IMC 2022: Leveraging Detector-Based and Detector-Free Methods

Armando Fortes
Tsinghua University
fmq22@mails.tsinghua.edu.cn

David Pissarra
Tsinghua University
pissarrad10@mails.tsinghua.edu.cn

Abstract

In this report, we describe our approach to the Image Matching task. We propose a solution that leverages both SuperGlue and LoFTR, two highly regarded architectures in this field. Additionally, we apply the novel QuadTree Attention mechanism to LoFTR, which improves the performance of the detector-free model. This report presents every further detail of our solution, as well as insightful ideas taken from it. Also, we present some qualitative match captures of the proposed model. Our solution performed very well in the Image Matching Challenge 2022 hosted by Google Research on Kaggle, placing us in the 34th position among 642 teams in total.

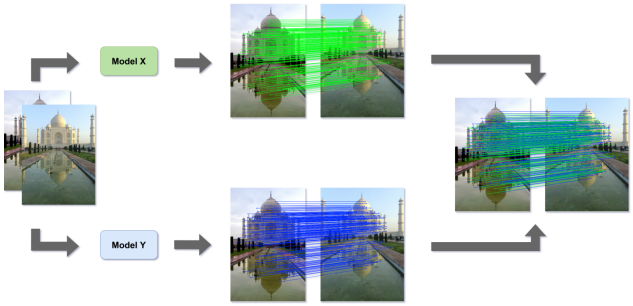

Figure 1: **Approach overview.** This example demonstrates an overview of our approach to this task. We focus on inferring different models, with the aim of extracting diverse interest points from pairs of images, and concatenating them afterward.

1. Introduction

Finding point-to-point correspondences between images is a crucial task for various 3D Computer Vision problems, such as Structure-from-Motion (SfM), Simultaneous Localization and Mapping (SLAM), and Visual Localization. Such correspondences are usually estimated by matching local features, with most existing methods consisting of the following three phases: feature detection, feature description, and feature matching. Given two input images, interest points (e.g. points where the direction of the boundary of an object suddenly changes, and intersection points between edge segments) are first detected. Upon discovering the interest points of both images, the respective local descriptors are extracted from neighborhood regions, which should be invariant under image transformation. Finally, point-to-point correspondences are found by nearest neighbor search or more advanced matching algorithms. A prominent example of a successful application of the classical architecture is composed of learned convolutional neural network based feature detector and descriptor SuperPoint [10], and graph neural network based feature matcher SuperGlue [27].

Nevertheless, the use of a feature detector may sometimes be detrimental, since it can fail to extract sufficient interest points which intersect both input images. Such behavior could be introduced under several factors, for example, poor texture, illumination and viewpoint changes, and motion blur. Accordingly, novel detector-free methods have also been developed to tackle those drawbacks. Sun et al. [29] have recently achieved state-of-the-art results in local feature matching by establishing pixel-wise dense matches at a coarse level and later refining the good matches at a fine level, using self and cross attention layers in Transformer to obtain feature descriptors conditioned on both images.

Google Maps uses SfM techniques for various aspects of its services, such as the creation of 3D models from StreetView and aerial imagery. Intending to accelerate research into this topic and leverage the data already available publicly, Google partnered with the University of British Columbia and the Czech Technical University to host the competition *Image Matching Challenge 2022* [17, 24] on Kaggle, in which we participated. This competition is part of the Image Matching: Local Features and Beyond workshop at the Conference on Computer Vision and Pattern Recognition (CVPR) in June 2022.

In this work, we propose an ensemble-based method as our solution to the aforementioned competition, composed of both traditional detector-based and novel detector-free approaches. We successfully outperform state-of-the-art methods in this competition and managed to win a Kaggle

silver medal, finishing 34th out of 642 teams on the final leaderboard.

2. Related Work

The traditional pipeline for local feature matching was developed in the 2000s and often based on keypoint detector and descriptor SIFT [21] or ORB [26]. Feature matching would usually be treated as a separate phase in the pipeline where descriptors are matched, followed by match filtering with Lowe's ratio test [21], the mutual check, and heuristics such as neighborhood consensus [31, 3, 5]. Additionally, a geometric transformation between the images could be found with a robust solver like RANSAC [14, 23], or one of its improved variants [4, 2, 6].

Recent advances in deep learning have led to the development of better sparse detectors and descriptors from convolutional neural networks (CNNs) [10, 11, 22, 32, 25]. Remarkably, SuperPoint [10] builds upon MagicPoint [9], one of the first learning-based methods for extracting local features, and proposes a self-supervised training approach through homographic adaptation. Most of the mentioned detection and description techniques use nearest neighbor (NN) search to find matches between the discovered interest points. Recently, Sarlin et al. [27] presented SuperGlue, a more sophisticated approach to matching, which simultaneously performs context aggregation, matching, and filtering in a single end-to-end architecture. SuperGlue uses attentional graph neural networks (i.e. a general form of Transformers [18]), achieving impressive results and setting the new state-off-the-art in local feature matching at the time.

From a different standpoint, the primal detection of image interest points might not always be an optimal solution. Feature detectors may fail to extract enough interest points in both images, due to various factors, such as poor texture, illumination and viewpoint changes, and motion blur. Since repeated points are needed to successfully match image features, detector-based models may create insufficient image matches in the mentioned situations. In an attempt to tackle this issue, detector-free methods remove the feature detector phase, directly producing dense feature matches with higher accuracy. Instances following this line of work are DRC-Net [19] and LoFTR [29], which focus on first establishing matches at a coarse level and later refining the good matches at a fine level. LoFTR achieves state-of-the-art results by utilizing a local feature CNN to extract both coarse-level and fine-level feature maps, followed by a linear transformer to process the dense local features, enforcing global consensus among the computed matches.

3. Method

Given two input images, we propose to send a copy of both images to two different local feature matching heads, one is composed of a SuperPoint + SuperGlue approach, and the other is an improved variant of the LoFTR model, where QuadTree Attention is used. Finally, matches from both heads are concatenated and fed into a consensus algorithm for inlier/outlier detection and fundamental matrix calculation. With this configuration, we attempt to leverage conceptually different detector-based and detector-free approaches together and therefore find more sparse and diverse matches throughout images for a more accurate fundamental matrix calculation.

3.1. SuperPoint and SuperGlue

Several local feature detectors and descriptors can be combined with SuperGlue but it works particularly well in conjunction with SuperPoint, which produces sparse and repeatable keypoints, therefore, enabling efficient matching.

SuperPoint. A fully-convolutional neural network that detects interest points accompanied by fixed-length descriptors from a full-sized image in a single forward pass. There is a single encoder that processes all input images and reduces their dimensionality. The architecture divides into two decoder heads after the encoder, one for keypoint detection and another for description.

SuperGlue. Contains two main components: an *attentional graph neural network* and an *optimal matching layer*. In the first component, keypoint positions and their visual descriptors (received from SuperPoint) are encoded into a single vector. This vector is then used to create more powerful representations by alternating self- and cross-attention layers (self-attention boosts the receptive field of local descriptors, while cross-attention enables communication and is inspired by the way humans look back and forth when matching images). The optimal matching layer creates a score matrix, augments it with dustbins, then uses the Sinkhorn algorithm [28, 8] to find the optimal partial assignment.

In summary, SuperPoint presents a fully-CNN architecture for interest point detection and description, trained using a self-supervised domain adaptation framework called Homographic Adaptation, and SuperGlue replaces handcrafted heuristics with a powerful neural model based on attentional-GNNs, which simultaneously performs context aggregation, matching, and filtering in a single unified architecture.

3.2. LoFTR

LoFTR (Detector-Free **Lo**cal **F**eature Matching with **Tr**ansformers), proposed by Sun et al. [29], intends to create quality matches by introducing a detector-free approach. This was achieved by substituting the traditional pipeline with a coarse-to-fine approach. In summary, LoFTR can be regarded as the aggregation of four components:

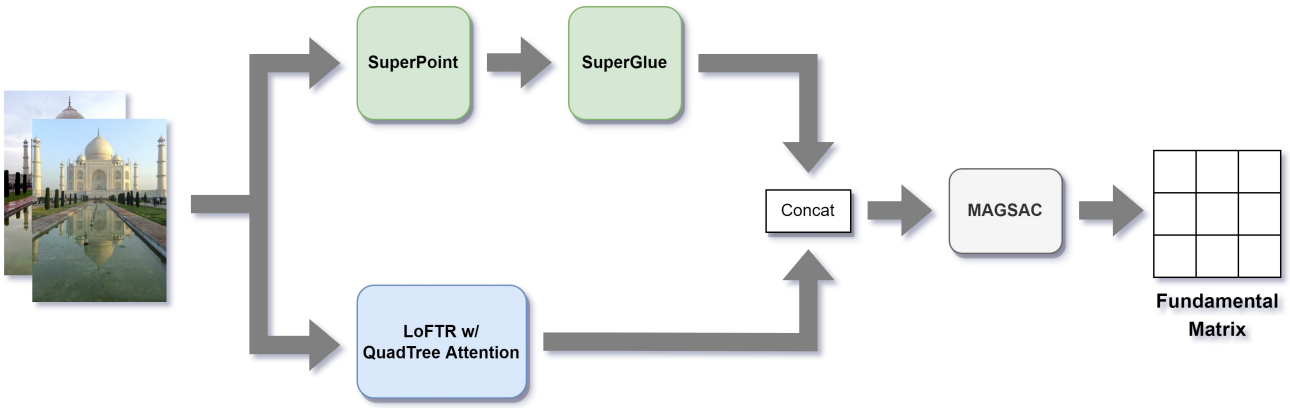

Figure 2: **Our overall model architecture.** Both input images are sent to two different local feature matching heads. The first is the state-off-the-art detector-based approach SuperPoint + SuperGlue, and the other is the state-off-the-art detector-free approach LoFTR, with the novel QuadTree Attention addition. Upon the resulting matches from both heads being found, they are concatenated and fed into the consensus algorithm MAGSAC for inlier/outlier detection and fundamental matrix calculation.

Local Feature CNN. A local feature CNN, using a modified version of ResNet-18 [16] as the backbone. Starts by extracting coarse-level alongside fine-level feature maps.

Coarse-Level Local Feature Transform. Coarse feature maps are now flattened and concatenated with the respective positional encodings to embed position-awareness. These features are then processed by the LoFTR module, which is composed of self- and cross-attention layers.

Matching Module. A matching layer calculates a confidence matrix, matching the transformed features. The resulting coarse matches are filtered according to a given confidence threshold and mutual-nearest-neighbor search.

Coarse-to-Fine Module. Computed coarse matches will be refined within a local window extracted from the fine-level feature maps, resulting in the final match prediction.

3.3. QuadTree Attention

Many hierarchical attention designs are widely adopted by transformers, being successful in many vision tasks. However, some might outperform others according to their specific application and purpose. In fact, most transformers have quadratic computational complexity in terms of the number of image patches. Intending to reduce the computational complexity from quadratic to linear, Tang et al. [30] proposed the QuadTree Attention. The introduced method builds token pyramids and therefore computes attention in a coarse-to-fine manner.

At each level, the QuadTree Attention will divide each image patch into four framed sub-patches. Subsequently, the attention is computed over these created patches, selecting the top-K ones according to the highest attention scores, and discarding the remaining image patches.

As introduced in the last subsection, LoFTR [29] uses the linear attention mechanism to calculate the existing dependencies. However, it has been concluded that it will restrict the attention in local windows in a single attention block, which may harm long-range dependencies. On the other hand, the QuadTree Attention can capture both fine image details and long-range dependencies. Hereupon, as an addition to the recent state-of-the-art work LoFTR, we replace the linear transformer in the LoFTR module with the QuadTree transformer.

4. Experiments

4.1. Evaluation Metric

Throughout our experiments, we evaluate our model by estimating the relative pose between two images, known as projective geometry. In computer vision tasks, this relationship is typically expressed as the fundamental matrix, a 3x3 matrix of rank 2. Thus, the fundamental matrix can be achieved given a set of sparse correspondences of these two perspectives, i.e., the computed matches from the model. Given the computed fundamental matrix and the ground truth matrix, the respective rotation and translation errors can be calculated. Accordingly, one may decide if a given matrix is accurate or not, by defining a threshold. Therefore, matrices are evaluated on the **mean Average Accuracy (mAA)** regarding the error threshold.

4.2. Experimental Details

The competition's evaluation data mostly contains photos of urban scenes. Therefore, the larger dataset

MegaDepth [20] is used to train both SuperGlue and LoFTR models, since it is composed of multi-view internet outdoor photo collections, combined with 3D reconstruction and semantic labeling methods, to generate large amounts of training data for various computer vision tasks. In addition, for simplicity, we decided to use the respective authors' defined hyperparameters, as well as model default configurations, for both SuperGlue and LoFTR.

4.3. Consensus Algorithms

From the produced image matches, the fundamental matrix can be extracted using a consensus algorithm. In fact, for most vision tasks these algorithms are crucial to estimate outlier 2D point correspondences, being capable of iteratively detecting these with at least seven points. From a big variety of consensus algorithms, we found that the most robust and publicly available algorithms were RANSAC [15], MAGSAC [1], and DEGENSAC [7]:

RANSAC. **Ran**dom **Sa**mple **C**onsensus firstly introduced to solve the Location Determination Problem (LDP), by applying the simple least squares method to a sub-sample of points, finding the inlier points.

MAGSAC. **Marg**inalizing **Sa**mple **C**onsensus optimizes previous algorithms by performing the weighted least-squares fitting, improving the accuracy of robust estimation significantly.

DEGENSAC. The algorithm works on top of the two-view geometry theorem, which was proved by the authors. A novel RANSAC-based algorithm is presented, achieving a more robust estimation of projective geometry from image point correspondences.

We provide the results regarding the consensus algorithms in Table 2. All algorithms were iterated 100.000 times. One may observe that MAGSAC outperforms the remaining methods, showing superior accuracy in detecting outlier image point correspondences for the competition.

4.4. Results

We present our final results in Table 1, according to different models. We tested models individually and our ensemble model outperformed any individual model. In addition, we introduce another slight improvement to the models, which consists of re-scaling the image matches back to the original image resolution after inference. This will cause the interest points to be on the same aspect ratio as the original image since LoFTR required resizing the original images for input to further computation. Thus, the consensus algorithm will calculate the fundamental matrix given more accurate correspondences in relation to original aspect ratios. To conclude, our best submission performed well compared to the competition panorama, achieving the 34th position out of 642 teams.

Matcher	Model	mAA		LB
		Public	Private	
SuperGlue	SuperPoint + SuperGlue	67.6	67.1	586th
LoFTR	LoFTR	72.6	73.6	386th
	LoFTR + re-scaling	74.1	74.2	123rd
	LoFTR w/ QTAttention	78.0	79.7	65th
	LoFTR w/ QTAttention + re-scaling	78.4	80.0	61st
SuperGlue + LoFTR	**Ensemble Model**	**81.5**	**82.2**	**34th**

Table 1: **Competition results.** The mean Average Accuracy (mAA) is reported. The models were tested in public and private datasets of the *Image Matching Challenge 2022* competition. In addition, we present the respective projected private leaderboard position (LB) for each model (out of 642 teams).

Consensus Algorithm	mAA		LB
	Public	Private	
RANSAC	72.1	75.8	102nd
MAGSAC	**81.5**	**82.2**	**34th**
DEGENSAC	73.2	76.7	90th

Table 2: **Consensus algorithm comparison.** Similarly to Tab. 1, we present the competition results, according to the respective algorithm.

5. Conclusion

This work presents our method for the Image Local Feature Matching task. Although most of the work is focused on the Image Matching models, i.e. LoFTR and SuperGlue, many other concepts are important to understand the projective geometry behind this problem. Also, consensus algorithms were studied to better compute the fundamental matrix out of the image correspondences. This way, we were able to successfully assess the models.

Despite our final solution achieving satisfactory results in the leaderboard, we believe that better results could be accomplished by ensembling more existent models, such as DKM [12] for matching or even DBSCAN [13] to generate precise bounding boxes for further post-computation. In fact, this year's edition of the competition did not yield a new singular state-off-the-art winning model, instead, it was observed that the top-10 contestants of the competition also created ensemble-based approaches, similar to ours. We believe that ensemble models work especially well in this task because the optimal solution would detect the most sparse and diverse accurate matches possible throughout a given pair of images, i.e. by leveraging different detector-based and detector-free feature matching algorithms, we are able to detect some distinct matches from each algorithm and further combine them together.

References

[1] Daniel Barath and Jiri Matas. MAGSAC: marginalizing sample consensus. *CoRR*, abs/1803.07469, 2018. 4

[2] Dániel Baráth and Jiri Matas. Magsac: Marginalizing sample consensus. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10189–10197, 2019. 2

[3] Jiawang Bian, Wen-Yan Lin, Yasuyuki Matsushita, Sai-Kit Yeung, Tan Dat Nguyen, and Ming-Ming Cheng. Gms: Grid-based motion statistics for fast, ultra-robust feature correspondence. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2828–2837, 2017. 2

[4] Eric Brachmann and Carsten Rother. Neural-guided ransac: Learning where to sample model hypotheses. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4321–4330, 2019. 2

[5] Jan Cech, Jiri Matas, and Michal Perdoch. Efficient sequential correspondence selection by cosegmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32:1568–1581, 2008. 2

[6] Ondřej Chum, Tomá*vs*. Werner, and Jiri Matas. Two-view geometry estimation unaffected by a dominant plane. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 1:772–779 vol. 1, 2005. 2

[7] Ondrej Chum, Tomas Werner, and Jiri Matas. Two-view geometry estimation unaffected by a dominant plane. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 772–779. IEEE, 2005. 4

[8] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *NIPS*, 2013. 2

[9] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Toward geometric deep slam. *ArXiv*, abs/1707.07410, 2017. 2

[10] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 337–33712, 2018. 1, 2

[11] Mihai Dusmanu, Ignacio Rocco, Tomás Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-net: A trainable cnn for joint detection and description of local features. *ArXiv*, abs/1905.03561, 2019. 2

[12] Johan Edstedt, Mårten Wadenbäck, and Michael Felsberg. Deep kernelized dense geometric matching. *CoRR*, abs/2202.00667, 2022. 4

[13] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231, 1996. 4

[14] Martin A. Fischler and Robert C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24:381–395, 1981. 2

[15] Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, jun 1981. 4

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. 3

[17] 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. *International Journal of Computer Vision*, 2020. 1

[18] Chaitanya Joshi. Transformers are graph neural networks. *The Gradient*, 2020. 2

[19] Xinghui Li, Kai Han, Shuda Li, and Victor Adrian Prisacariu. Dual-resolution correspondence networks. *CoRR*, abs/2006.08844, 2020. 2

[20] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. *CoRR*, abs/1804.00607, 2018. 4

[21] G LoweDavid. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2004. 2

[22] Yuki Ono, Eduard Trulls, Pascal V. Fua, and Kwang Moo Yi. Lf-net: Learning local features from images. In *NeurIPS*, 2018. 2

[23] Rahul Raguram, Jan-Michael Frahm, and Marc Pollefeys. A comparative analysis of ransac techniques leading to adaptive real-time random sample consensus. In *ECCV*, 2008. 2

[24] Google Research. Image matching challenge 2022. https://www.kaggle.com/competitions/image-matching-challenge-2022/. 1

[25] Jérôme Revaud, Philippe Weinzaepfel, César Roberto de Souza, No'e Pion, Gabriela Csurka, Yohann Cabon, and M. Humenberger. R2d2: Repeatable and reliable detector and descriptor. *ArXiv*, abs/1906.06195, 2019. 2

[26] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary R. Bradski. Orb: An efficient alternative to sift or surf. *2011 International Conference on Computer Vision*, pages 2564–2571, 2011. 2

[27] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4937–4946, 2020. 1, 2

[28] Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21:343–348, 1967. 2

[29] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. *CVPR*, 2021. 1, 2, 3

[30] Shitao Tang, Jiahui Zhang, Siyu Zhu, and Ping Tan. Quadtree attention for vision transformers. *CoRR*, abs/2201.02767, 2022. 3

[31] Tinne Tuytelaars and Luc Van Gool. Wide baseline stereo matching based on local, affinely invariant regions. In *BMVC*, 2000. 2

[32] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal V. Fua. Lift: Learned invariant feature transform. *ArXiv*, abs/1603.09114, 2016. 2

Appendix

A. Inference Examples

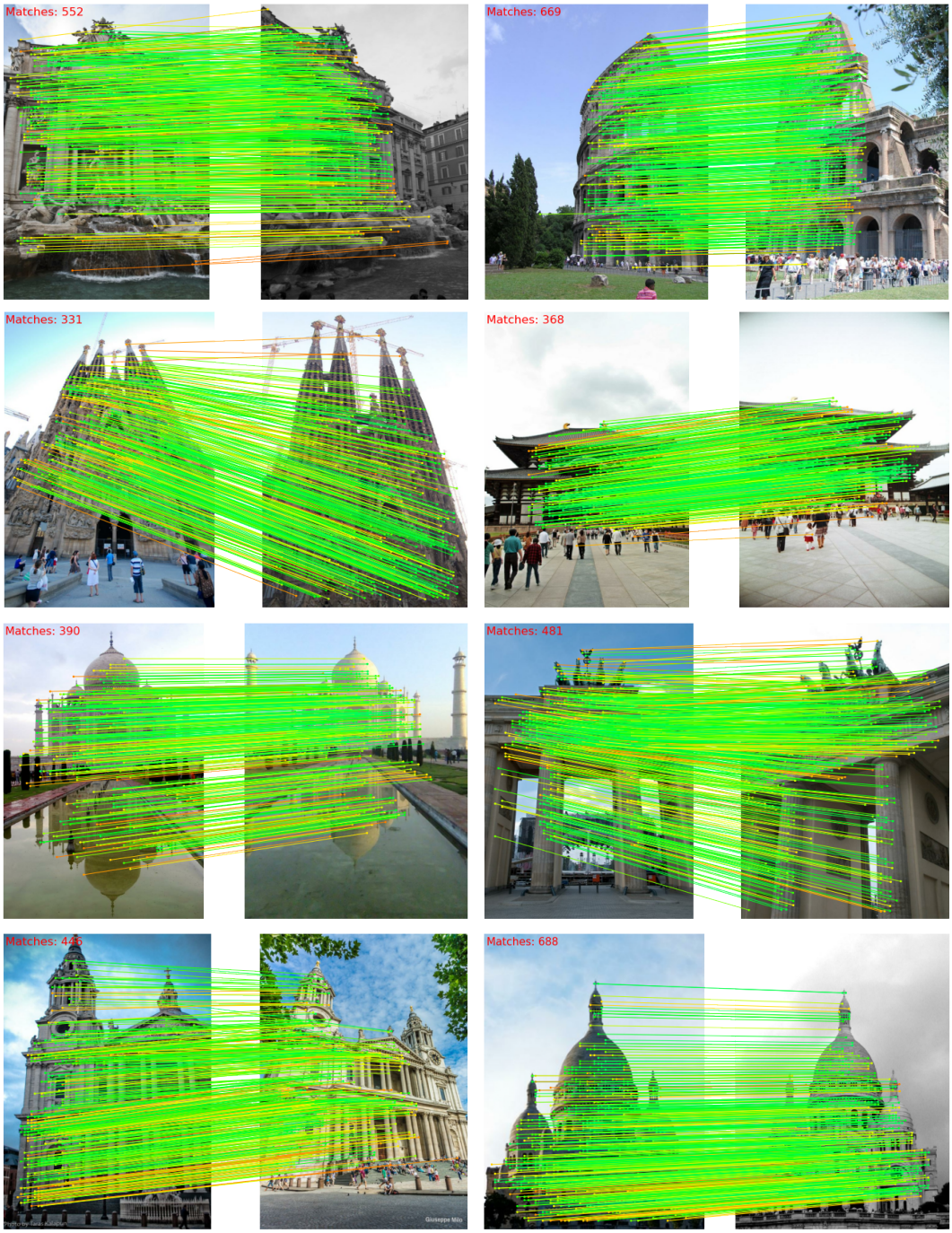

Figure 3: **Final solution inference examples.** We show some inference examples of our model, using photos from various well-known monuments around the world. Highly-confident matches are represented by green lines and low-confident matches by red lines. Therefore, matches with intermediate confidence are yellow lines. Additionally, the number of matches is presented at the top-left corner of each photo.