Visual Localization using Imperfect 3D Models from the Internet
-
Supplementary Material

Vojtech Panek[1,2] Zuzana Kukelova[3] Torsten Sattler[2]
[1] Faculty of Electrical Engineering, Czech Technical University in Prague
[2] Czech Institute of Informatics, Robotics and Cybernetics, Czech Technical University in Prague
[3] Visual Recognition Group, Faculty of Electrical Engineering, Czech Technical University in Prague
{vojtech.panek,torsten.sattler}@cvut.cz kukelzuz@fel.cvut.cz

This supplementary material is organized as follows: Sec. 1 describes the practical issues we encountered with the downloaded Internet models, mentioned in Sec. 3 of the main paper. Sec. 2 shows mean and maximum Dense Correspondence Re-Projection Error (DCRE) plots for both versions of the ground truth (*cf*. Sec. 5 of the main paper). Sec. 3 describes in detail the setup of the experiments on the evaluation of geometric fidelity. It shows extended results from Sec. 5 of the main paper for both the experiment with the database of real images and the database images rendered from a stretched 3D model. Sec. 4 presents experiment on the influence of simplification of model geometry and appearance on localization accuracy.

1. Issues with the Internet Models

In this section, we describe the practical issues we encountered when collecting the models described in Sec. 3 in the main paper. For convenience we replicate the Tab. 1 from the main paper in Tab. 1 and the individual scenes from Fig. 2 from the original paper in Fig. 14 (Notre Dame), Fig. 15 (Pantheon), Fig. 16 (Reichstag), Fig. 17 (St. Peter's Square), Fig. 18 (St. Vitus Cathedral) and Fig. 19 (Aachen).

All models downloaded from 3D Warehouse were created using the SketchUp 3D modeling software. The software allows mesh faces to have a material (color or texture) from both sides. Other formats, *e.g*., Wavefront OBJ, are not able to represent the double-sided textures, and therefore the texture of the face's back side is lost during the format conversion. The problem can be prevented during modeling time by orienting the front side of the faces outwards from the model and assigning the texture only to those. This is considered good practice in the SketchUp community; however a large fraction of the models we downloaded did not follow it. The same problem prevented us from extracting textures from Reichstag model F and St. Peters Square model D.

Few other models (Pantheon C and D) contained generic textures, which did not correspond to reality (see Fig. 1). Therefore we decided to use just the raw geometry of these models.

2. Localization Accuracy with Internet Models and Used Metrics

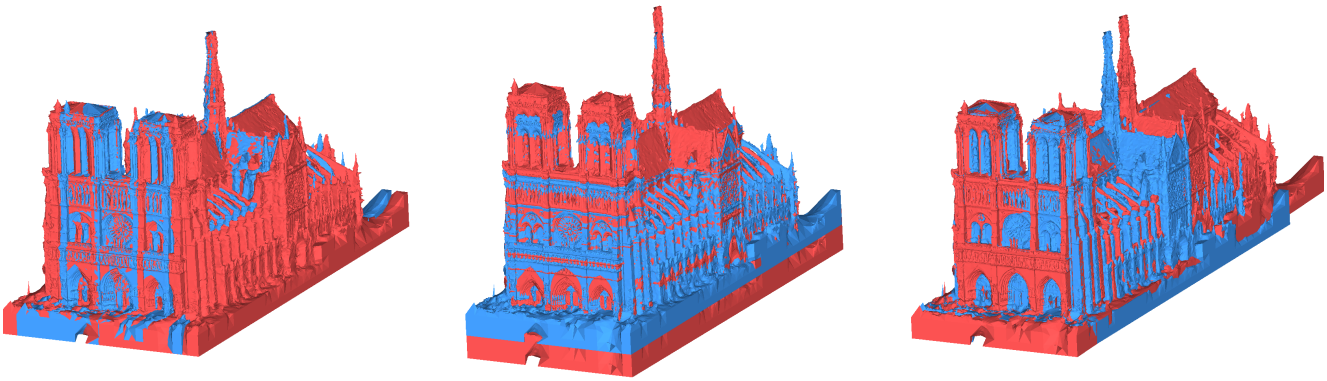
Sec. 5 of the main paper focused on showing results for the mean Dense Correspondence Re-Projection Error (mean DCRE) for both the global alignment (GA) and local refinement (LR) versions of the ground truth. For maximum DCRE results, Sec. 5 pointed to the supplementary material. These results will be presented in this section. For convenience and to facilitate easier comparisons, the following shows results obtained using the MeshLoc pipeline [8] for both possible DCRE aggregation functions (mean and maximum) and both ground truth methods (global alignment (GA) and local refinement (LR)) (*cf*. Sec. 4 of the main paper). We replicate the first row of Fig. 4 from the main paper, which shows the experiments with mean DCRE aggregation and global alignment (GA) ground truth in Fig. 4, and the second row of Fig. 4 of the main paper, which shows mean DCRE and local refinement (LR), in Fig. 6. Fig. 5 shows max DCRE and global alignment (GA) results. Fig. 7 shows max DCRE and local refinement (LR) results.

Regarding the difference between mean DCRE and maximum DCRE curves, we can observe two types of behavior. For the first type, the maximum DCRE does not alter much from the mean DCRE curve (*e.g*., Notre Dame A, B, and E). For the second type, the drop is much more significant (*e.g*., Notre Dame C, D). The first type corresponds to the models with more accurate geometry (see Fig. 3 in the main paper).

Naturally, measuring the maximum instead of the mean DCRE per image leads to lower performance. Still, the

(a) The real state of the landmark (b) Pantheon - model C (c) Pantheon - model D

Figure 1. Comparison of the real state of the Pantheon landmark to models containing generic textures.



(a) Stretching model width (120%) (b) Stretching model height (120%) (c) Stretching model depth (120%)

Figure 2. Visualization of the non-uniform scaling used for the evaluation of the impact of geometric fidelity. Original model in blue, scaled model in red.

Scene	ID	Model author	Model source	License	Model type	Color type	Size [MB]
Notre Dame (Front Facade) [6, 7, 13] 189 queries	A	Miguel Bandera	Sketchfab (https://skfb.ly/6QWu7)	CC BY-NC-SA 4.0	MVS	texture	22.4
	B	Chigirinsky	Sketchfab (https://skfb.ly/6Rn9M)	CC BY 4.0	CAD	texture	4.8
	C	Alejandro Diaz	Sketchfab (https://skfb.ly/31dba)	CC BY 4.0	CAD	texture	3.1
	D	Little-Goomba	3D Warehouse (https://bit.ly/3QWeOxY)	3DW: GML	CAD	texture	0.6
	E	MiniWorld3D	MyMiniWorld (https://mmf.io/o/91899)	BY-ND-NC-EX	CAD	raw	30.4
	F	giotis	3D Warehouse (https://bit.ly/3QOTQ41)	3DW: GML	CAD	raw	0.7
	G	Jul	3D Warehouse (https://bit.ly/3ThrhOt)	3DW: GML	CAD	raw	0.1
Pantheon (Exterior) [6, 7, 13] 141 queries	A	Fovea	Sketchfab (https://skfb.ly/6RZHt)	CC BY 4.0	MVS	texture	84.2
	B	brnipon	3D Warehouse (https://bit.ly/3CCwTwP)	3DW: GML	CAD	texture	5.5
	C	Ultima Ratio	3D Warehouse (https://bit.ly/3AuN2BK)	3DW: GML	CAD	raw	61.5
	D	Adsman007	3D Warehouse (https://bit.ly/3ASv10b)	3DW: GML	CAD	raw	24.5
	E	Emanuele Viani	Sketchfab (https://skfb.ly/EAKB)	CC BY 4.0	CAD	raw	36.0
Reichstag [6, 7, 13] 75 queries	A	Emperor Heer 99	3D Warehouse (https://bit.ly/3wzFhci)	3DW: GML	CAD	texture	13.3
	B	Emperor Heer 99	3D Warehouse (https://bit.ly/3ATX2Wk)	3DW: GML	CAD	texture	7.5
	C	Emperor Heer 99	3D Warehouse (https://bit.ly/3ASdFSr)	3DW: GML	CAD	texture	27.3
	D	Emperor Heer 99	3D Warehouse (https://bit.ly/3ctJvvp)	3DW: GML	CAD	texture	6.0
	E	Klaus T.	3D Warehouse (https://bit.ly/3cme5qV)	3DW: GML	CAD	raw	5.3
	F	SH	3D Warehouse (https://bit.ly/3AP4CBM)	3DW: GML	CAD	raw	0.1
St. Peter's Square [6, 7, 13] 126 queries	A	Brian Trepanier	Sketchfab (https://skfb.ly/or8Ip)	CC BY 4.0	MVS	texture	230.1
	B	Dounia B.	3D Warehouse (https://bit.ly/3CCAYkk)	3DW: GML	CAD	texture	131.5
	C	mstochl	3D Warehouse (https://bit.ly/3RhqEmc)	3DW: GML	CAD	texture	4.2
	D	Antonino G.	3D Warehouse (https://bit.ly/3Rd3KMC)	3DW: GML	CAD	raw	24.5
St. Vitus Cathedral (own data) 213 queries	A	Brian Trepanier	Sketchfab (https://skfb.ly/o8n8D)	CC BY 4.0	MVS	texture	109.4
	B	Brian Trepanier	Sketchfab (https://skfb.ly/o8n8D)	CC BY 4.0	MVS	texture	284.9
	C	Pera	3D Warehouse (https://bit.ly/3Tf7Bum)	3DW: GML	CAD	texture	3.7
	D	Hrusak	3D Warehouse (https://bit.ly/3Rja6tJ)	3DW: GML	CAD	texture	1.0
Aachen [10, 11, 14], 1015 queries	A	[5]	-	CC BY-NC-SA 4.0	CAD	texture	21.6

Table 1. We show Tab. 1 from the main paper for convenience. List of scenes and 3D models used for the evaluation. The query images for the scenes were obtained from the Image Matching Challenge (IMC) 2021 [6, 7, 13], the Aachen Day-Night v1.1 dataset [10, 11, 14], and our own recordings. We distinguish between models directly created from images via MVS and models created from human input (CAD).

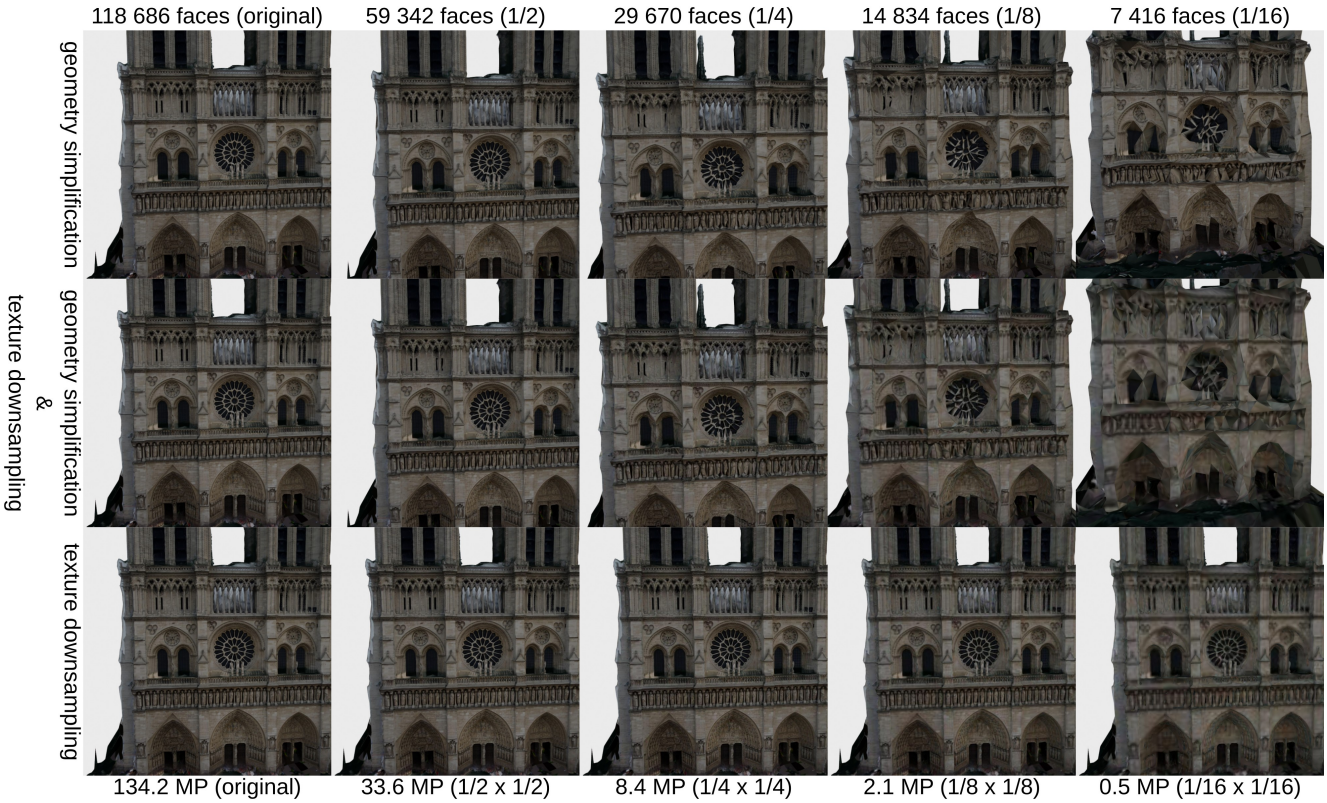


Figure 3. Renderings of the simplified Notre Dame A model. The first column contains renderings of the original model. All other columns correspond to models with reduced geometric and / or texture detail.

relative ranking between different models is mostly preserved, especially for more accurate poses with a mean / max. DCRE of 15% of the image diagonal or smaller.

3. Isolating the Impact of Geometric Fidelity

To isolate the impact of geometric fidelity on localization performance, we experimented with changing the geometry used in the MeshLoc pipeline while fixing the appearance. To this end, we matched each query image against other real images (*cf*. Sec. 5 of the main paper). Both image retrieval with AP-GeM [4, 9] descriptors and local feature matching is done using a database of real images. Renderings of the different 3D models are only used to obtain the depth maps used by MeshLoc [8] to establish 2D-3D matches.

All IMC (Image Matching Challenge) 2021 [6, 7, 13] scenes, except Reichstag, contain a large number of images, from which we use only a small part as queries (*cf*. Tab. 1 for the resulting sizes of the query subsets). The rest of the images were used as an image database in this experiment. For Notre Dame and St. Peters Square, we use every 20th image as a query, for Pantheon every 10th, and for St. Vitus Cathedral every 4th. The Reichstag scene contains only 75 images; therefore, we decided to use all of them as queries and performed the experiment in a leave-one-out manner, *i.e*., when localizing one of the images, we used all the other queries as the image database.

We show the results with mean DCRE and global alignment (GA) ground truth in Fig. 8 (which is a reproduction of Fig. 6 from the main paper) and with local refinement (LR) in Fig. 10. We also show maximum DCREs with GA in Fig. 9 and with LR in Fig. 11. Tab. 1 associates the model IDs to the individual models.

Compared to the results presented in Sec. 2, the gap between the mean and maximum DCRE curves is significantly smaller when matching against real images and only using the 3D geometry of the Internet models (in the form of rendered depth maps). As already mentioned in the main paper, the results show that finding sufficiently many matches to facilitate accurate pose estimation seems to be the main bottleneck, even if the underlying geometry is rather coarse. Thus, we can attribute the majority of the outliers skewing the maximum DCRE curves shown in Sec. 2 to the feature matching stage.

To directly observe the influence of the geometric fidelity, we further experimented with non-uniformly scaling the most precise model (Notre Dame A) to measure the impact of a changing aspect ratio on localization accuracy. The non-uniform scaling in width and height was done relative to the center of the model bounding box. The scal-

ing in depth direction had a center in the main plane of the building's front facade (see Fig. 2 for visualization). Fig. 12 extends the results from Fig. 7 in the main paper by using intermediate scaling factors. The main conclusion drawn in the paper, that changing the aspect ratio significantly reduces localization accuracy, remains valid.

4. Ablation Study: Simplifying the Representation

To better understand the influence of the level of geometric and visual fidelity on localization accuracy, we experiment with reducing the geometric resolution (number of faces) and texture resolution of the Notre Dame A model already used in the main paper for ablation studies.

We used Quadric Mesh Collapse Decimation algorithm [2, 3], implemented in MeshLab [1], for geometry simplification. Note that even when we use the version of the algorithm that is supposed to be more suitable for meshes with textures [3], the textures are significantly distorted during the simplification. Therefore, the appearance fidelity is not completely isolated from the geometric simplification. The other way around, the reduction of texture resolution does not influence the geometry of the model. The texture simplification was done by downsampling all the texture files in the model. We also combined the models with the simplified geometry with the downsampled textures at the same simplification ratio, *e.g.*, the model with half the number of the original faces is combined with the textures with half the original width and height.

The renderings of the simplified models are shown in Fig. 3. Note the distortion of the texture present in the models with higher levels of geometry simplification. We did not observe such artifacts in models available on the Internet. As such, the results obtained for these severe distortions are not indicative of real-world performance.

The localization results of the MeshLoc [8] pipeline with LoFTR [12] are show in Fig. 13. The localization method uses both the rendered images and depth maps. The localization pipeline copes surprisingly well, even with very high levels of geometric and appearance simplification. We can see a major drop in accuracy only after the combination of simplified geometry and downsampled texture at a very high simplification ratio of 1/16. Note that the simplified model is significantly more compact than the original one (18.4 MB original vs. 0.7 MB at 1/16 ratio), which suggests a potential use of the simplified meshes as very compact scene representations.

Note that the experiment was done using a MVS mesh reconstructed from images. Automatically simplifying the geometry of manually created CAD models can result in a complete collapse of the model geometry even for very low simplification ratios, as the CAD models are often composed of a small number of planar walls.

References

[1] Paolo Cignoni, Marco Callieri, Massimiliano Corsini, Matteo Dellepiane, Fabio Ganovelli, and Guido Ranzuglia. MeshLab: an Open-Source Mesh Processing Tool. In *Eurographics Italian Chapter Conference*, 2008. 4

[2] Michael Garland and Paul S Heckbert. Surface simplification using quadric error metrics. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, pages 209–216, 1997. 4

[3] Michael Garland and Paul S Heckbert. Simplifying surfaces with color and texture using quadric error metrics. In *Proceedings Visualization'98 (Cat. No. 98CB36276)*, pages 263–269. IEEE, 1998. 4

[4] Albert Gordo, Jon Almazan, Jerome Revaud, and Diane Larlus. End-to-end learning of deep visual representations for image retrieval. *International Journal of Computer Vision*, 124(2):237–254, 2017. 3

[5] Martin Habbecke and Leif Kobbelt. Linear Analysis of Nonlinear Constraints for Interactive Geometric Modeling. *Comput. Graph. Forum*, 31(2pt3):641–650, 2012. 2

[6] Jared Heinly, Johannes L. Schönberger, Enrique Dunn, and Jan-Michael Frahm. Reconstructing the world* in six days. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3287–3295, 2015. 2, 3

[7] Yuhe Jin, Dmytro Mishkin, Anastasiia Mishchuk, Jiri Matas, Pascal Fua, Kwang Moo Yi, and Eduard Trulls. Image Matching across Wide Baselines: From Paper to Practice. *IJCV*, 2021. 2, 3

[8] Vojtech Panek, Zuzana Kukelova, and Torsten Sattler. Meshloc: Mesh-based visual localization. In *ECCV*, 2022. 1, 3, 4, 13

[9] Jerome Revaud, Jon Almazán, Rafael S Rezende, and Cesar Roberto de Souza. Learning with average precision: Training image retrieval with a listwise loss. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5107–5116, 2019. 3

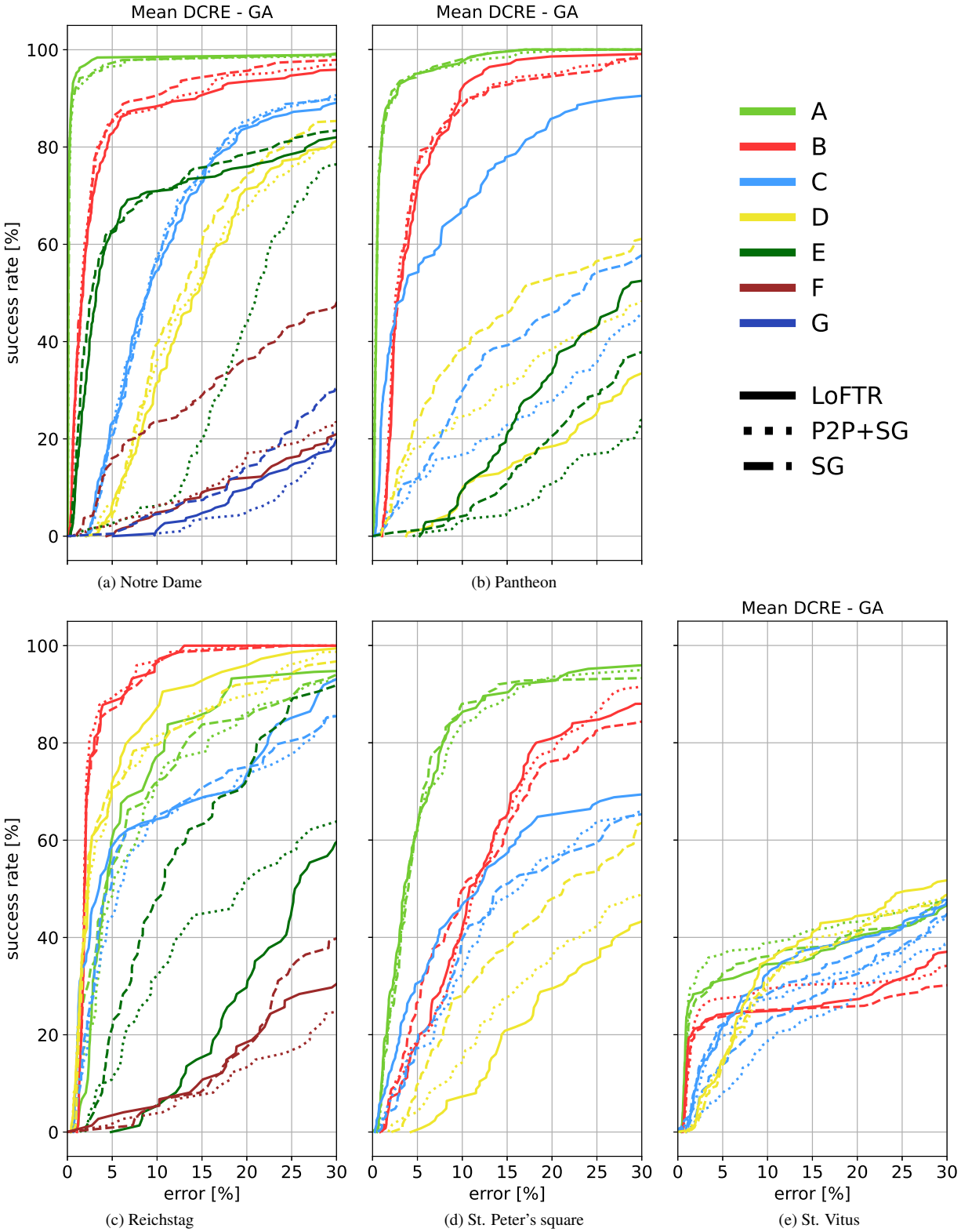
[10] Torsten Sattler, Will Maddern, Carl Toft, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, Fredrik Kahl, and Tomas Pajdla. Benchmarking 6DOF Urban Visual Localization in Changing Conditions. In *CVPR*, 2018. 2

[11] Torsten Sattler, Tobias Weyand, Bastian Leibe, and Leif Kobbelt. Image Retrieval for Image-Based Localization Revisited. In *BMVC*, 2012. 2

[12] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-Free Local Feature Matching With Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 4

[13] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl S. Ni, Douglas N. Poland, Damian Borth, and Li-Jia Li. YFCC100M: the new data in multimedia research. *Commun. ACM*, 59:64–73, 2016. 2, 3

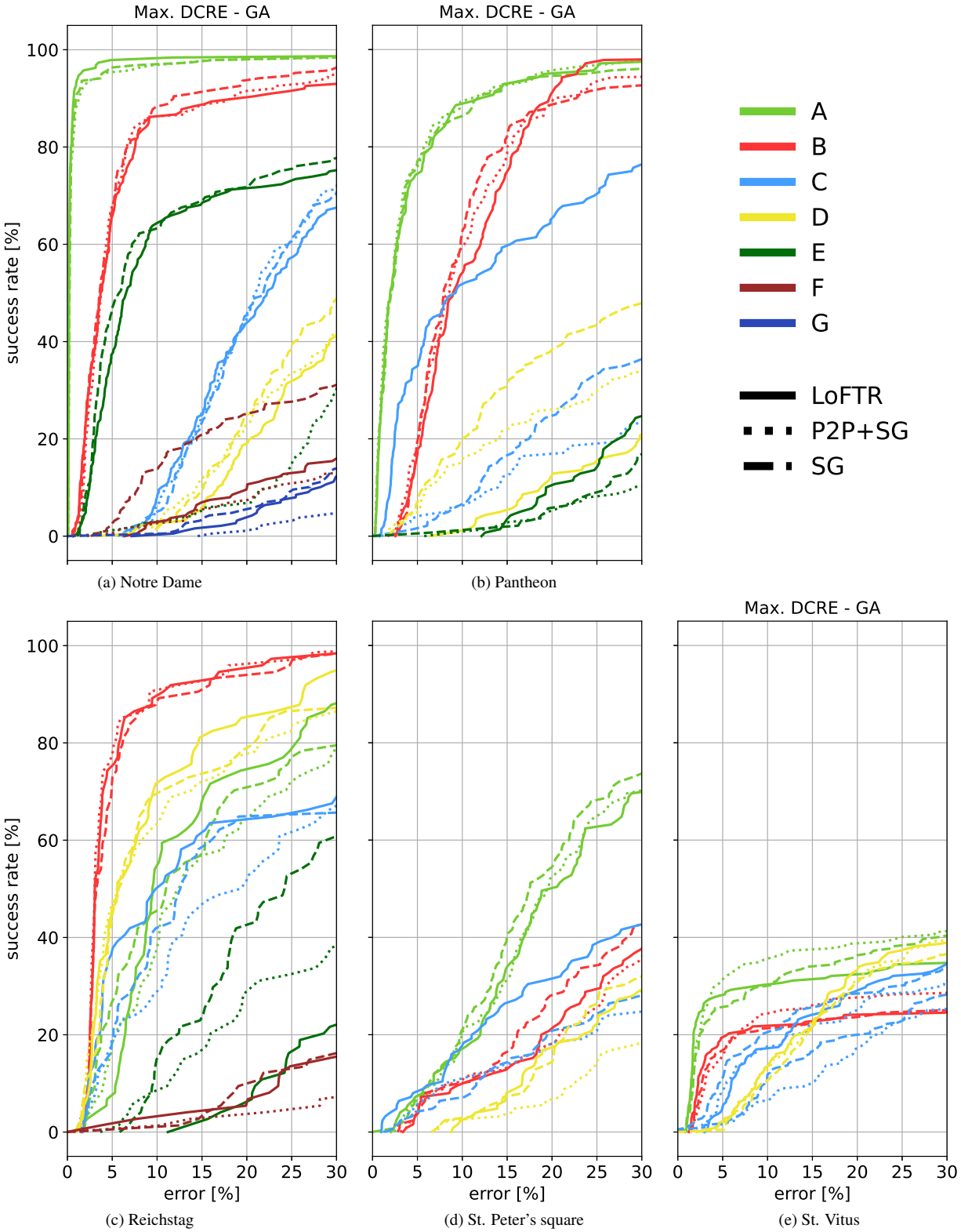
[14] Zichao Zhang, Torsten Sattler, and Davide Scaramuzza. Reference Pose Generation for Long-term Visual Localization via Learned Features and View Synthesis. *IJCV*, 2020. 2



(a) Notre Dame (b) Pantheon

(c) Reichstag (d) St. Peter's square (e) St. Vitus

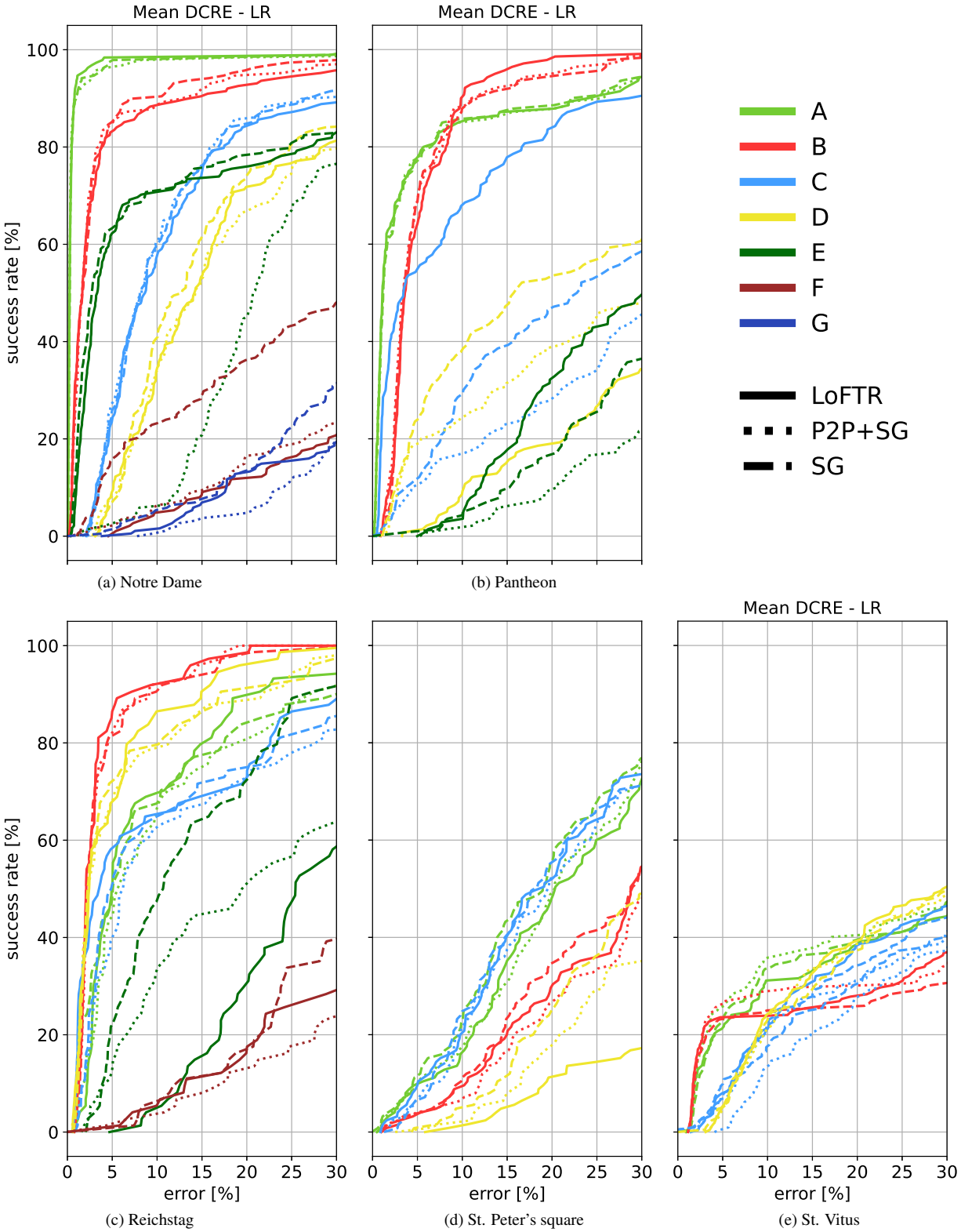
Figure 4. Reproduction of the first row of Fig. 4 from the main paper. Cumulative histograms of the mean DCRE over all query images in a scene for the ground truth poses obtained via global alignment (GA). We show the DCRE as the percentage of the image diagonal.



(a) Notre Dame (b) Pantheon

(c) Reichstag (d) St. Peter's square (e) St. Vitus

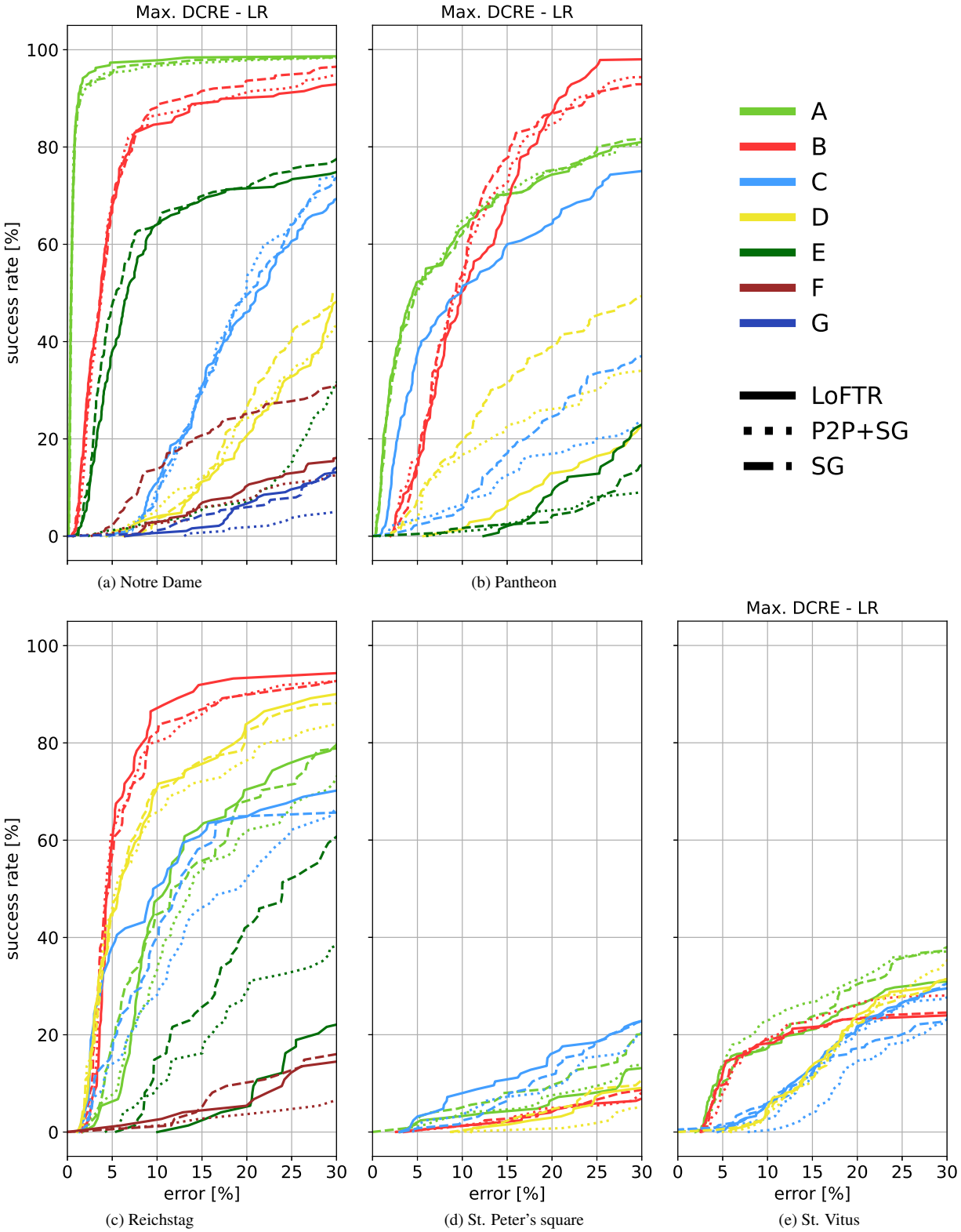
Figure 5. Cumulative histograms of the maximum DCRE over all query images in a scene for the ground truth poses obtained via global alignment (GA). We show the DCRE as the percentage of the image diagonal.



(a) Notre Dame (b) Pantheon

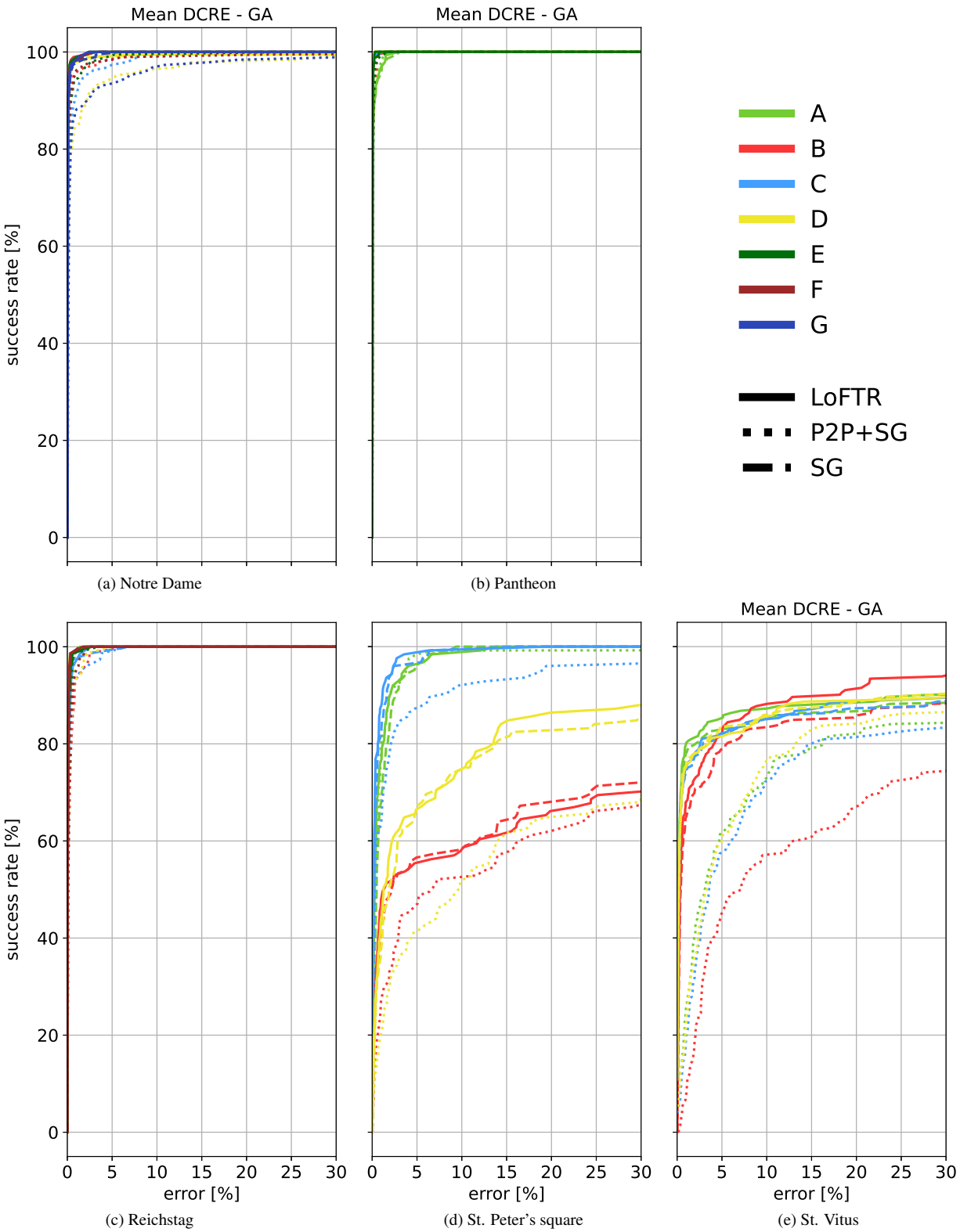
(c) Reichstag (d) St. Peter's square (e) St. Vitus

Figure 6. Reproduction of the second row of Fig. 4 from the main paper. Cumulative histograms of the mean DCRE over all query images in a scene for the ground truth poses obtained via local refinement (LR). We show the DCRE as the percentage of the image diagonal.



(a) Notre Dame (b) Pantheon (c) Reichstag (d) St. Peter's square (e) St. Vitus

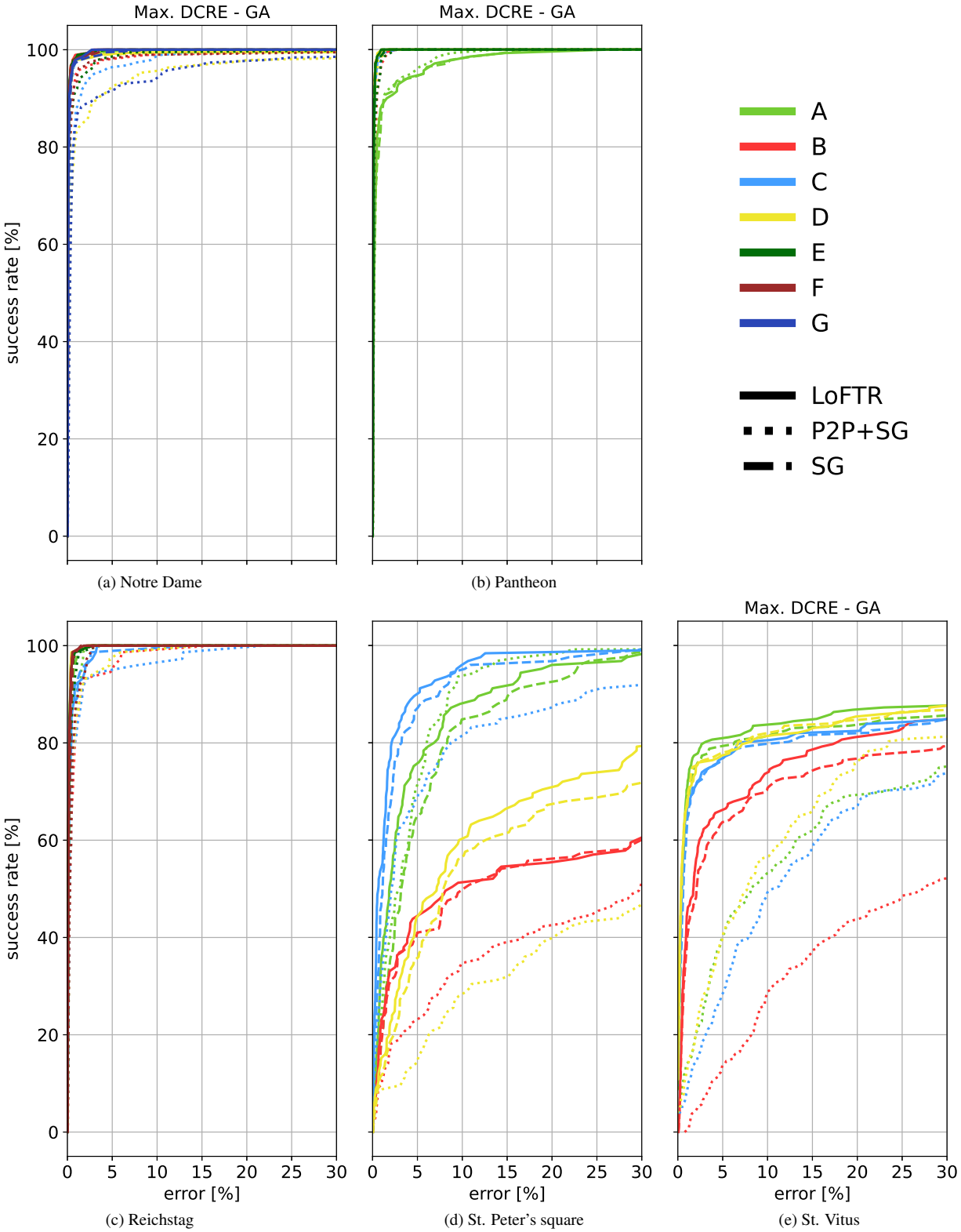
Figure 7. Cumulative histograms of the maximum DCRE over all query images in a scene for the ground truth poses obtained via local refinement (LR). We show the DCRE as the percentage of the image diagonal.



(a) Notre Dame (b) Pantheon

(c) Reichstag (d) St. Peter's square (e) St. Vitus

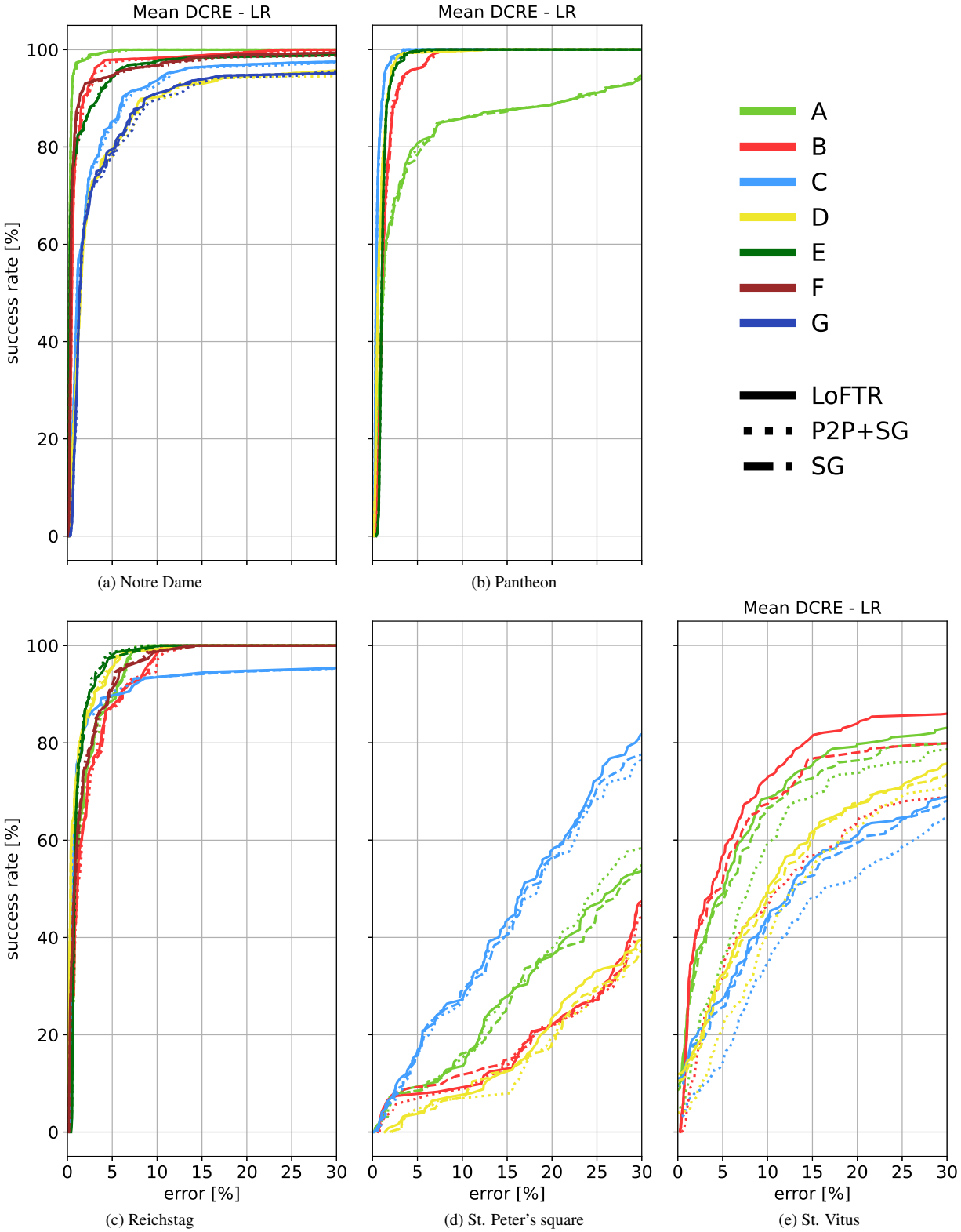
Figure 8. Isolating the impact of geometric fidelity by combining real images with geometry from the Internet models. We show cumulative histograms of the mean DCRE, as a percentage of the image diagonal, over all query images in a scene for the ground truth poses obtained via global alignment (GA).



(a) Notre Dame (b) Pantheon

(c) Reichstag (d) St. Peter's square (e) St. Vitus

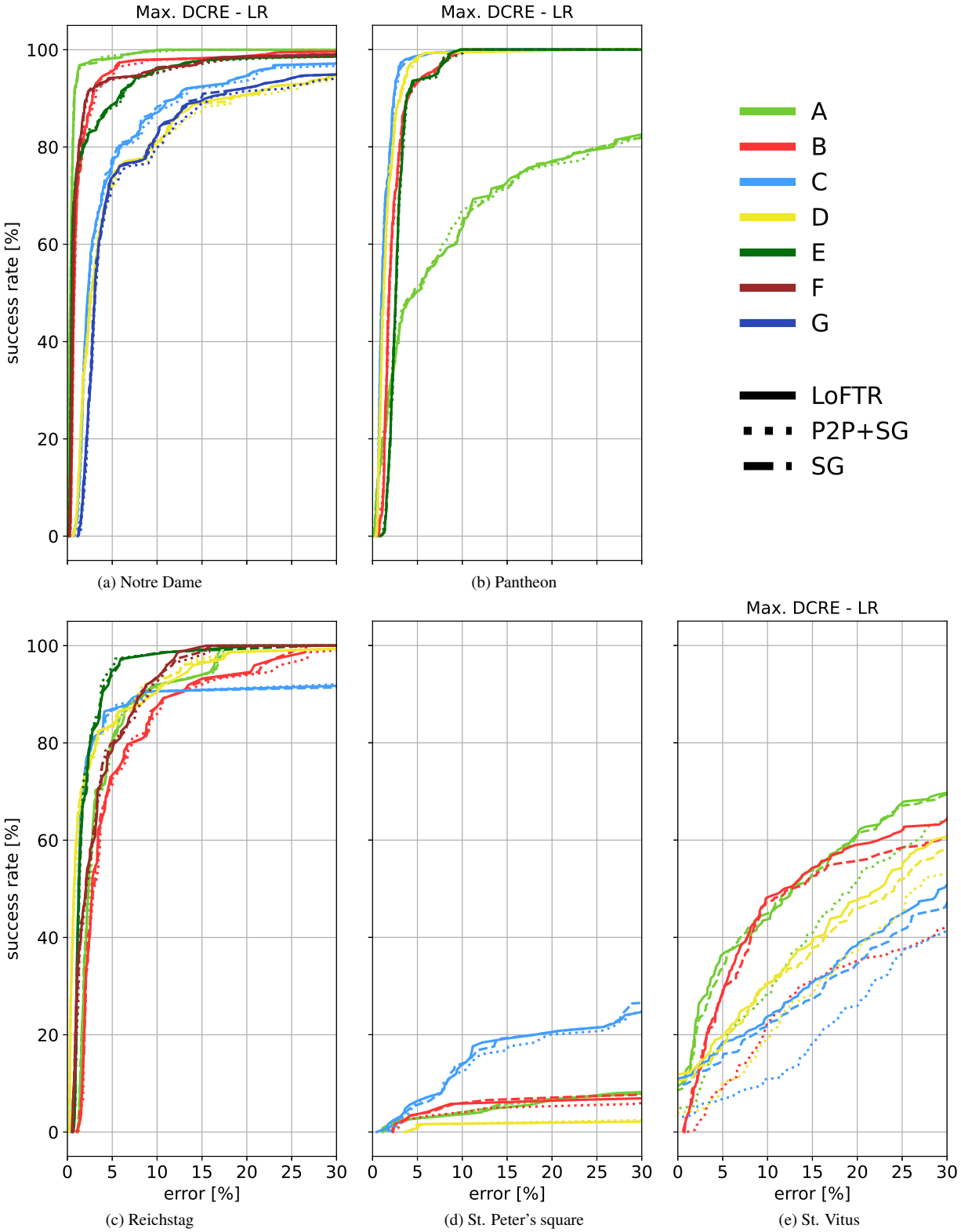
Figure 9. Isolating the impact of geometric fidelity by combining real images with geometry from the Internet models. We show cumulative histograms of the maximum DCRE, as a percentage of the image diagonal, over all query images in a scene for the ground truth poses obtained via global alignment (GA).



(a) Notre Dame (b) Pantheon

(c) Reichstag (d) St. Peter's square (e) St. Vitus

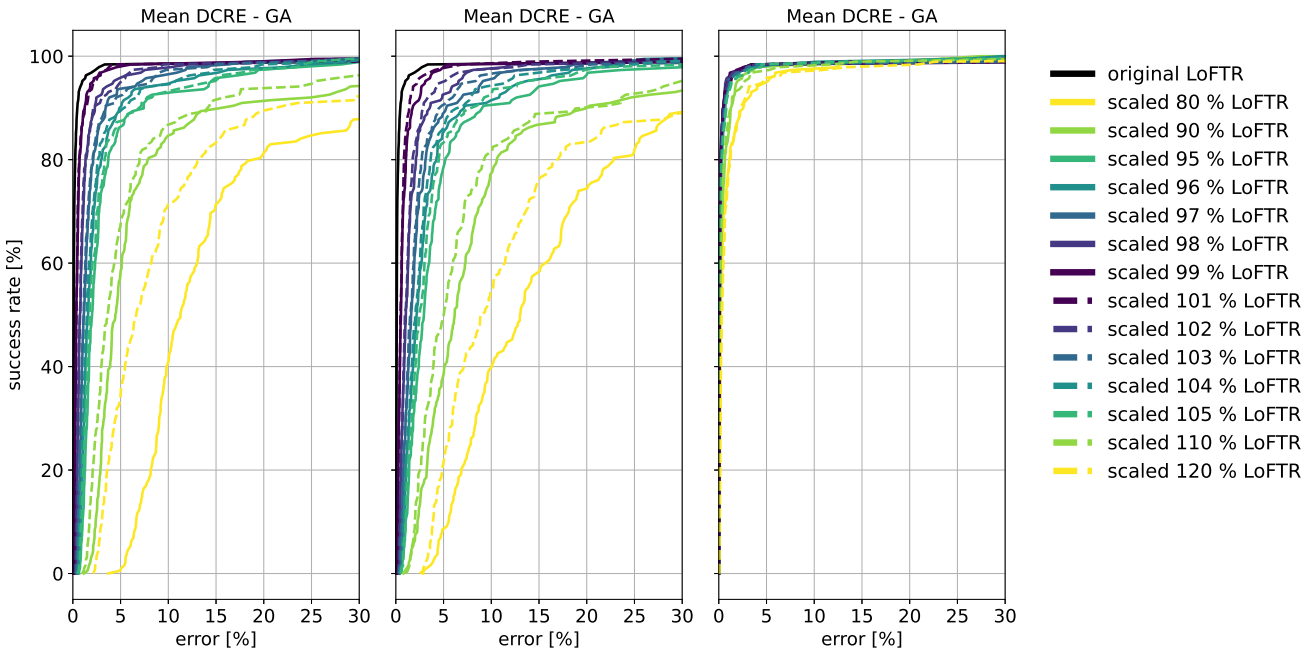
Figure 10. Isolating the impact of geometric fidelity by combining real images with geometry from the Internet models. We show cumulative histograms of the mean DCRE, as a percentage of the image diagonal, over all query images in a scene for the ground truth poses obtained via local refinement (LR).



(a) Notre Dame (b) Pantheon

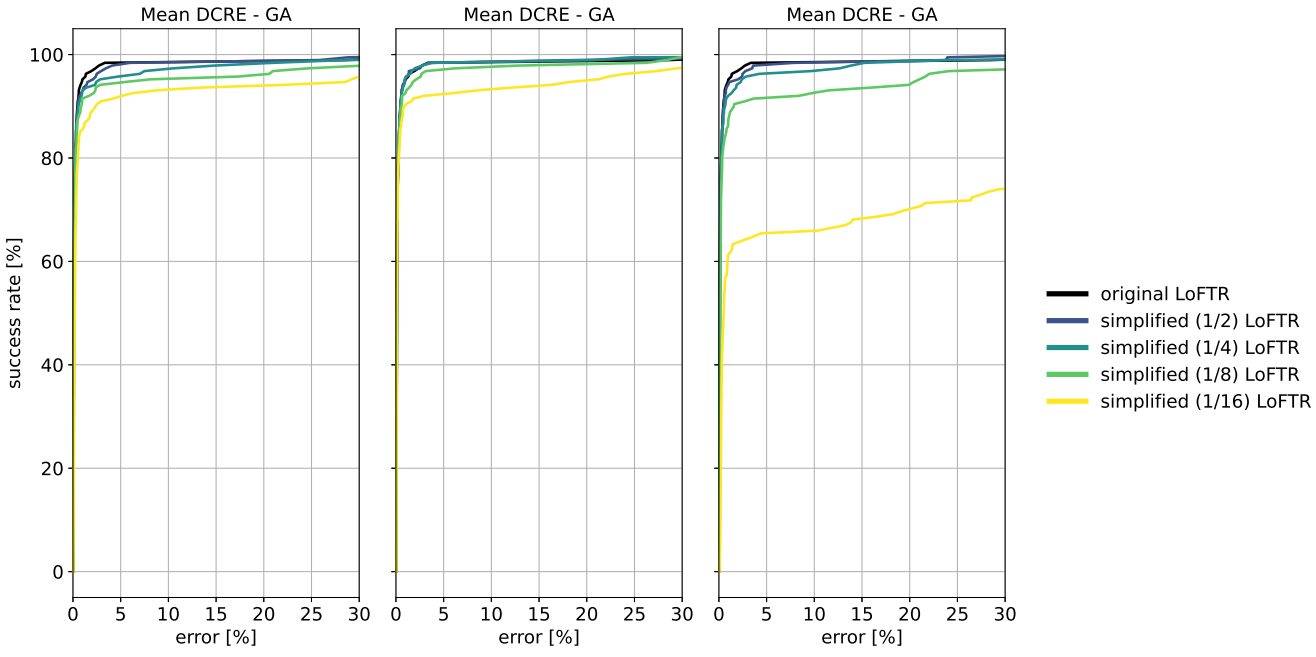
(c) Reichstag (d) St. Peter's square (e) St. Vitus

Figure 11. Isolating the impact of geometric fidelity by combining real images with geometry from the Internet models. We show cumulative histograms of the maximum DCRE, as a percentage of the image diagonal, over all query images in a scene for the ground truth poses obtained via local refinement (LR).



(a) width scaling (b) height scaling (c) depth scaling

Figure 12. Results from Fig. 7 in the main paper, using more intermediate scaling steps. Isolating the impact of geometric fidelity by applying non-uniform scaling on the 3D model. We show cumulative histograms of the mean DCRE, as a percentage of the image diagonal, over all query images in a scene for the ground truth poses obtained via global alignment (GA).



(a) geometry simplification (b) texture downsampling (c) simplification of both

Figure 13. Localization performance of MeshLoc [8] using LoFTR when reducing the geometric and / or texture resolution of the Notre Dame A model.

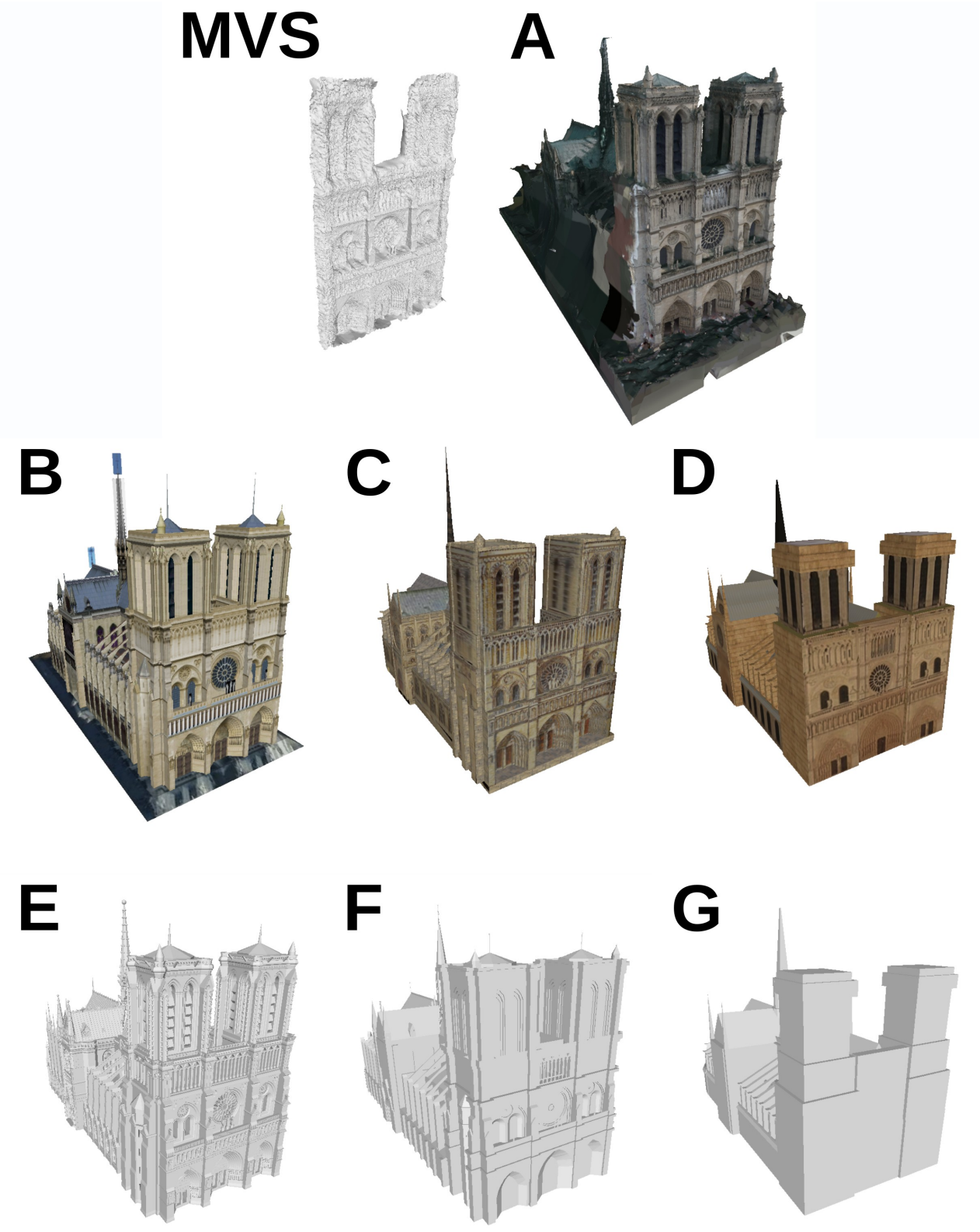


Figure 14. Enlarged Notre Dame models from Fig. 2 in the main paper.

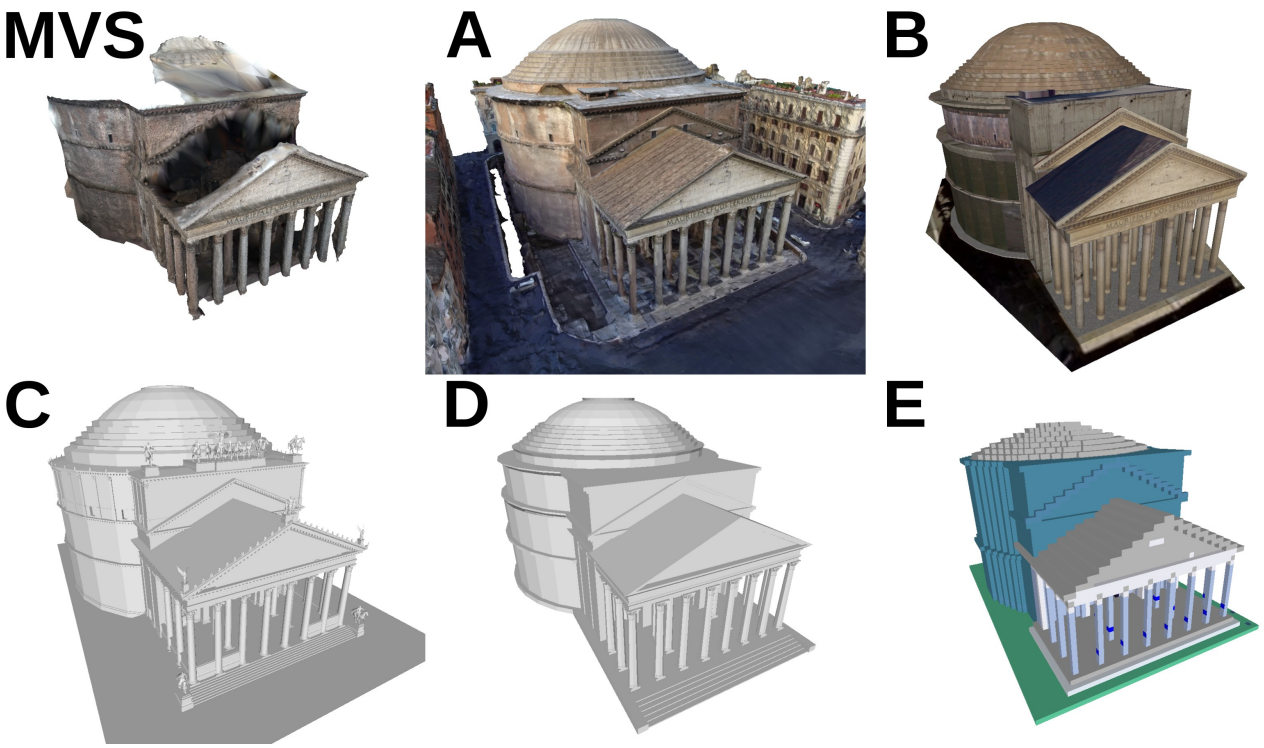


Figure 15. Enlarged Pantheon models from Fig. 2 in the main paper.

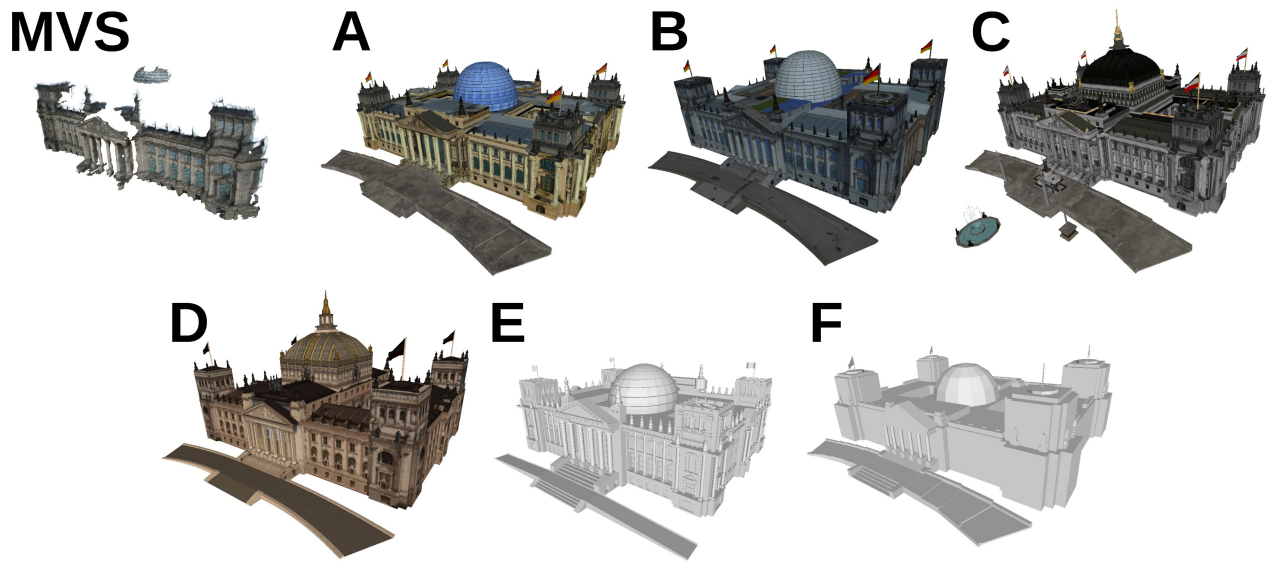


Figure 16. Enlarged Reichstag models from Fig. 2 in the main paper.

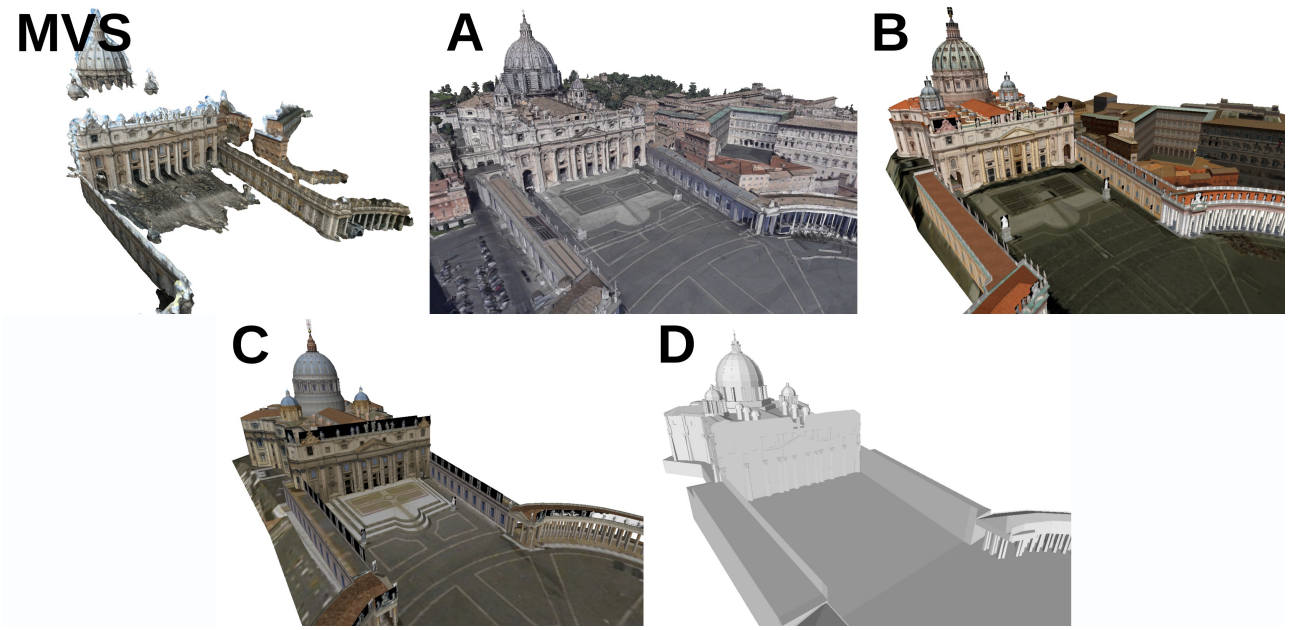


Figure 17. Enlarged St. Peter's Square models from Fig. 2 in the main paper.

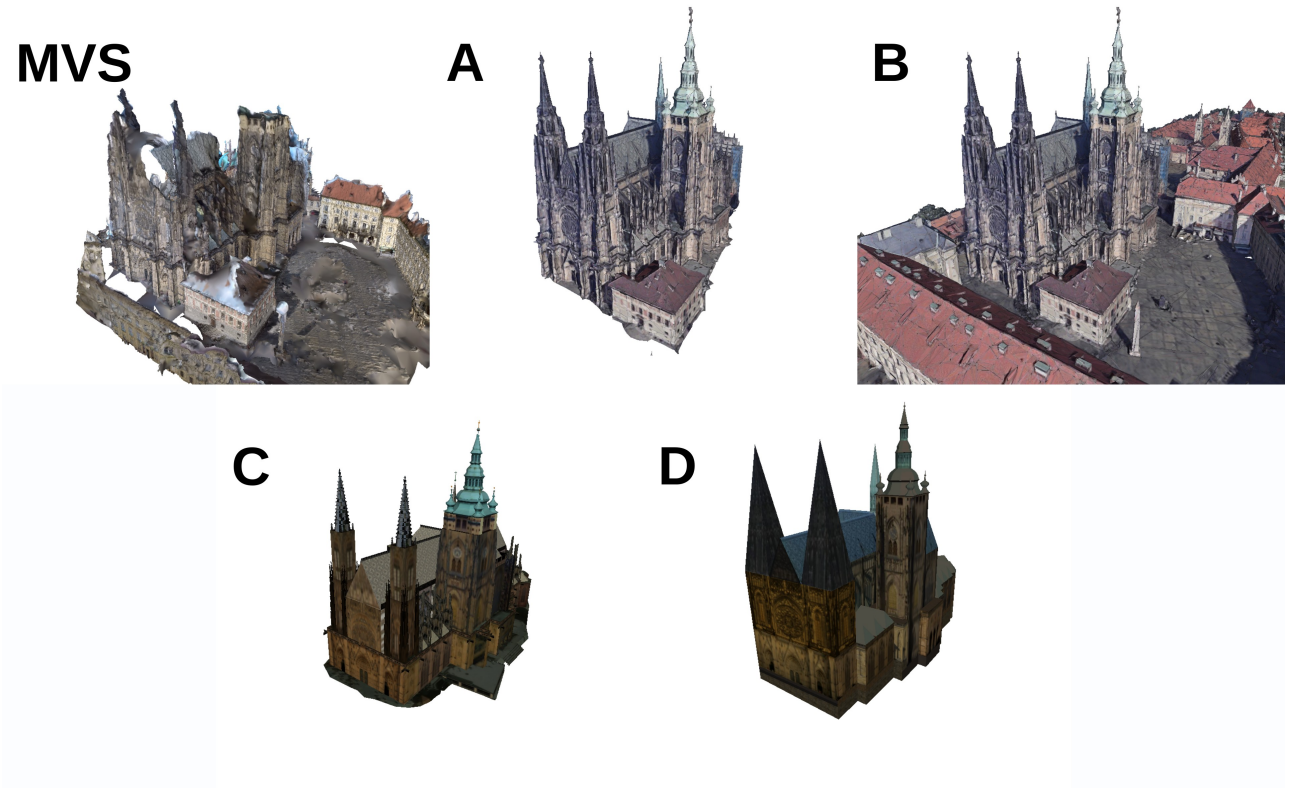


Figure 18. Enlarged St. Vitus Cathedral models from Fig. 2 in the main paper.

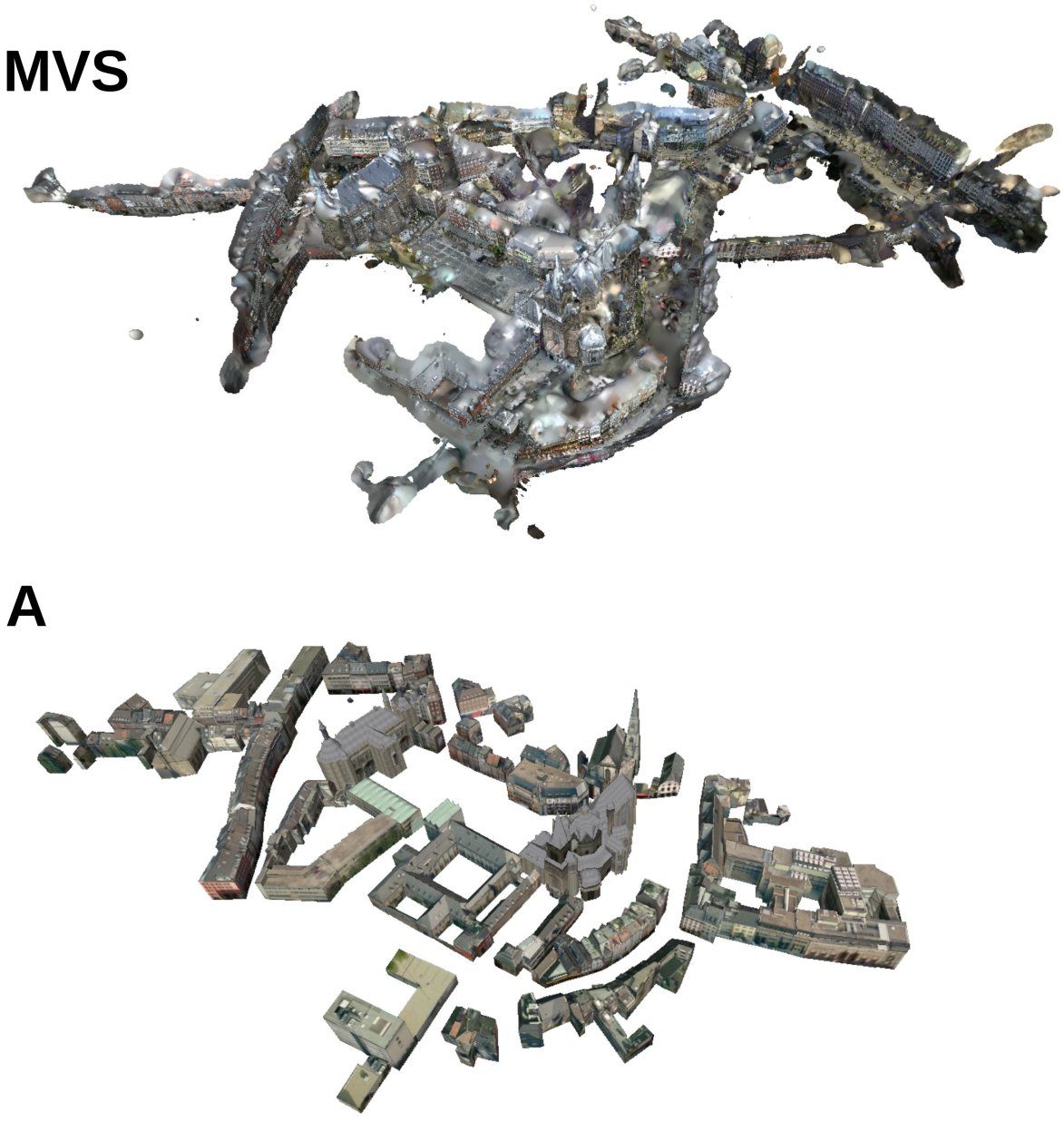


Figure 19. Enlarged Aachen models from Fig. 2 in the main paper.