2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

Wide-Area Geolocalization with a Limited Field of View Camera

Lena M. Downes[1,2,3], Ted J. Steiner[2], Rebecca L. Russell[2] and Jonathan P. How[1]

***Abstract*— Cross-view geolocalization, a supplement or replacement for GPS, localizes an agent within a search area by matching images taken from a ground-view camera to overhead images taken from satellites or aircraft. Although the viewpoint disparity between ground and overhead images makes cross-view geolocalization challenging, significant progress has been made assuming that the ground agent has access to a panoramic camera. For example, our prior work (WAG) introduced changes in search area discretization, training loss, and particle filter weighting that enabled city-scale panoramic cross-view geolocalization. However, panoramic cameras are not widely used in existing robotic platforms due to their complexity and cost. Non-panoramic cross-view geolocalization is more applicable for robotics, but is also more challenging. This paper presents Restricted FOV Wide-Area Geolocalization (ReWAG), a cross-view geolocalization approach that generalizes WAG for use with standard, non-panoramic ground cameras by creating pose-aware embeddings and providing a strategy to incorporate particle pose into the Siamese network. ReWAG is a neural network and particle filter system that is able to globally localize a mobile agent in a GPS-denied environment with only odometry and a 90° FOV camera, achieving similar localization accuracy as what WAG achieved with a panoramic camera and improving localization accuracy by a factor of 100 compared to a baseline vision transformer (ViT) approach.**

I. INTRODUCTION

GPS is an external system for localization that is susceptible to failure through jamming, spoofing, and signal dropout due to dense foliage or urban canyons. Cross-view geolocalization [1]–[5] is a localization method that only requires images from a ground-view camera and preexisting overhead imagery, with or without GPS measurements. Cross-view geolocalization measures the similarity between a ground image and all of the satellite images in a search area to determine the location that the ground image was taken from. Satellite imagery at some resolution is widely available for most of the planet, even in forests and urban areas where GPS signals can be weaker.

Cross-view geolocalization with ground and overhead images is challenging due to the wide difference in viewpoints. Many existing works [5]–[8] rely upon the use of a panoramic ground camera because it effectively decreases the problem dimensionality by reducing the impact of heading. Although a panoramic ground camera can have different headings, the heading only affects the alignment of the image, not the content, whereas the heading of a limited field of view (FOV) camera affects the visible content of the image. The use of panoramic ground cameras also simplifies the problem by maintaining as much semantic similarity as possible between the two viewpoints—overhead images show 360° of the surroundings of a ground agent, as do panoramic ground cameras. However, in practice panoramic cameras are rarely used due to their high monetary cost (resulting in lesser availability) and their difficulty to mount without occlusion. As a result, few real-world systems can benefit from panoramic-based localization. Widespread adoption of cross-view geolocalization technology will require its applicability to platforms without panoramic imaging capabilities, like that shown in Fig. 1.

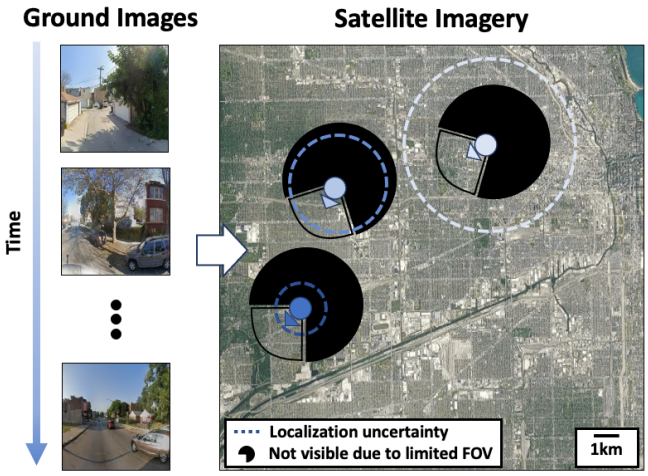


Fig. 1. ReWAG is a cross-view geolocalization system that takes in a series of non-panoramic ground-view camera images and satellite imagery of the search area to accurately localize the agent on a search area scale that was not possible with previous works.

Most recent work on cross-view geolocalization takes a deep learning approach to the problem by using Siamese networks [3]–[7], [9]. A Siamese network consists of a pair of neural networks with matching architectures that simultaneously learn embedding schemes for ground and overhead images. The Siamese network is trained to embed images so that images that were taken in the same location are close together in embedding space while images that were taken in different locations are far apart in embedding space. Although Siamese networks have improved geolocalization performance beyond what was demonstrated with hand-crafted features, their accuracy can be improved further by integrating their measurements over time with a particle filter [4]–[6], [10], [11]. However, existing particle filter geolocalization works are highly constrained— requiring some level of GPS data [6], [10], 180° or greater FOV of the ground camera [4]–[6], [10], [11], perfect initial location knowledge [5], or a search area of less than 2.5 km^2 [4], [5]. These works all impose additional constraints because they

[1]Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, Cambridge, MA, USA
[2] Perception and Embedded ML Group, Draper, Cambridge, MA, USA
[3] Draper Scholar. Research funded by Draper. lmdownes@mit.edu

are not able to efficiently geolocalize a limited FOV camera across a large search area. Localization with a limited FOV camera increases both the difficulty of the problem and the computational requirements.

Our contribution. Our approach, Restricted FOV Wide-Area Geolocalization (ReWAG), builds upon our previous work in [11] to enable efficient wide-area geolocalization with a restricted FOV camera through two key changes. The first change is a computationally efficient method for matching limited FOV ground images to satellite images by appending relative pose information to the intermediate network embedding before inputting it to the Spatial Aware Feature Aggregator (SAFA) [2]. The second change is the dual incorporation of relative pose into both the Siamese network and the particle filter, which enables the probability distribution to be modeled more accurately. The only additional information these changes necessitate at runtime is a noisy heading from a compass, and they produce a cross-view geolocalization system that is capable of localizing across city-scale search areas using a 90° FOV camera. In summary, in this paper we demonstrate the following contributions of ReWAG for restricted FOV localization:

1) Efficient pose-aware embedding generation,
2) A particle filter system that more accurately models the probability distribution, resulting in lower average and final localization error, and
3) Faster particle filter convergence than a ViT baseline [12].

II. RELATED WORKS

Ground-to-aerial cross-view geolocalization. Cross-view geolocalization derives from previous work in the areas of scene recognition and image retrieval. Ground-to-aerial cross-view geolocalization pushes previous work to higher levels of difficulty due to the vast difference in viewpoints between ground and aerial images. Previous works have attempted to solve this problem with hand-crafted features and traditional computer vision techniques [8], [13]–[15], but recent works [2]–[5], [7], [9], [16]–[21] have mostly applied deep learning in the form of Siamese networks [22]. In recent years, recall at top-1 has been steadily rising, but most works [5]–[8] focus on panoramic ground images and report high recall at top-1 for these panoramic ground images. Recall at top-1 for limited FOV ground images is significantly lower than that for panoramic ground images. However, limited FOV cameras are much more common than panoramic cameras for robotic applications.

Orientation-aware cross-view geolocalization. Non-panoramic ground cameras make cross-view geolocalization more difficult due to the reduced number of visible features in the image and due to the matching satellite features being concentrated within one area of the satellite image instead of spread throughout it. The unknown orientation of the ground camera is a key factor in this problem. Some previous works have developed methods to incorporate an understanding of orientation into the cross-view geolocalization system, like by appending orientation maps to images to be input to the Siamese network [7], by using Dynamic Similarity Matching (DSM) to calculate the correlation between ground and satellite images [23], or by jointly embedding the full satellite image as well as the satellite image portion that is visible in the limited FOV ground image [24]. Instead of jointly determining the most highly matching satellite image and the orientation, [25] assumes that the satellite image has already been determined, and they then use pose optimization to estimate the pose within that satellite image. *These works, not designed for mobile robotics constraints, require many search iterations, polar transformations and data augmentations at runtime and hence may be too computationally demanding for real-time robotics.*

Orientation-blind cross-view geolocalization. More recent works tend to treat orientation as an aspect of the problem that can be solved at the last step [9], or do not directly encode or estimate it at all [2], [12]. In [9] orientation-invariant embedding schemes are learned through a combination of global mining, binomial loss, and training data augmentation with random rotations. Spatial Aware Feature Aggregation (SAFA) [2] is an attention mechanism that helps the network to learn image descriptors regardless of the large viewpoint difference between ground and satellite views. TransGeo [12] uses a vision transformer instead of the typical convolutional neural network (CNN) approach. This attention-focused approach embeds patches of the images into tokens with learnable position tokens, which gives it a more flexible method for learning about orientation and position while embedding images. *Although orientation-blind embedding schemes improve image retrieval when orientation is unknown, these methods do not have a mechanism by which the ground camera pose can be input when modeling with a particle filter.*

Particle filter implementations. Some previous works have combined image retrieval with particle filters to enable cross-view geolocalization over time as an agent moves through a search area [4], [5], [8], [10], [11], [13]. Particle filters in these systems use random discrete particles to model a probability distribution of the agent location given odometry, satellite images of the search area, and ground images. However, existing works have localized with particle filters that use wide angle ground cameras [5], [8], [11], [13]. Although [4], [10] localize a ground agent with a 180° ground camera instead of a panoramic 360° camera, commonly available cameras like those found in cell phones have FOV of 90° or less, making much fewer features visible in their images. *These existing particle filter cross-view geolocalization systems are not able to accurately localize with extremely limited FOV images.*

III. METHODS

A. Overview of Approach

ReWAG builds upon WAG, the system developed by [11], to enable localization across a wide search area with restricted FOV ground images. ReWAG uses WAG's strategies of creating a coarse satellite image database, generating embeddings with Siamese networks based on VGG-16 [26] and SAFA [2], and localizing over time with a particle filter

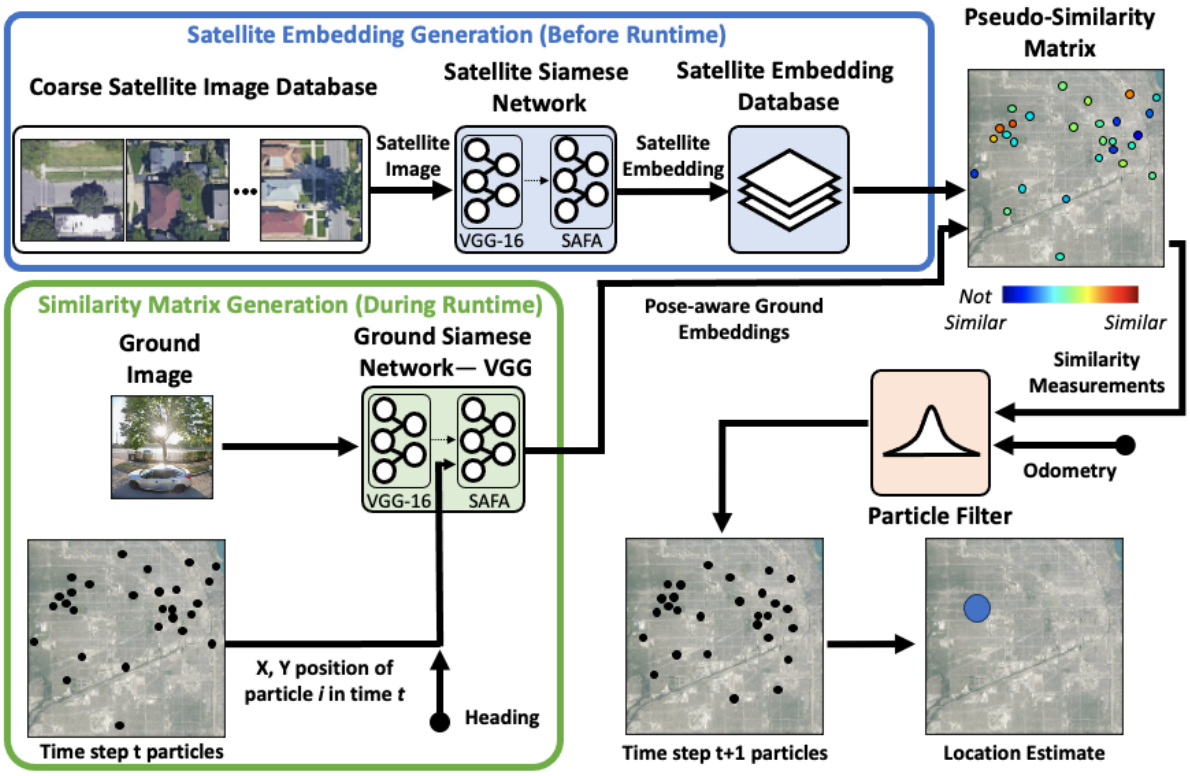


Fig. 2. Diagram of ReWAG. Satellite embeddings are generated before runtime with the coarsely sampled satellite image database and the Siamese network that was trained with trinomial loss. During runtime, a pose-aware ground embedding is generated for each particle at each time step and combined with the satellite embeddings to create a pseudo-similarity matrix, which is the similarity of each particle with its location in the search area. Odometry and measurements from the pseudo-similarity matrix are input to the particle filter with a Gaussian measurement model to generate a location estimate.

(see Fig. 2). However, ReWAG differs from WAG in two major ways: first, ReWAG generates pose-aware image embeddings in a computationally efficient manner, and second, ReWAG generates more informative similarity measures and hence more accurately models the agent location probability distribution by incorporating pose information from each particle into the Siamese network input. For non-panoramic ground imagery it is necessary to incorporate this additional information due to the increased difficulty of the problem.

ReWAG first coarsely samples the search area to construct a database of satellite images that are preprocessed with a Siamese network before runtime to generate satellite embeddings. During runtime, a generic ground embedding is generated by the VGG-16 portion of the Siamese network for each ground image, and for each particle a pose-aware embedding is produced from the generic embedding and the particle's pose. The similarity between each particle's pose-aware ground embedding and its corresponding satellite embedding is calculated to produce the pseudo-similarity matrix, a probabilistic representation of the ground image's similarity across the search area. The particle filter receives odometry and measurements from the pseudo-similarity matrix to produce a location estimate at each time step.

B. Pose-Aware Embeddings

We have developed a method to train the ground Siamese network to generate pose-aware embeddings in a computationally efficient manner while minimally modifying the architecture. Like WAG, our Siamese network architecture is derived from that of [6], which consists of a VGG-16 backbone and a SAFA module to increase the network's spatial understanding. In ReWAG, the ground Siamese network is modified to append the particle pose to the intermediate embedding that is output by the VGG-16 backbone, as shown in Fig. 3. This intermediate embedding appended with the particle pose is then input to SAFA, which learns a spatial-aware representation of the ground image.

ReWAG's computationally efficient benefit comes from the ability to generate one base embedding for each ground image, and then append any pose to the base embedding to efficiently generate a pose-aware embedding for each particle with the much lighter-weight SAFA. When combined with a particle filter, this design enables the VGG-16 inference to be done once per time step instead of once for each particle for each time step. In contrast, [24] generates pose-aware embeddings through a joint global and local pipeline, hence the full embedding generation must be done for each particle at each timestep.

C. Siamese Network and Particle Filter Integration

Our key observation is that cross-view geolocalization can be improved by more thorough integration between the Siamese network and the particle filter. Previous works have built systems with mostly one-way connections between the Siamese network and the particle filter—for each time step, the location of each particle determines which satellite image

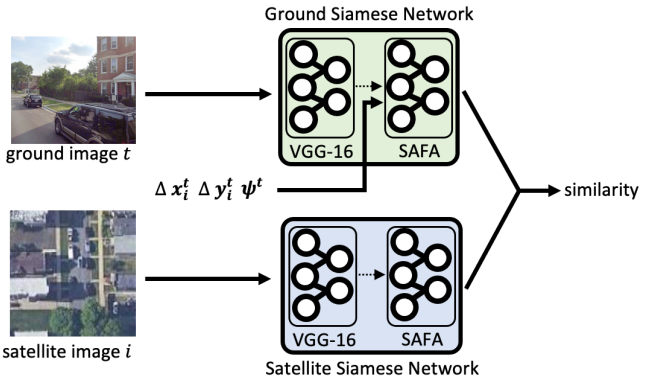


Fig. 3. Our method requires pose to be incorporated in only the SAFA portion of the ground Siamese network, which reduces computation and allows faster inference at run time.

Fig. 4. Particles are dispersed through search area, which is segmented into satellite tiles whose embeddings are precomputed before runtime. At each time step, the true heading of the ground agent is approximately known and the x, y location of each particle within its satellite tile is given by its displacement from the center of the tile.

will be compared with that time step's ground image, and that similarity is used to adjust the weight of that particle. However, additional information can be integrated into the Siamese network-particle filter connection. In addition to a corresponding satellite image, each particle also has a location within that satellite image, as shown in Fig. 4. the particle filter is modeling a probability distribution that includes the location within the satellite image, but traditional architectures do not factor that information into the similarity measure and hence it is not reflected in the particle weights.

We have developed a method by which we incorporate particle pose information into the similarity measure through the pose-aware embeddings. The pose of a particle i at time t consists of x and y displacements Δx_i^t and Δy_i^t, which are determined from the particle's location within its satellite tile, and the heading ψ^t, which is determined from sensor measurements at time step t. We assume that the heading measurement will be fairly accurate to the true ground agent heading based on the typical error of a compass. At each time step, the ground image is used to generate one generic intermediate embedding, and for each particle at that time step a pose-aware embedding is generated from the generic embedding and each particle pose. This method increases the computation required for each particle filter update, but the computationally efficient method by which we generate pose-aware embeddings helps to offset this increase. The incorporation of the particle pose aids the Siamese network in identifying where within the satellite image there should be corresponding features if the image pair is a positive match. Without pose-aware embeddings, a negative image pair could incorrectly find matching features anywhere within the satellite image. With pose-aware embeddings, a negative image pair will be encouraged to only look for matching features within a specified portion of the satellite image, decreasing the opportunity to find false matches.

IV. RESULTS

A. Experimental Setup

To demonstrate ReWAG's performance relative to WAG [11], we perform limited FOV ground image localization experiments with the same test paths from [11] but, instead of using panoramic ground images, we crop the ground images to 90° FOV. These test paths are a large scale localization experiment with very noisy location initialization across the entire city of Chicago using simulated data. The simulated data consists of ground images and overhead satellite images from Google Maps Static API and odometry measurements from the ground-truth displacement between images with added noise proportional to displacement. A satellite image database was generated by sampling the search area approximately every 60 meters into a 256 × 256 grid of non-overlapping satellite image tiles. This grid size maintains a similar image size and resolution that the network was trained on; satellite images of size 640 × 640 pixels with a resolution of approximately 0.1 m/pixel.

We use a version of the neural network architecture from [6] with the VGG-16 backbone and the SAFA module, modified to generate pose-aware embeddings. The satellite network architecture is unchanged from [11]. We train both ReWAG's Siamese network and our comparison baseline, a stage-1 TransGeo [12], on the VIGOR dataset [6] with the ground images cropped to 90°. We train the TransGeo baseline for 50 epochs with the training parameters described in [12], and ReWAG for 30 epochs with triplet loss [5]:

$$\mathcal{L}_{\text{triplet}} = \log\left(1 + e^{-\alpha(d_{pos} - d_{neg})}\right) \tag{1}$$

with α loss parameter set to 10. Then we train ReWAG for 15 additional epochs with trinomial loss [11]:

$$\mathcal{L}_{\text{t}} = \frac{\log\left(1 + e^{-\alpha_{\text{p}}(S_{\text{p}} - m_{\text{p}})}\right)}{N_{\text{p}}\alpha_{\text{p}}} + \frac{\log\left(1 + e^{\alpha_{\text{n}}(S_{\text{n}} - m_{\text{n}})}\right)}{N_{\text{n}}\alpha_{\text{n}}} + \frac{\log\left(1 + e^{-\alpha_{\text{semi}}(S_{\text{semi}} - m_{\text{semi}})}\right)}{N_{\text{semi}}\alpha_{\text{semi}}} \tag{2}$$

with the parameter values used in [11]. The filter has 30,000 particles and uses the Gaussian measurement model from [11]. We initialized the particle filter with Gaussian distributions, centered 1.3 km from the true initial location (standard deviation of 900 m) for C-1 and C-2, and centered 600 m from the true initial location (standard deviation of 300 m) for C-3. We add 2% noise to the ground-truth odometry and 1% noise to the ground-truth heading at each time step.

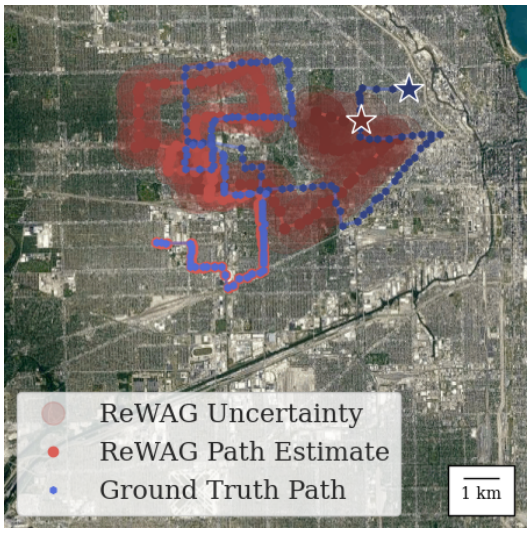


Fig. 5. The ground-truth path in Chicago and ReWAG's path estimate, which accurately converges upon the ground-truth. Uncertainty bubble sizes are scaled down by raising to 0.75 to improve interpretability. Increasing brightness of path indicates passing of time. Start marked with stars.

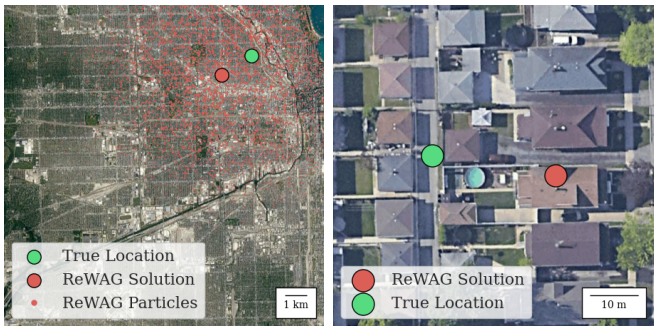
B. Large-scale Test: Chicago

Experiment details. The true path the simulated agent travels in Chicago is shown in Fig. 5 with ReWAG's estimated location at each time step. We ran this experiment with ReWAG and a baseline that uses stage-1 TransGeo [12], a Vi-T approach, for its Siamese network combined with the same particle filter as ReWAG.

Estimation error. Even with an initialization as far from the agent as Fig. 6(a), ReWAG has a final estimation error of 26 m (visible in Fig. 6(b)). This estimation error is only 5 meters greater than that which was achieved with 360° ground images in WAG. Fig. 7 compares the estimation error of ReWAG's particle filter and the TransGeo baseline as the simulated agent moves. The error is the Euclidean distance between the actual location and the weighted average of the particle locations. Over the duration of the experiment, ReWAG has an average estimation error of 925 m, versus 2.2 km for the baseline. ReWAG achieves a final estimation error of 26 m compared to the baseline of 2.2 km.

Convergence. Figure 8 compares the system convergence measured as the mean squared error (MSE) of the particle locations at each time step. ReWAG converges to an MSE of less than 60 m (the satellite image size) after 133 filter updates, while the baseline does not reach that level of convergence in the time period tested. Fig. 9 shows the particle filter distribution that ReWAG converges to as red dots in the inset of the figure while the baseline terminates with the particle distribution shown in blue dots, which still shows significant estimation error.

Ablation. We performed a small ablation study to determine the benefit of including both heading and location information in the pose-aware embeddings as opposed to only including heading, or including neither as in WAG. We trained a Siamese network with the same architecture, training regime and parameters as ReWAG, with the exception that the base embeddings input to SAFA only had heading



(a) Initial distribution supplied to particle filter. True location is over 1 km from initial particle filter estimate.

(b) Final particle filter solution and true location. True location is approximately 26 m from final particle filter estimate.

Fig. 6. ReWAG is able to accurately localize an agent to within 26 meters of its true location across nearly 200 km^2 of Chicago after being initialized to a Gaussian distribution centered 1.3 km from the true location.

TABLE I
ABLATION

Metric and System Type		C-1
Final Error (m)	ReWAG	26
	ReWAG without Position	375
	WAG	192
Convergence Time (time steps)	ReWAG	133
	ReWAG without Position	-
	WAG	-

appended to them. We also tested limited FOV images with WAG retrained on limited FOV (without heading or position appended). We tested the systems on the C-1 test path and the results are summarized in Table I We attribute ReWAG without position's performance to the fact that heading alone does not dictate what content is visible in a ground image and hence may be misleading on its own.

Multiple path result summary. Table II shows a summary of ReWAG's localization performance compared to the baseline on three test paths in Chicago. These paths are the same test paths that were used in [11]. The first, C-1, is the path discussed previously (Fig. 5). The particle filter for C-1 and C-2 was initialized 1.2 km away from the ground truth, and for C-3 was initialized 600 m away from the ground truth due to the challenging urban scenery in this path. On all paths, ReWAG outperforms the baseline in final estimation error, final standard deviation and convergence time.

C. Small-scale Test: KITTI

We also demonstrate ReWAG's localization performance on a KITTI test path in Fig. 10. This experiment demonstrates ReWAG's ability to localize with more accurate initialization information across a smaller search area and its robustness towards localizing on images in a different city than those that it was trained on. We sample images and pose data from the residential "2011_0_30_drive_0028" path and reduce the FOV to 90° by cropping the KITTI images. The satellite images of the search area are obtained from the Google Static API; the search area is divided into a grid of 32 $\times$ 32 satellite images at zoom level 20. We initialize the particle filter with a Gaussian distribution approximately 80

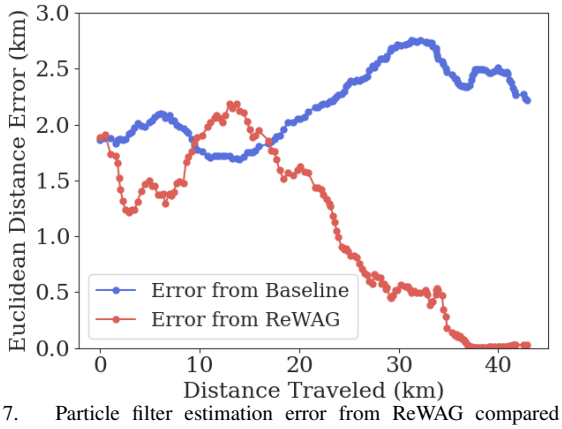


Fig. 7. Particle filter estimation error from ReWAG compared to a TransGeo baseline. ReWAG has lower final and average error.

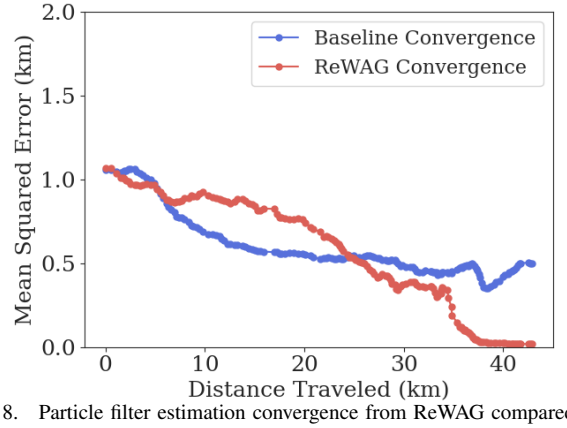


Fig. 8. Particle filter estimation convergence from ReWAG compared to a TransGeo baseline. ReWAG accurately converges, the baseline does not.

TABLE II
COMPARISON OF RESULTS ON CHICAGO TEST PATHS– BASELINE: WITH TRANSGEO

Metric and System Type		C-1	C-2	C-3
Final Error (m)	Baseline	2218	2259	300
	ReWAG	26	16	17
Final Standard Deviation (m)	Baseline	500	1321	169
	ReWAG	18	10	10
Convergence Time (time steps)	Baseline	-	-	-
	ReWAG	133	61	41

m away from the true location, and ReWAG's final estimation error is 12 m after 34 ground image updates.

V. CONCLUSION

ReWAG redesigns the Siamese network-particle filter architecture for increased hardware flexibility, a key component of applying cross-view geolocalization to mobile robotics. Previous works have largely focused on geolocalization with ground cameras of unrealistic FOVs for existing robotics platforms. This work uses all available information to generate Siamese network embeddings and accurately reweight particles, hence enabling faster and more accurate particle filter convergence with limited FOV cameras.

ReWAG accurately localizes across several hundred square kilometers of Chicago, maintaining the same level of accuracy that WAG [11] previously demonstrated with panoramic ground cameras. ReWAG maintains the same benefits of WAG in terms of reducing the size of the satellite image database required. Additionally, it has lower final error and

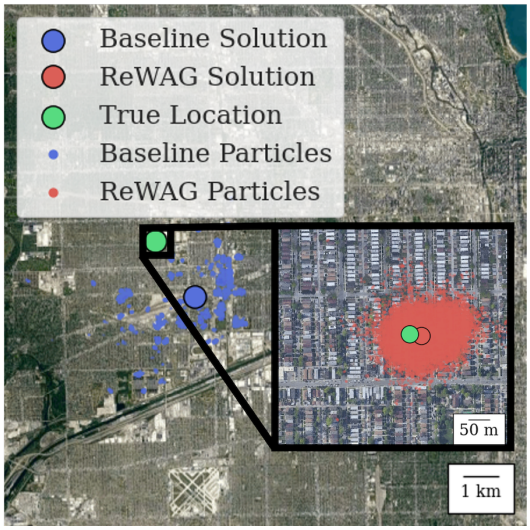


Fig. 9. Final particle filter dispersion with the baseline system and with ReWAG on the Chicago test path (C-1). Baseline does not successfully converge to a location estimate, while ReWAG converges to within 18 m of standard deviation.

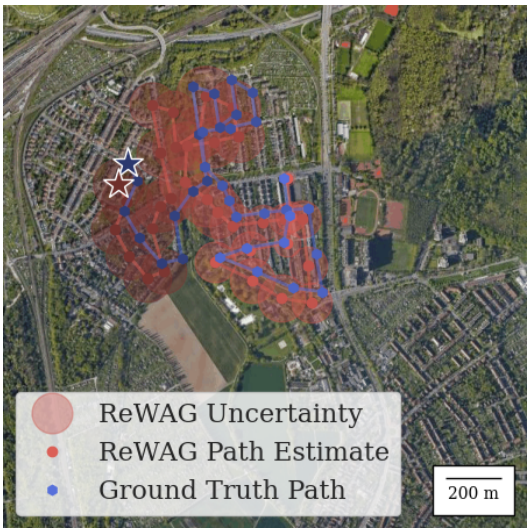


Fig. 10. Ground-truth and estimated path in the KITTI test area. Increasing brightness of path indicates passing of time. Start marked with stars.

faster convergence than the TransGeo baseline in the Chicago test area, which we attribute to its pose-awareness. It also demonstrates its ability to generalize to a city it was not trained on by successfully converging with only 12 m error in the KITTI test path.

In the short term, future work on this topic includes improving computational speed, testing on real-time physical platforms, and further restriction of the FOV. Domain shift also remains an open challenge in the field— localizing in areas with different appearances than training image pairs, on satellite images that were taken in different seasons than the ground images, and in rural areas without as many identifiable landmarks as residential or urban areas.

In summary, ReWAG lifts the heavy hardware requirements that were inherent in other cross-view geolocalization systems and enables accurate localization using a camera with a narrow FOV. It is a step forward in making cross-view geolocalization a widespread, cross-platform tool.

REFERENCES

[1] Y. Tian, C. Chen, and M. Shah, "Cross-view image matching for geo-localization in urban environments," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1998–2006.

[2] Y. Shi, L. Liu, X. Yu, and H. Li, "Spatial-aware feature aggregation for image based cross-view geo-localization," in *NeurIPS*, 2019.

[3] S. Cai, Y. Guo, S. Khan, J. Hu, and G. Wen, "Ground-to-aerial image geo-localization with a hard exemplar reweighting triplet loss," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 8390–8399.

[4] D.-K. Kim and M. R. Walter, "Satellite image-based localization via learned embeddings," *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2073–2080, 2017.

[5] S. Hu and G. H. Lee, "Image-based geo-localization using satellite imagery," *International J. of Computer Vision*, pp. 1205–1219, 2020. [Online]. Available: https://doi.org/10.1007/s11263-019-01186-0

[6] S. Zhu, T. Yang, and C. Chen, "Vigor: Cross-view image geo-localization beyond one-to-one retrieval," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5316–5325, 2021.

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

[8] A. Viswanathan, B. R. Pires, and D. F. Huber, "Vision based robot localization by ground to satellite matching in gps-denied situations," *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 192–198, 2014.

[9] S. Zhu, T. Yang, and C. Chen, "Revisiting street-to-aerial view image geo-localization and orientation estimation," in *2021 Winter Conf. on Applications of Computer Vision*, 01 2021, pp. 756–765.

[10] Z. Xia, O. Booij, M. Manfredi, and J. F. P. Kooij, "Cross-view matching for vehicle localization by learning geographically local representations," *IEEE Robotics and Automation Letters*, vol. 6, pp. 5921–5928, 2021.

[11] L. M. Downes, D.-K. Kim, T. J. Steiner, and J. P. How, "City-wide street-to-satellite image geolocalization of a mobile ground agent," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022, forthcoming in IROS 2022. [Online]. Available: https://arxiv.org/abs/2203.05612

[12] S. Zhu, M. Shah, and C. Chen, "Transgeo: Transformer is all you need for cross-view image geo-localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1162–1171.

[13] A. Viswanathan, B. R. Pires, and D. Huber, "Vision-based robot localization across seasons and in remote locations," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*, 2016, pp. 4815–4821.

[14] N. Jacobs, S. Satkin, N. Roman, R. Speyer, and R. Pless, "Geolocating static cameras," in *2007 IEEE 11th International Conference on Computer Vision*, 2007, pp. 1–6.

[15] M. Bansal, H. S. Sawhney, H. Cheng, and K. Daniilidis, "Geo-localization of street views with aerial image databases," in *Proceedings of the 19th ACM International Conference on Multimedia*, ser. MM '11. New York, NY, USA: Association for Computing Machinery, 2011, p. 1125–1128. [Online]. Available: https://doi-org.libproxy.mit.edu/10.1145/2072298.2071954

[16] S. Workman, R. Souvenir, and N. Jacobs, "Wide-area image geolocalization with aerial reference imagery," *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 3961–3969, 2015.

[17] N. N. Vo and J. Hays, "Localizing and orienting street views using overhead imagery," in *ECCV*, 2016.

[18] R. Rodrigues and M. Tani, "Are these from the same place? seeing the unseen in cross-view image geo-localization," in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*. Los Alamitos, CA, USA: IEEE Computer Society, jan 2021, pp. 3752–3760. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/WACV48630.2021.00380

[19] R. Cao, J. Zhu, Q. Li, Q. Zhang, Q. Li, B. Liu, and G. Qiu, "Learning spatial-aware cross-view embeddings for ground-to-aerial geolocalization," in *ICIG*, 2019.

[20] T.-Y. Lin, S. Belongie, and J. Hays, "Cross-view image geolocalization," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 891–898.

[21] T.-Y. Lin, Y. Cui, S. Belongie, and J. Hays, "Learning deep representations for ground-to-aerial geolocalization," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 5007–5015.

[22] J. Bromley, J. Bentz, L. Bottou, I. Guyon, Y. Lecun, C. Moore, E. Sackinger, and R. Shah, "Signature verification using a "siamese" time delay neural network," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 7, p. 25, 08 1993.

[23] Y. Shi, X. Yu, D. Campbell, and H. Li, "Where am i looking at? joint location and orientation estimation by cross-view matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4064–4072.

[24] R. Rodrigues and M. Tani, "Global assists local: Effective aerial representations for field of view constrained image geo-localization," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 3871–3879.

[25] Y. Shi and H. Li, "Beyond cross-view image retrieval: Highly accurate vehicle localization using satellite image," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 010–17 020.

[26] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.