trained on DG, our model shows clear improvements across all three test sets. Specifically, on the in-domain DG test set, our model achieves higher AP and ROC AUC, indicating that it maintains high precision across all recall levels, while also being less sensitive to the decision threshold τ with better class separation. On the out-of-domain test sets (VisymScenes and Mapillary), we observe 25% to 65% improvements in both AP and ROC AUC, highlighting the generalizability of our method.

Although our model demonstrates improved performance even when trained solely on the DG dataset, its precision outside of the training domain is suboptimal for practical use in SfM, where high precision is essential. In the second setting, we expand the training set by including VisymScenes data. Both methods maintain similar performance on the DG test set, with our model experiencing a slight drop in recall when precision is set to 0.99. On the VisymScenes test set, both models show improvements across all metrics, with ours reaching 99% in both AP and ROC AUC. Notably, adding VisymScenes to the training set does not enhance DG-OG's performance on the out-of-domain Mapillary test set, while ours continues to benefit from increased training diversity. These results indicate that the prior approach struggles to generalize effectively, whereas ours shows greater generalization with more varied training data.

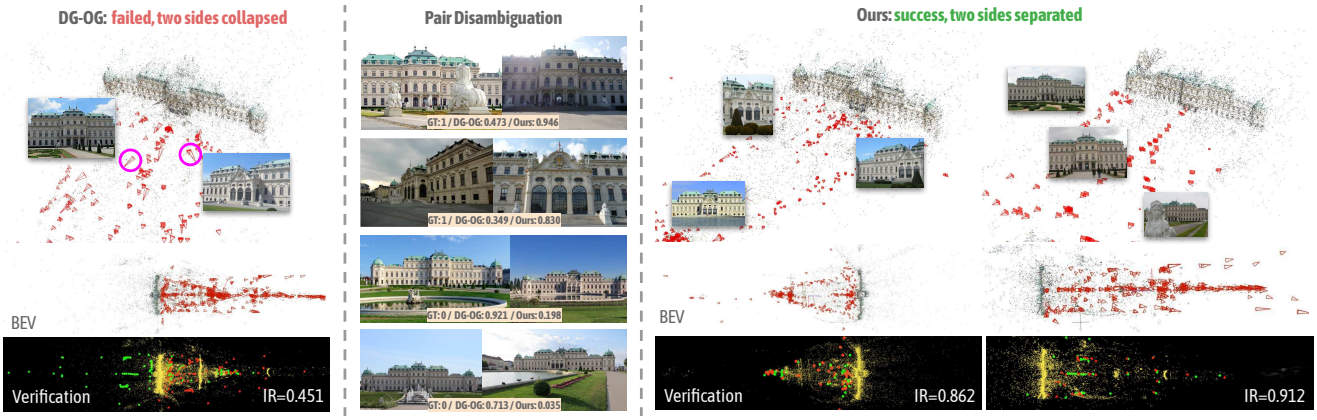

Figure 5. **SfM Disambiguation on MegaScenes [14]. (White background)** SfM results from DG-OG [3] and ours. **(Black background)** Verification using geo-tagged images, red points represent registered cameras and green points represent geolocations, inlier ratio (IR) is labeled on the bottom right. DG-OG fails to disambiguate this scene, predicting incorrect scores for image pairs. Our method correctly splits the model into two clean components.

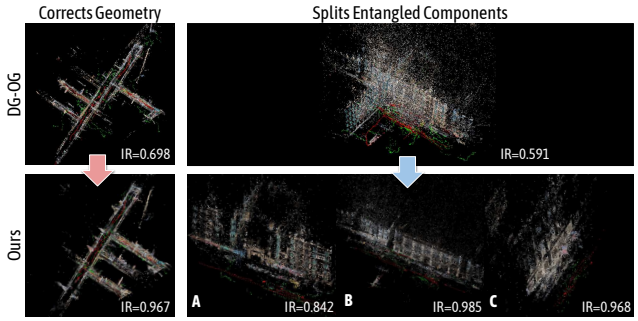

Figure 6. **SfM disambiguation on VisymScenes.** We show that our classifier is more robust than DG-OG [3] on test scenes from new domains, like everyday street scenes. DG-OG has difficulty disambiguating such scenes, leading to incorrect geometry.

Test Data	Method	Metrics (trained on DG / trained on DG + VisymScenes)			
		AP↑	ROC AUC↑	Prec@Recall=0.85↑	Recall@Prec=0.99↑
DG	DG-OG	0.954 / 0.956	0.944 / 0.947	0.901 / 0.910	0.611 / 0.614
	Ours	0.980 / 0.981	0.981 / 0.981	0.972 / 0.982	0.690 / 0.642
VisymScenes	DG-OG	0.816 / 0.938	0.726 / 0.921	0.498 / 0.831	0.340 / 0.623
	Ours	0.936 / 0.991	0.909 / 0.990	0.892 / 0.999	0.542 / 0.901
Mapillary	DG-OG	0.566 / 0.692	0.581 / 0.701	0.523 / 0.572	0.003 / 0.000
	Ours	0.950 / 0.968	0.944 / 0.958	0.927 / 0.942	0.310 / 0.736

Table 1. **Evaluation of pairwise disambiguation.** We evaluate DG-OG [3] and our method trained on DG [3] only and DG +VisymScenes (two numbers per cell) on three test sets. Both DG-OG and ours benefit from dataset expansion, whereas ours gained more generalizability on out-of-domain test set (Mapillary) after training on both. Our classifier constantly demonstrates better precision, recall across all settings.

4.4. Structure from Motion disambiguation

We integrate our classifier trained on DG and VisymScenes into COLMAP's SfM pipeline [12], and evaluate its performance on reconstructing scenes with duplicated and symmetric structures. We compare the results with vanilla COLMAP and Cai *et al*. [3] (DG-OG). Quantitative results are presented in Tab. 2. Our approach registers more images across all scenes than DG-OG and operates without threshold tuning. In contrast, DG-OG relies on scene-specific thresholds, such as $\tau = 0.97$ for the Church on Spilled Blood to correctly separate both sides of the church, and $\tau = 0.6$ for Ponte di Rialto [14] to avoid over-segmentation and maintain completeness. Notably, DG-OG fails to disambiguate Belvedere (Vienna), while our method succeeds with a consistent threshold ($\tau = 0.8$ across all scenes). Our approach also consistently achieves a higher inlier ratio than the baselines, indicating greater accuracy and completeness in the reconstructed models. As an example, we qualitatively show the results of reconstructing Belvedere (Vienna) in Fig. 5, along with verification results using geo-tagged images.

We also evaluate on scenes from the VisymScenes test set. Since VisymScenes images come with geolocation metadata, we do not need additional Mapillary images; instead, we directly compute the inlier ratio between SfM camera positions and the geolocation metadata after RANSAC. Results show that our model effectively prunes spurious pairs and improves the reconstructed model quality, whereas DG-OG sometimes fails to distinguish doppelganger pairs. Fig. 6 shows examples where DG-OG encounters difficulties.

In Fig. 7, we show that while MASt3R's features are powerful, MASt3R-SfM is not free from doppelganger issues. Although our classifier was trained on image pairs mined through COLMAP's feature matching module (*i.e*. w/ SIFT features), it effectively prunes incorrect matches generated by MASt3R, restoring accurate reconstruction results.

4.5. Ablation

In this section, we study the effectiveness of our designs from the following aspects: 1) fine-tuning the entire model (vs. only new heads), 2) one classification head (vs. two separate heads), 3) the architecture of our classification head, and 4) final-layer decoder features (vs. multi-layer decoder

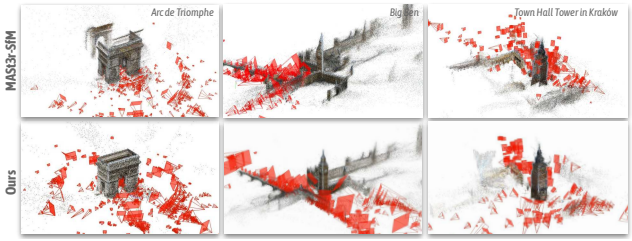

Figure 7. **MASt3R-SfM w/ Doppelgangers++.** MASt3R-SfM also suffers from doppelganger issues. Our classifier effectively prunes false positive pairs and correctly reconstructs challenging scenes.

Test Scenes	# SfM-registered Images			Inlier Ratio (Sec. 3.3)		
	COLMAP	DG-OG	Ours	COLMAP	DG-OG	Ours
Alexander Nevsky Cathedral [6]	447	444	447	0.565	1.0	1.0
Arc de Triomphe [6]	424	384	423	0.697	0.966	0.990
Berliner Dom [6]	1603	1588	1606	0.709	0.987	0.992
Big Ben [6]	398	379	394	0.564	0.827	0.831
Church on Spilled Blood [6]	273	84+94 (τ=0.97)	157+106	0.542	0.881	0.962
Radcliffe Camera [6]	281	91+84	94+186	0.495	0.955	0.970
Seville [18]	1498	585+272+515	615+303+552	0.450	0.772	0.854
Brandenburg Gate [14]	2137	1361+570	1398+603	0.440	0.900	0.909
Palacio de Comunicaciones [14]	727	307+80 (τ=0.6)	308+84	0.229	0.823	0.934
Ponte di Rialto [14]	652	538+101 (τ=0.6)	540+107	0.627	0.834	0.844
York Minster [14]	636	200+362	206+284	0.727	0.858	0.901
Town Hall Tower, Kraków [14]	298	255	280	0.609	0.731	0.838
Belvedere (Vienna) [14]	1038	851 (fail)	457+500	0.521	0.451	0.874
Reichstag (building) [14]	1504	997+310	1024+356	0.469	0.804	0.862
St. Vitus Cathedral [14]	752	673	692	0.853	0.909	0.933
Royal Liver Building [14]	212	171	180	0.7	1.0	1.0
VisymSite0010	1544	520+290	1446	0.770	0.820	0.913
VisymSite0023	849	471+81	566+82	0.867	0.848	0.942
VisymSite0028	450	238+179	267+120	0.818	0.711	0.909
VisymSite0042	540	206+207	467	0.863	0.924	0.963
VisymSite0109	1245	237+458+78	239+612+127	0.857	0.862	0.927

Table 2. **Evaluation of SfM results**. We compare reconstructions from COLMAP [12], DG-OG [3], and our method. τ=0.8 is used unless otherwise stated. The '+' symbol indicates split reconstruction components; *e.g.*, DG-OG and our method split the Radcliffe Camera reconstruction into two components. Because VisymScenes scenes are large, we report statistics on the largest reconstruction component produced by COLMAP, and identify the corresponding (split) components in the results of DG-OG and ours.

features). We evaluate models trained on DG and VisymScenes datasets and show PR curves in Fig. 8 on the three test sets with respect to the pairwise classification task. A full ablation table can be found in the supplementary.

Fine-tuning MASt3R. As discussed in Sec.3.2, we choose to train a lighter output head on top of MASt3R features rather than fine-tuning the entire model weights. Curves in Fig. 8 show that training only the head achieves comparable performance to full model fine-tuning. The drop in recall with full model fine-tuning suggests potential overfitting to training data, making the model more conservative and less generalizable to unseen inputs. This tendency also holds on models trained solely on DG data: on the VisymScenes test set, our approach achieved 0.03 higher AP than fine-tuning. On Mapillary test set, ours achieved comparable AP but with 0.2 higher recall at precision 0.99.

Single classification head. Our design uses two classification heads to process features from the two branches separately. An alternative approach would be to combine the

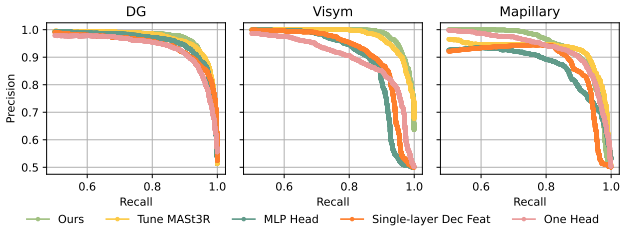

Figure 8. **Precision-Recall curves of ablation studies.** Metrics are evaluated on models trained with DG and VisymScenes. Full tables can be found in the supplementary.

features from both decoder branches and use a single head to predict the doppelganger score. As shown in Fig. 8, this alternative degrades performance consistently. This may be due to the asymmetric nature of the MASt3R design, where each branch captures different kinds of information that are better to be processed separately. On the other hand, our voting mechanism between the two heads helps enhance class distinction, outputting lower scores for negative samples and higher scores for positive samples.

Head architecture. While the decoder features of MASt3R retain rich geometric 3D information, designing a classifier with sufficient capacity to effectively leverage this information for the doppelganger prediction task is non-trivial. We compare the performance of our Transformer-based classifier with that of an MLP. Fig. 8 shows that our Transformer model outperforms the MLP across all metrics and test sets.

Single-layer decoder feature. Our classifier head takes multi-layer decoder features as input. Here we substitute with only the final-layer decoder feature. Results in Fig. 8 show that the multi-layer decoder features are consistently superior to the single-layer ones, because the classifier is able to analyze doppelganger factors from different aspects and levels. See supplementary for more detailed analysis.

5. Conclusion

We propose Doppelgangers++ as an effective approach to handling visual aliasing in 3D reconstruction. We introduce a new VisymScenes dataset, featuring images from diverse daily scenes, and develop rules to mine doppelganger data from the dataset to enrich our training data. We train a Transformer-based classifier that leverages geometric 3D features to classify image pairs with high precision and recall across various scenes. Doppelgangers++ integrates seamlessly into existing SfM pipelines, enhancing reconstruction accuracy without extensive parameter tuning. Further, we propose a validation method using geo-tagged map images, offering a more comprehensive and automatic way to assess SfM accuracy and model completeness. Extensive experiments show that Doppelgangers++ significantly improves visual disambiguation and 3D reconstruction quality in complex scenes.

6. Acknowledgment

We thank Joseph Tung and Brandon Li for their valuable discussion and help with the webviewer. This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/Interior Business Center (DOI/IBC) contract number 140D0423C0035. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/IBC, or the U.S. Government. This research was also supported in part by the National Science Foundation under award IIS-2212084.

References

[1] Eric Brachmann, Jamie Wynn, Shuai Chen, Tommaso Cavallari, Áron Monszpart, Daniyar Turmukhambetov, and Victor Adrian Prisacariu. Scene coordinate reconstruction: Posing of image collections via incremental learning of a relocalizer. In *Eur. Conf. Comput. Vis.*, pages 421–440. Springer, 2025. 3

[2] J. Byrne, G. Castanon, Z. Li, and G. Ettinger. Fine-grained activities of people worldwide. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023. 2, 3, 1

[3] Ruojin Cai, Joseph Tung, Qianqian Wang, Hadar Averbuch-Elor, Bharath Hariharan, and Noah Snavely. Doppelgangers: Learning to disambiguate images of similar structures. In *Int. Conf. Comput. Vis.*, pages 34–44, 2023. 1, 2, 3, 5, 6, 7, 8, 4

[4] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proc. CVPR Workshops*, pages 224–236, 2018. 2

[5] Bardienus Pieter Duisterhof, Lojze Žust, Philippe Weinzaepfel, Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Mast3r-sfm: a fully-integrated solution for unconstrained structure-from-motion. *ArXiv*, abs/2409.19152, 2024. 2, 3, 6, 5

[6] Jared Heinly, Enrique Dunn, and Jan-Michael Frahm. Correcting for duplicate scene structure in sparse 3d reconstruction. In *Eur. Conf. Comput. Vis.*, pages 780–795. Springer, 2014. 2, 6, 8

[7] Nianjuan Jiang, Ping Tan, and Loong-Fah Cheong. Seeing double without confusion: Structure-from-motion in highly ambiguous scenes. In *Conf. Comput. Vis. Pattern Recog.*, pages 1458–1465, 2012. 2

[8] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 6

[9] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. *ArXiv*, abs/2406.09756, 2024. 2, 3, 6, 5

[10] David G Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.*, 60:91–110, 2004. 2

[11] Ignacio Rocco, Mircea Cimpoi, Relja Arandjelović, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Neighbourhood consensus networks. In *Adv. Neural Inform. Process. Syst.*, 2018. 2

[12] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conf. Comput. Vis. Pattern Recog.*, pages 4104–4113, 2016. 2, 3, 5, 6, 7, 8, 1

[13] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Conf. Comput. Vis. Pattern Recog.*, pages 8922–8931, 2021. 2

[14] Joseph Tung, Gene Chou, Ruojin Cai, Guandao Yang, Kai Zhang, Gordon Wetzstein, Bharath Hariharan, and Noah Snavely. Megascenes: Scene-level view synthesis at scale. *ArXiv*, abs/2406.11819, 2024. 5, 6, 7, 8, 1, 3

[15] Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. Vggsfm: Visual geometry grounded deep structure from motion. In *Conf. Comput. Vis. Pattern Recog.*, pages 21686–21697, 2024. 3

[16] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jérôme Revaud. DUSt3R: Geometric 3d vision made easy. In *Conf. Comput. Vis. Pattern Recog.*, pages 20697–20709, 2023. 2, 3

[17] Frederik Warburg, Søren Hauberg, Manuel López-Antequera, Pau Gargallo, Yubin Kuang, and Javier Civera. Mapillary street-level sequences: A dataset for lifelong place recognition. In *Conf. Comput. Vis. Pattern Recog.*, pages 2623–2632, 2020. 6

[18] Kyle Wilson and Noah Snavely. Network principles for sfm: Disambiguating repeated structures with local context. In *Int. Conf. Comput. Vis.*, pages 513–520, 2013. 6, 8

[19] Kyle Wilson and Noah Snavely. Network principles for sfm: Disambiguating repeated structures with local context. In *Int. Conf. Comput. Vis.*, pages 513–520, 2013. 2

[20] Qingan Yan, Long Yang, Ling Zhang, and Chunxia Xiao. Distinguishing the indistinguishable: Exploring structural ambiguities via geodesic context. In *Conf. Comput. Vis. Pattern Recog.*, pages 3836–3844, 2017. 2

[21] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. Lift: Learned invariant feature transform. In *Eur. Conf. Comput. Vis.*, pages 467–483, 2016. 2

[22] Christopher Zach, Arnold Irschara, and Horst Bischof. What can missing correspondences tell us about 3d structure and motion? In *Conf. Comput. Vis. Pattern Recog.*, pages 1–8, 2008. 2

[23] Christopher Zach, Manfred Klopschitz, and Marc Pollefeys. Disambiguating visual relations using loop constraints. In *Conf. Comput. Vis. Pattern Recog.*, pages 1426–1433, 2010. 2

Doppelgangers++: Improved Visual Disambiguation with Geometric 3D Features

Supplementary Material

7. VisymScenes Dataset Details

The VisymScenes dataset consists of 258K ground images with GPS/IMU metadata, recorded at 149 sites in 42 cities and 15 countries. Each site is specified as a 200m $\times$ 200m area containing a man-made structure in an urban, suburban or rural environment. VisymScenes was collected by freelancers from July 2023 to March 2024 using the Visym Collector platform [2], a framework for distributed collection of image and video datasets. Freelancers from around the world were tasked with selecting an interesting site near them, then collecting imagery while walking around their site using the Collector mobile app. Imagery was collected at 1Hz intervals from a first-person perspective, outdoors, in daylight, with the device in either portrait or landscape orientation. Each image is accompanied by additional metadata, including GPS coordinates, device compass direction, intrinsic camera calibration and bounding boxes for people and vehicles. All personally identifiable information of faces and license plates have been redacted.

Fig. 9 shows a visualization of the collected dataset. The map shows the site locations worldwide, and the circle size and color are proportion to the number of images collected at that site. The map highlights three sites in the dataset, an urban high-rise in Dubai, a suburban office park in Woburn, Massachusetts and a rural building in Nairobi. Each example shows a sample of imagery organized as a montage, highlighting the variation in viewpoints covering the site. Sites were revisited over a nine month period including new weather conditions (snow, rain, fog), time of day/illumination changes and seasonal appearance variations. The VisymScenes dataset will be available for download under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

8. VisymScenes Pairs Mining

The VisymScenes dataset includes GPS and device orientation collected while recording imagery. While this metadata can be noisy, it still provides valuable information for identifying potential doppelganger pairs. We develop a series of filtering rules applied to all matched image pairs identified by the COLMAP feature matching module [12]. Note that these rules are designed according to the capture style of the Visym Collector, where cameras consistently focus on the scene of interest and maintain a reasonable distance from it. For a given pair of images, we use the (metric) distance between the camera centers r, angle between their viewing directions θ, and camera intrinsics $\mathcal{K}$, all derived from camera metadata, to identify confident negative (doppelganger) and positive (correct match) pairs. To identify confident negative pairs, we first identify spatially distant matching pairs, and classify them as doppelgangers. Then, we classify the remaining (nearby) pairs into three cases: the intersection point of the viewing directions is 1) in front of both cameras, 2) behind both cameras; or 3) in front of one camera and behind the other. We use the term 'intersection point of the viewing directions' to refer to the point of closest approach between the viewing rays from the two cameras. While the rays typically do not intersect, the computed point represents our best estimate of their intersection. To determine the intersection point of a pair of images (I_p, I_q), let $(\mathbf{d}_p, \mathbf{d}_q)$ be their viewing directions, and $(\mathbf{t}_p, \mathbf{t}_q)$ be their positions, respectively. We define the matrix $\mathbf{A}$ and vector $\mathbf{b}$ as follows:

$$\mathbf{A} = \begin{bmatrix} -\mathbf{d}_p & \mathbf{d}_q \end{bmatrix}, \quad \mathbf{b} = \mathbf{t}_p - \mathbf{t}_q. \tag{5}$$

We need to find the parameter vector that minimizes the least-square error between $\mathbf{Ax}$ and $\mathbf{b}$, denoted by $\mathbf{s}$:

$$\mathbf{s} = \arg\min_{\mathbf{x}} \|\mathbf{Ax} - \mathbf{b}\|_2^2 . \tag{6}$$

which can be solved as a least squares problem. The resulting $\mathbf{s}$ is a two-dimensional vector $\mathbf{s} = [s_1, s_2]$, where s_1 indicates how far along viewing direction $\mathbf{d}_a$ from $\mathbf{t}_a$ is the closest point to the other camera location, with s_2 describing the same with role of the cameras reversed. For case 1), *i.e.* $s_1 > 0$ and $s_2 > 0$, we label pairs with very large view angles (*e.g.*, $> 160^\circ$) as negative, since they likely capture different 3D surfaces or mostly non-overlapping content. For case 2), *i.e.* $s_1 < 0$ and $s_2 < 0$ if their view angle exceeds the diagonal field of view, then their view frustums are unlikely to overlap, so we label them as doppelgangers. For case 3), *i.e.* $s_1 \times s_2 \leq 0$, we check for frustum overlap using camera intrinsics; if no intersection is detected, the pair is considered a doppelganger. Alg. 1 and Alg. 2 explain algorithm details of mining negative and positive matching pairs respectively.

9. Experiment Results

Attention weights visualization. To better understand what our model focuses on when assessing the matching probability of an image pair, we visualize the attention weights of our prediction head on example pairs from landmarks in the MegaScenes dataset [14] in Fig. 10 (top). Notably, traditional SfM methods fail on these three landmarks due to doppelganger issues. While DG-OG can correctly disambiguate the Brandenburg Gate with a threshold $\tau = 0.8$ and

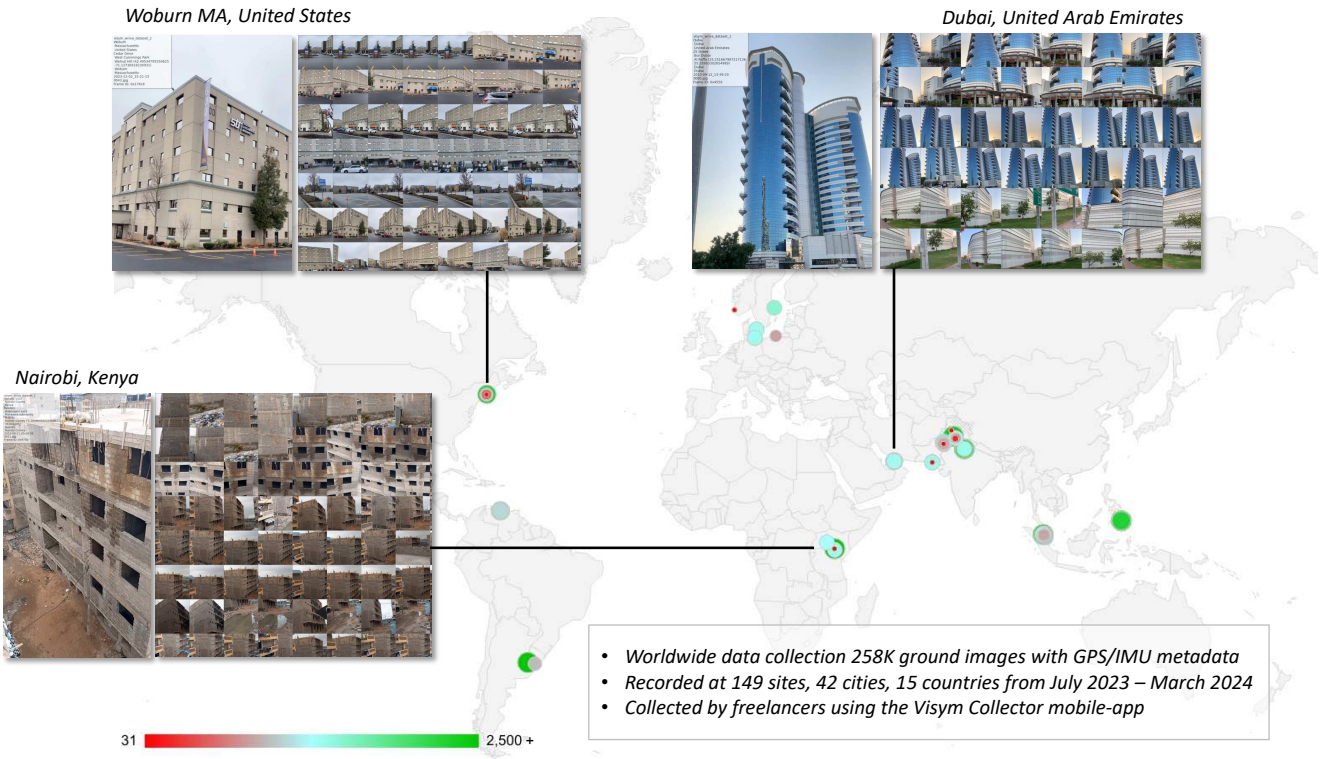

Figure 9. The VisymScenes dataset contains 258K images with GPS/IMU metadata, recorded at 149 sites in 42 cities and 15 countries. This dataset includes everyday rural and suburban scenes designed to complement the notable landmarks available on Wikimedia Commons.

the Siena Cathedral with $\tau = 0.9$, it was unsuccessful with the Belvedere Palace at all thresholds tried. In contrast, our model successfully reconstructs all three landmarks using a common threshold of $\tau = 0.8$. Additionally, in Fig. 10 (bottom), we show pairs of images from the VisymScenes dataset. Here, we observe that our two heads focus on different aspects of each scene, effectively highlighting details that are critical to determining whether the image pairs capture different 3D surfaces. The two head and symmetrized design allows our model to extract complementary clues for accurate disambiguation.

Full Ablations. We study the effectiveness of our design by analyzing the following aspects: 1) fine-tuning the entire model (vs. only new heads), 2) one classification head (vs. two separate heads), 3) the architecture of our classification head, and 4) final-layer decoder features (vs. multi-layer decoder features). We evaluate models trained on DG and VisymScenes datasets. Tab. 3 shows quantitative results.

We also look into the choice of 3D geometric features. Our method uses features from MASt3R [9]. As a comparison, we also train our doppelganger classifier on DUSt3R [16] features whilst keeping all other settings unchanged. We report average precision (AP) of models trained on DG and DG+VisymScenes on the three test sets. Results are shown in Tab. 4. One can see that training with DUSt3R features is not as good as with training with MASt3R features. This can possibly be ascribed to MASt3R's advantage of finding explicit matches over DUSt3R, which benefits our doppelganger detection task. On the other hand, training with DUSt3R still outperforms DG-OG.

More Qualitative Results. We provide additional qualitative results in Fig. 11, Fig. 12 and Fig. 13.

10. Limitations and Future Work

Our approach demonstrates superior performance in visual disambiguation tasks and significantly aids in correcting 3D reconstructions. Extensive experiments have been conducted across various types of scenes, alongside a thorough analysis of our method's design. We observe that performance can be further enhanced with improved 3D geometric features, as evidenced by the experiments in Table 4. From a data perspective, most of the DG and VisymScenes datasets currently consist of ground-level imagery, leaving their adaptability to other sources, such as drone views, uncertain. Incorporating data from such alternative perspectives could further boost performance.

In this work, we highlight our visual disambiguation capability by pruning spurious matches from the structure-from-

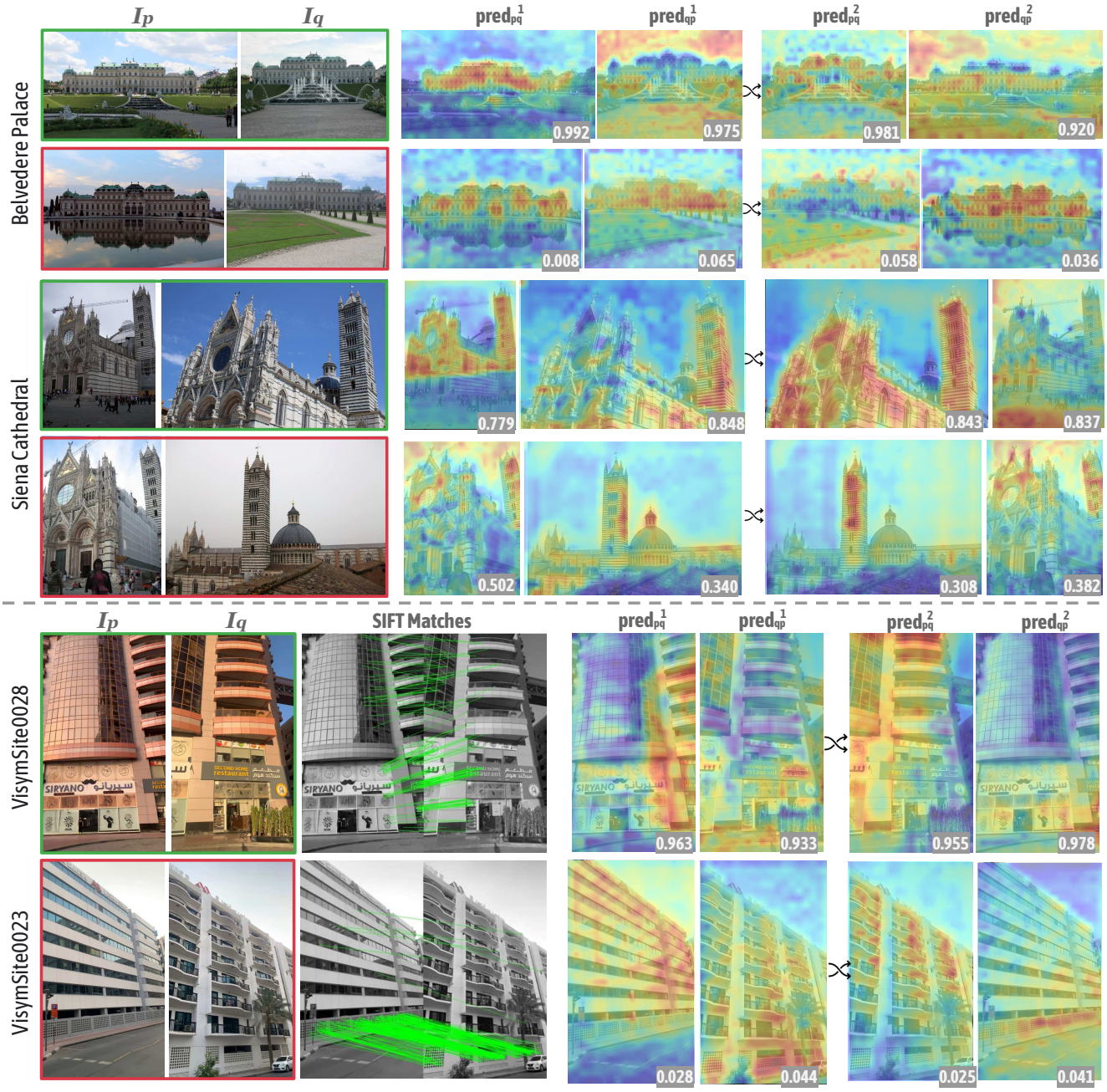

Figure 10. We display example positive and negative pair examples from MegaScenes [14] and VisymScenes. To better understand what the model is attending to, we also visualize the attention map from our two doppelganger classification heads, with their predicted scores labeled on the bottom right of each attention map. The switch symbol indicates symmetrized image pairs. (Top) Belvedere Palace (MegaScenes): DG-OG fails to disambiguate this scene due to subtle differences between distinct surfaces; Siena Cathedral: DG-OG requires a strict threshold at $\tau = 0.9$ to correct this model. Our method can successfully correct these two cases with a consistent threshold $\tau = 0.8$. (Bottom) Challenging cases from VisymScenes data, where DG-OG fails to correctly classify, giving opposite predictions.

motion (SfM) scene graph. Beyond this application, our method also holds potential for improving the training of 3D geometric models like MASt3R. For example, if a pair of input images is classified as doppelgangers, their point maps are unlikely to overlap, nor should they have matches. By leveraging this insight, we can make 3D foundation models more robust and accurate.

Algorithm 1: Negative Pair Selection Algorithm

Input: Distance between views (r)
Viewing angle (θ)
Solution vector ($\mathbf{s} = [s_1, s_2]$)
Diagonal Field of View (ϕ_{dia})
Camera frustums ($\text{FR}_{\text{target}}$, FR_{query})
Output: If two views are doppelganger pairs.

if $r > 70$ **then**
  return *No common scene surface.*
end
else
  if $s_1 > 0$ ***and*** $s_2 > 0$ **then**
    if $\theta > 160°$ **then**
      return *No common scene surface, or have unreliable matches*
    end
  end
  else if $s_1 < 0$ ***and*** $s_2 < 0$ **then**
    if $\theta > \phi_{dia}$ **then**
      return *Non-overlapping views*
    end
  end
  else if $s_1 \times s_2 \leq 0$ **then**
    if *Frustums do not intersect* **then**
      return *Non-overlapping views.*
    end
  end
end

Algorithm 2: Positive Pair Detection Algorithm

Input:
- Distance between views (r)
- Viewing angle (θ)
- Solution vector ($\mathbf{s} = [s_1, s_2]$)
- Horizontal Field of View (ϕ_h)
- Camera frustums ($\text{FR}_{\text{target}}$, FR_{query})

Output: If two views are true match pairs.

if $r < 20$ ***or*** *(*$r < 50$ ***and*** *Frustums intersect)* **then**
  if $s_1 > 0$ ***and*** $s_2 > 0$ ***and*** $\theta < 90°$ **then**
    return *Overlapping views.*
  end
  if $s_1 < 0$ ***and*** $s_2 < 0$ ***and*** $\theta < \phi_h$ **then**
    return *Overlapping views.*
  end
  if $s_1 \times s_2 \leq 0$ ***and*** $\theta < \phi_h$ **then**
    return *Overlapping views.*
  end
end

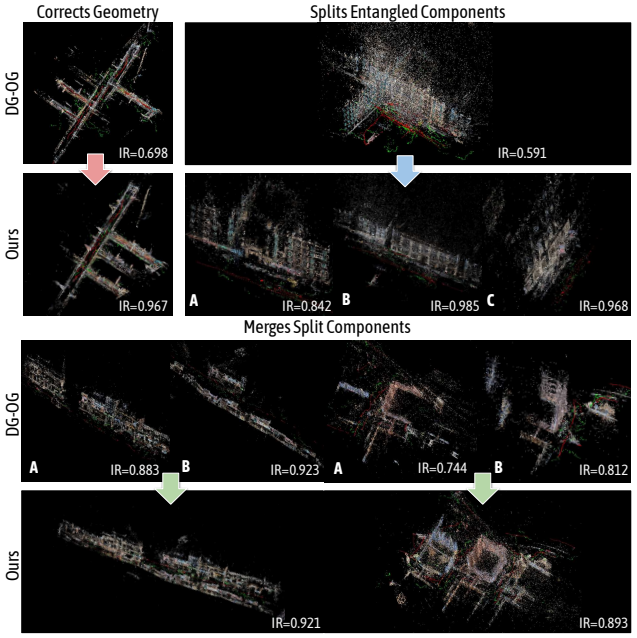

Figure 11. More examples on VisymScenes where DG-OG [3] produces less ideal reconstructions and ours achieves better results. Inlier ratio (higher is better) is reported on the bottom right.

Test Data	Method	Metrics			
		AP ↑	ROC AUC ↑	Precision@Recall=0.85 ↑	Recall@Precision=0.99 ↑
DG	Ours	0.981	0.981	0.982	0.642
	Tune MASt3R	0.982	0.981	0.973	0.728
	Mix two-branch dec feat	0.976	0.975	0.964	0.579
	MLP head	0.973	0.973	0.952	0.543
	Single-layer dec feat	0.970	0.971	0.944	0.496
VisymScenes	Ours	0.991	0.990	0.999	0.901
	Tune MASt3R	0.988	0.987	0.990	0.851
	Mix two-branch dec feat	0.982	0.976	0.974	0.888
	MLP head	0.944	0.914	0.908	0.686
	Single-layer dec feat	0.955	0.934	0.926	0.875
Mapilary	Ours	0.968	0.958	0.942	0.736
	Tune MASt3R	0.967	0.958	0.939	0.396
	Mix two-branch dec feat	0.963	0.947	0.929	0.583
	MLP head	0.906	0.918	0.872	0.333
	Single-layer dec feat	0.918	0.904	0.923	0.513

Table 3. **Full ablation table.** We ablate our designs individually to study their efficacy.

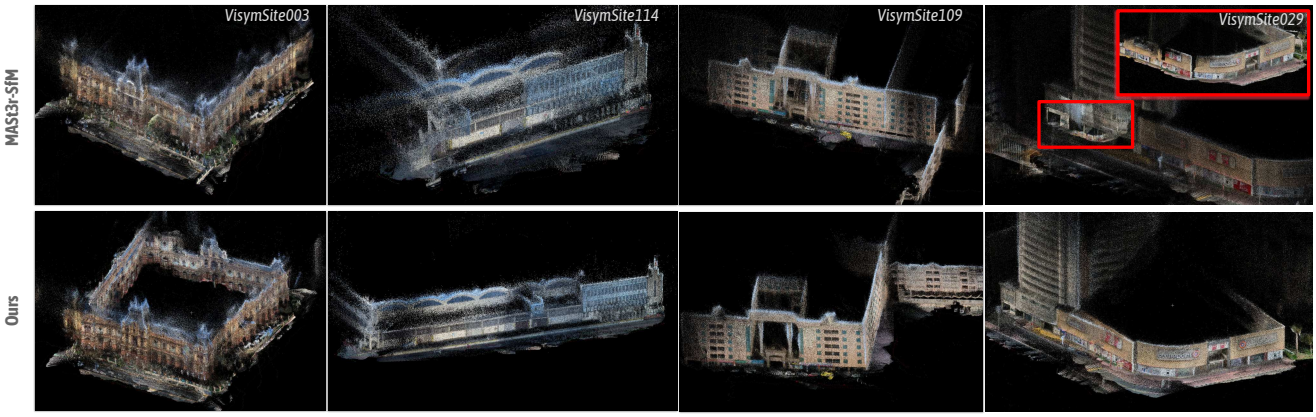

Figure 12. Failure cases of MASt3R-SfM [5] on VisymScenes. Our method can help restore collapsed structure and correct scene geometry. Specific failure of MASt3R-SfM are: different sides of VisymSite003 are collapsed; two ends of VisymSite114 are matched togather; the side of VisymSite109 building is flipped to the frontal side; and VisymSite029 has a nested structure.

Test Data	Method	AP
DG	Ours	0.980 / 0.981
	w/ DUSt3R feat.	0.971 / 0.981
VisymScenes	Ours	0.936 / 0.991
	w/ DUSt3R feat.	0.886 / 0.991
Mapillary	Ours	0.950 / 0.968
	w/ DUSt3R feat.	0.915 / 0.950

Table 4. Comparison of using different 3D geometric features. Ours are trained using MASt3R [9] features, and we switch to DUSt3R features whilst keeping all other settings unchanged. We report metrics on models trained on DG [3] and trained on DG+VisymScenes (separated by '/'). While training with DUSt3R features is not as good as training with MASt3R features, DUSt3R features still outperform the baselines.

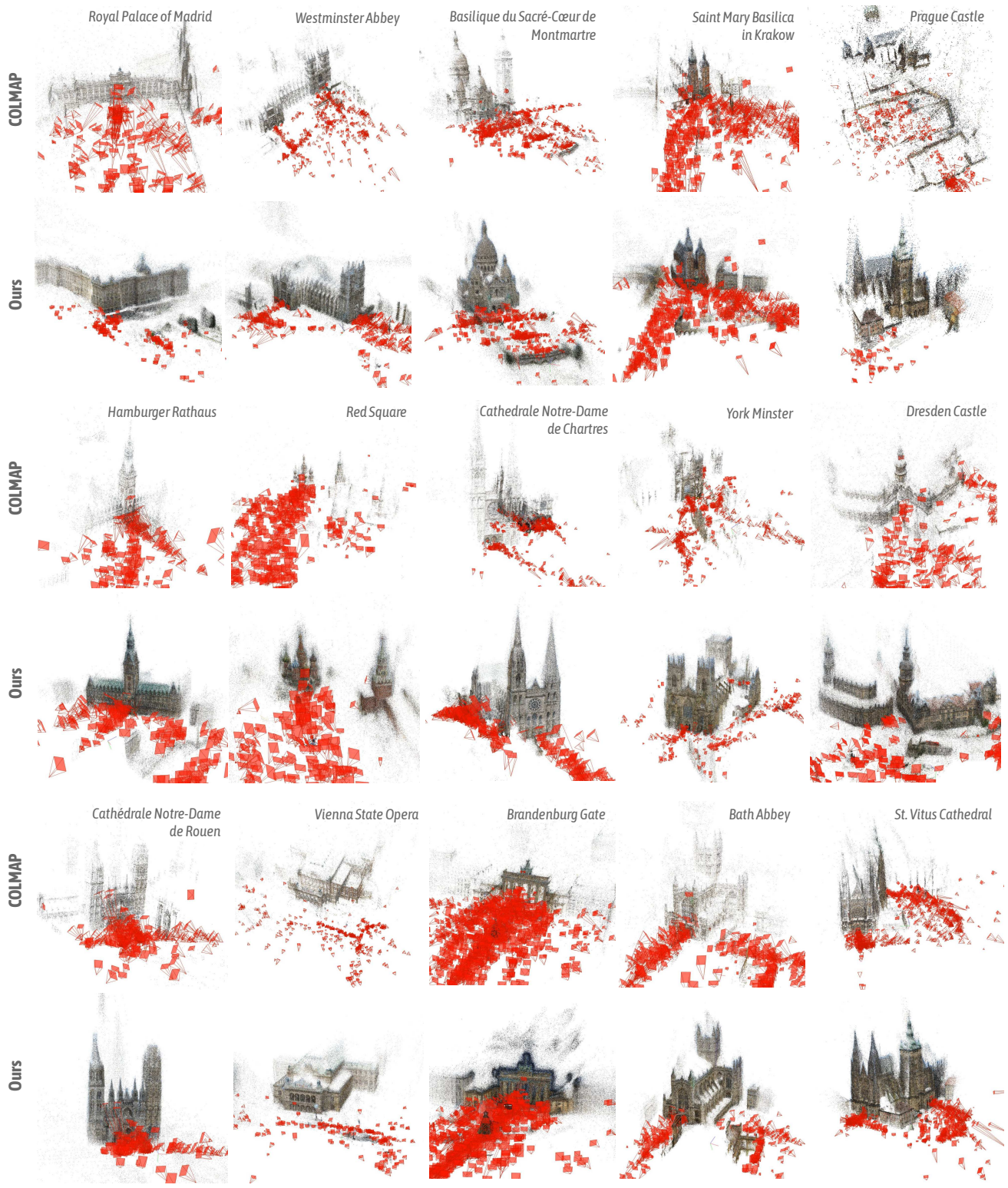

Figure 13. More results on MegaScenes [14] where COLMAP struggles to correctly reconstruct, resulting in entangled components (*e.g.* Hamburger Rathaus, Dresden Castle, Vienna State Opera), ghost structure (*e.g.* Basilique du Sacre Coeur Montmartre, York Minster, St. Vitus Cathedral) etc.