Around the World in 80 Timesteps: A Generative Approach to Global Visual Geolocation

Nicolas Dufour [1,2] David Picard [1] Vicky Kalogeiton [2] Loic Landrieu[1]

[1] LIGM, Ecole des Ponts, IP Paris, CNRS, UGE [2] LIX, Ecole Polytechnique, IP Paris

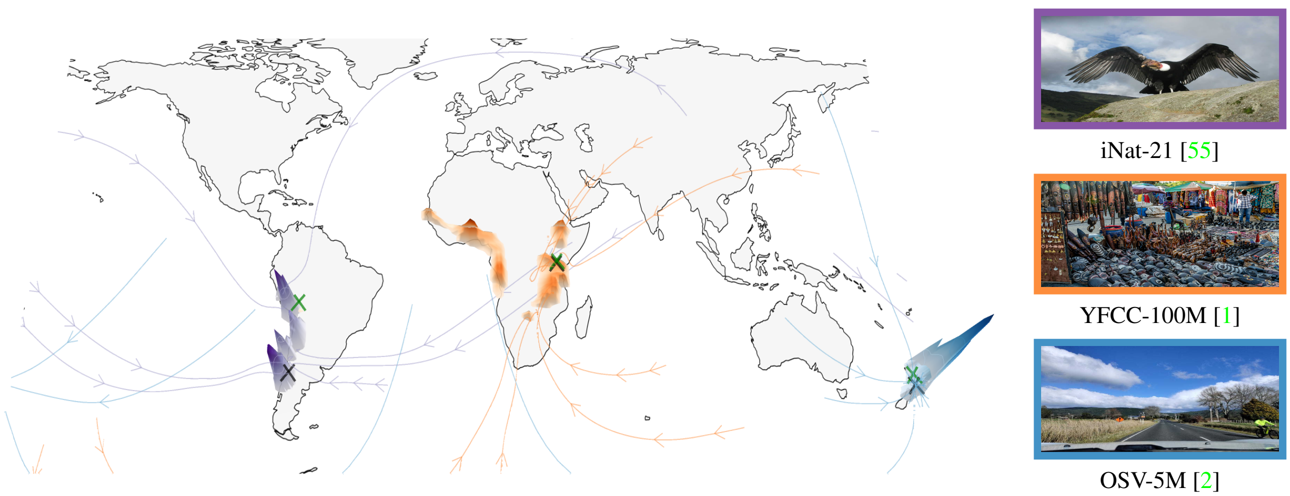


Figure 1. **Geolocation as a Generative Process.** We use diffusion/flow matching to denoise random locations into estimates, yielding trajectories on the Earth's surface and location probability densities. Examples show trajectories and log-densities for images from iNat21 [55], YFCC-100M [1], and OSV-5M [2]. Predicted: ✖, True: ✖.

Abstract

Global visual geolocation consists in predicting where an image was captured anywhere on Earth. Since not all images can be localized with the same precision, this task inherently involves a degree of ambiguity. However, existing approaches are deterministic and overlook this aspect. In this paper, we propose the first generative approach for visual geolocation based on diffusion and flow matching, and an extension to Riemannian flow matching. Our model achieves state-of-the-art performance on three visual geolocation benchmarks: OpenStreetView-5M, YFCC-100M, and iNat21. In addition, we introduce the task of probabilistic visual geolocation, where the model predicts a probability distribution over all locations instead of a single point.

1. Introduction

Knowing where an image was captured is crucial for applications like cultural heritage [9], forensics [3], and archive management [40], yet most images lack geotags [15]. This motivates the visual geolocation challenge: inferring location from image content [19, 56]. Localization precision, or *localizability* [2, 25], varies greatly (Fig. 1): landmarks like the Eiffel Tower are precise, while featureless beaches are ambiguous. Current methods (regression [2], classification [57], retrieval [41]) often ignore this inherent ambiguity, though modeling it has proven useful in vision [12, 36, 59]. Inspired by generative models like diffusion [22] and flow matching [33], we propose a novel generative approach. We use diffusion/flow-matching to denoise random locations into estimates conditioned on image features, extending manifold techniques [5] to operate on the Earth's sphere. This allows computing location likelihoods [33] and quantifying localizability. Our approach achieves state-of-the-art accuracy on OpenStreetView-5M [2], iNat21 [55], and YFCC-100M [1]. We introduce probabilistic visual geolocation (predicting location distributions) with metrics and baselines, demonstrating our method's ability to capture ambiguity.

Our contributions include:

- Introducing the first diffusion and Riemannian flow matching methods for visual geolocation.
- Extending density estimation for flow matching to geolocation for likelihood/localizability computation.
- Achieving SOTA results by explicitly modeling geoloca-

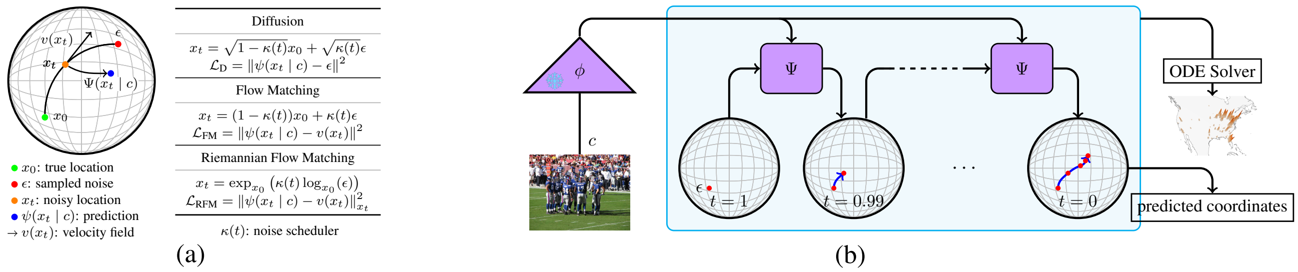


Figure 2. (a) **Generative Framework.** Comparison of diffusion $\mathbb{R}^3$, flow matching $\mathbb{R}^3$, and Riemannian flow matching $\mathcal{S}_2$, with their noising processes and losses. (b) **Inference Pipeline.** An image is embedded, noise is sampled, and iterative denoising from $t = 1$ to 0 using reverse diffusion/flow matching yields the predicted location. The model can also output a probability distribution via an ODE.

tion ambiguity.

- Proposing the probabilistic visual geolocation task with metrics and baselines.

2. Related Work

Visual geolocation [20] predicts image coordinates using retrieval (handcrafted [19, 35, 41] or deep features [56]), classification over global cells (grids [57], adaptive [7], semantic [53], admin [18, 46]), or hybrid methods [2, 18, 28, 56]. Uncertainty estimation [27], vital for localization [11, 12, 29, 36, 43, 54], leverages Bayesian [26, 37, 60] or distribution prediction [25] techniques. Generative models, including diffusion [8, 21, 22, 44, 45, 47, 48, 50, 52] and flow matching [16, 32], excel at modeling uncertainty [4, 14, 24, 31, 34, 39, 58], learning on manifolds [6], and are increasingly adapted for discriminative tasks [30]. We propose leveraging their ability to learn the data distribution manifold for superior visual geolocation.

3. Method

We first present our diffusion-based approach (Sec. 3.1) and extend it to Riemannian flow matching (Sec. 3.2), see Fig. 2. We then describe predicting location distributions (Sec. 3.3) and detail implementation choices (Sec. 3.4).

Notations. Given an image c, we predict its location x_0 on Earth, modeled as the unit sphere $\mathcal{S}_2 \subset \mathbb{R}^3$. We aim to model the conditional distribution $p(y \mid c)$ for any $y \in \mathcal{S}_2$. ϵ denotes noise, x_t noisy coordinates at time t, and ψ the network to optimize.

3.1. Geographic Diffusion

Training. We adapt diffusion models [22, 52] for geolocation. Given a coordinate-image pair (x_0, c) from a dataset Ω of geotagged images, and random coordinates ϵ from $\mathcal{N}(0, \mathbf{I_3}) \in \mathbf{R}^3$, we define noisy coordinates $x_t = \sqrt{1-\kappa(t)}x_0 + \sqrt{\kappa(t)}\epsilon$, where $\kappa(t) : [0, 1] \to [0, 1]$ with $\kappa(0) = 0$ and $\kappa(1) = 1$ is the noise scheduler. We train $\psi(x_t \mid c)$ to predict ϵ by minimizing the diffusion loss:

$$\mathcal{L}_{\text{D}} = \mathbb{E}_{x_0,c,\epsilon,t}\left[\|\psi(x_t \mid c) - \epsilon\|^2\right] , \tag{1}$$

where the expectation is over $(x_0, c) \sim \Omega$, $\epsilon \sim \mathcal{N}(0, \mathbf{I})$, and $t \sim \mathcal{U}[0, 1]$, the uniform distribution over $[0, 1]$.

Inference. To predict the likely locations for a new image c, we start by sampling a random coordinate $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ and initialize $x_1 = \epsilon$. We then iteratively refine the coordinate x_t over N timesteps from $t = 1$ to $t = 0$ using the Denoising Diffusion Implicit Models (DDIM) sampling procedure [51]. At the end of the denoising process ($t = 0$), we project the predicted location to the the Earth's surface $\mathcal{S}_2$. See Fig. 2 for an illustration of the inference process.

3.2. Extension to Riemannian Flow Matching

We extend our approach to flow matching [33], first on $\mathbb{R}^3$, then on the sphere $\mathcal{S}_2$.

Flow Matching in $\mathbb{R}^3$. We define a mapping from the true coordinates x_0 to random noise ϵ: $x_t = (1-\kappa(t))x_0 + \kappa(t)\epsilon$, inducing the velocity field $v(x_t) = \frac{dx_t}{dt} = \dot{\kappa}(t)(\epsilon - x_0)$, where $\dot{\kappa}$ the derivative of κ with respect to t. We train ψ to predict this velocity field conditionally to the image c:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{x_0,c,\epsilon,t}\left[\|\psi(x_t \mid c) - v(x_t)\|^2\right] , \tag{2}$$

with the expectation taken over the same distributions as in Eq. (1). During inference, we solve the Ordinary Differential Equation (ODE) initialized at a random coordinate ϵ, integrating backward from $t = 1$ to $t = 0$ using the predicted velocity field $\psi(x_t \mid c)$. At the end of the integration, we project x_0 onto the sphere.

Riemannian Flow Matching on the Sphere. Since our data lies on the sphere $\mathcal{S}_2$, we use Riemanian flow matching [5] to constrain the flow matching process to $\mathcal{S}_2$. This implies three conditions: (i) all true coordinates x_0 lie on $\mathcal{S}_2$, which is naturally satisfied since we are working with coordinates on the Earth's surface; (ii) the noise samples ϵ lie on $\mathcal{S}_2$, which we achieved by sampling ϵ uniformly on $\mathcal{S}_2$; and (iii) the noisy coordinates x_t remain on $\mathcal{S}_2$. We define the noisy coordinates along the geodesic between the true coordinate x_0 and the noise sample ϵ, parameterized by $\kappa(t)$: $x_t = \exp_{x_0}\left(\kappa(t)\log_{x_0}(\epsilon)\right)$, where $\log_{x_0}$ is the logarithmic map mapping point of $\mathcal{S}_2$ to the tangent space at x_0, and $\exp_{x_0}$ is the exponential map, mapping tangent vectors back to the manifold. This parametrization induces a velocity field $v(x_t)$ defined on the tangent space of x_t: $v(x_t) = \dot{\kappa}(t) \cdot D(x_t)$, where $D(x_t)$ is the tangent vector at x_t pointing along the geodesic from x_0 to ϵ, with magnitude equal to the geodesic distance between x_0 and ϵ. We train our model ψ to approximate this velocity field by minimizing

$$\mathcal{L}_{\text{RFM}} = \mathbb{E}_{x_0,c,\epsilon,t}\left[\|\psi(x_t|c) - v(x_t)\|^2_{x_t}\right] , \tag{3}$$

with $(x_0, c) \sim \Omega$, $\epsilon \sim \mathcal{U}(\mathcal{S}_2)$ $t \sim \mathcal{U}[0, 1]$, and $\|\cdot\|_{x_t}$ denotes the norm induced by the Riemannian metric on the tangent space at x_t. During inference, we solve the ODE starting from a random point $\epsilon \in \mathcal{S}_2$ and integrating backward from $t = 1$ to $t = 0$ using the predicted velocity and projecting the iterates on the manifold at each step. This ensures that the trajectory remains on the sphere $\mathcal{S}_2$ throughout the integration process.

3.3. Guidance and Density Prediction

We adapt classifier-free guidance [23] and compute location likelihoods $p(y \mid c)$.

Guided Geolocation. Train ψ on both $p(y \mid c)$ and $p(y|\varnothing)$ by randomly dropping condition c. Inference uses adjusted velocity $\hat{\psi}(x_t \mid c) = \psi(x_t \mid c) + \omega(\psi(x_t \mid c) - \psi(x_t \mid \varnothing))$, with guidance scale $\omega \geq 0$. $\omega > 0$ sharpens conditioning.

Predicting Distributions. Following [33], we compute $\log p(y \mid c)$ by solving an ODE system derived from mass conservation principles. For a location y, solve for $(x_t, f(t))$ from $t = 0$ to 1:

$$\frac{d}{dt}\begin{bmatrix} x_t \\ f(t) \end{bmatrix} = \begin{bmatrix} \psi(x(t) \mid c) \\ -\operatorname{div}\ \psi(x_t \mid c) \end{bmatrix} \text{ with } \begin{bmatrix} x_0 \\ f(0) \end{bmatrix} = \begin{bmatrix} y \\ 0 \end{bmatrix} , \tag{4}$$

where $f(t)$ accumulates negative divergence. Then $\log p(y \mid c) = \log p_\epsilon(x(1) \mid c) - f(1)$, with p_ϵ the noise distribution.

3.4. Implementation

Scheduler. We use a skewed sigmoid scheduler $\kappa(t) = \frac{\sigma(\alpha) - \sigma(\alpha + t(\beta - \alpha))}{\sigma(\alpha) - \sigma(\beta)}$ (with $\alpha = -3, \beta = 7$) that prioritizes early timesteps (closer to x_0) to focus on fine-grained cues, where $\sigma(t) = 1/(1 + \exp(-t))$ is the sigmoid function.

Model Architecture. 6 Blocks use MLPs w/ GELU and AdaLN for conditioning. We input noisy coordinates x_t, embedding c (from frozen ϕ), and PE features of $\kappa(t)$.

Table 1. **Geolocation Performance.** Comparison of geolocation precision for traditional, generative, and our proposed approaches.

		OSV-5M [2]					iNat21 [55]
		geos. ↑ /5000	dist ↓ (km)	accuracy ↑ (in %) country	region	city	dist ↓ (km)
deterministic	SC 0-shot [17]	2273	2854	38.4	20.8	14.8	
	Regression [2]	3028	1481	56.5	16.3	0.7	
	ISNs [38]	3331	2308	66.8	39.4	4.2	
	Hybrid [2]	3361	1814	68.0	39.4	5.9	
	SC Retrieval [17]	3597	1386	73.4	**45.8**	**19.9**	
generative	Uniform	131	10052	2.4	0.1	0.0	10,010
	vMF	2776	2439	52.7	17.2	0.6	6270
	vMFMix [25]	1746	5662	34.2	11.1	0.3	4701
	Diff $\mathbb{R}^3$ (ours)	3762	1123	75.9	40.9	3.6	3057
	FM $\mathbb{R}^3$ (ours)	3688	1149	74.9	40.0	4.2	2942
	RFM $\mathcal{S}_2$ (ours)	**3767**	**1069**	**76.2**	**44.2**	5.4	**2500**
		YFCC-4k [1, 56]					
		geos. ↑ /5000	dist ↓ (km)	accuracy ↑ (in %) 25km	200km	750km	2500km
deterministic	PlaNet [57]			14.3	22.2	36.4	55.8
	CPlaNet [49]			14.8	21.9	36.4	55.5
	ISNs [38]			16.5	24.2	37.5	54.9
	Translocator [46]			18.6	27.0	41.1	60.4
	GeoDecoder [7]			24.4	33.9	50.0	68.7
	PIGEON [18]			24.4	40.6	**62.2**	**77.7**
generative	Uniform	131.2	10052	0.0	0.0	0.3	3.8
	vMF	1847	3563	4.8	15.0	30.9	53.4
	vMFMix [25]	1356	4394	0.4	8.8	20.9	41.0
	Diff $\mathbb{R}^3$ (ours)	2845	2461	11.1	37.7	54.7	71.9
	FM $\mathbb{R}^3$ (ours)	2838	2514	22.1	35.0	53.2	73.1
	RFM $\mathcal{S}_2$ (ours)	2889	2461	23.7	36.4	54.5	73.6
	RFM$_{10M}$ $\mathcal{S}_2$ (ours)	**3210**	**2058**	**33.5**	**45.3**	61.1	**77.7**

4. Experiments

We evaluate global visual geolocation in Sec. 4.1, and probabilistic visual geolocation in Sec. 4.2.

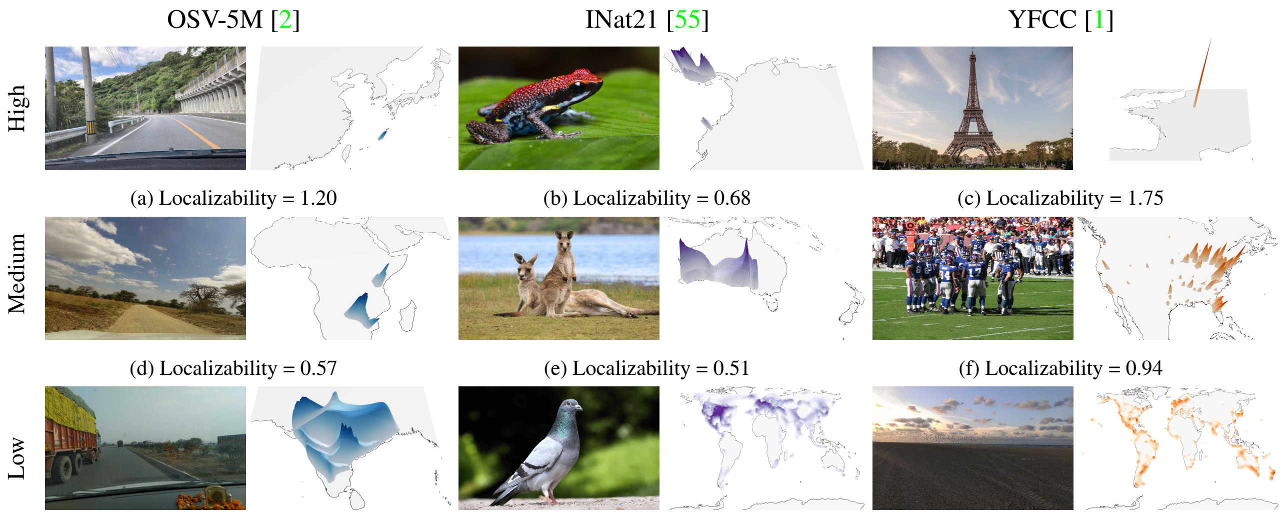
Datasets: We use **OpenStreetView-5M (OSV-5M)** [2] (5M street views), **iNat21** [55] (2.7M animal images), and **YFCC** [1] (48M geotagged images, evaluated on YFCC4k [56]).

Model Parameterization. We evaluate our three generative approaches: diffusion and flow matching in $\mathbb{R}^3$ (**Diff** $\mathbb{R}^3$ and **FM** $\mathbb{R}^3$), and Riemannian Flow-Matching on the sphere (**RFM** $\mathcal{S}_2$). Models train for 1M iterations (except **RFM$_{10M}$** $\mathcal{S}_2$ at 10M) on respective dataset training sets. Backbone ϕ is DINOv2-L [42] w/ registers [10], except OSV-5M uses StreetCLIP [17] ViT-L [13] (**SC**). Network ψ has 36M params (9.2M for iNat21). Guidance scale $\omega = 2$ for location prediction, $\omega = 0$ for distribution (Sec. 4.2).

4.1. Visual Geolocation Performance

Metrics. We use: **Distance** (Haversine km); **GeoScore** ($5000 \exp(-\delta/1492.7)$, [18]); **Accuracy** (% within country/region/city/distance).

Results. Table 1 shows our models achieve SOTA geolocation performance, surpassing existing methods and our baselines. Our generative approach significantly outperforms non-retrieval methods (e.g., +406 GeoScore vs. Astruc *et*



(a) Localizability = 1.20 (b) Localizability = 0.68 (c) Localizability = 1.75

(d) Localizability = 0.57 (e) Localizability = 0.51 (f) Localizability = 0.94

(g) Localizability = 0.49 (h) Localizability = 0.41 (i) Localizability = 0.47

Figure 3. **Estimating Localizability.** We use the entropy of the predicted distribution as a proxy for the localizability of images. For each dataset, we present examples of high, medium, and low localizability, which correlate well with human perception.

Table 2. **Probabilistic Visual Geolocation.** Evaluation of predicted distribution quality. Note: $\mathbb{R}^3$ and $\mathcal{S}_2$ likelihoods are not directly comparable. Log-likelihoods/entropies can be negative for continuous distributions. Generative metrics only shown for iNat21 for space.

	OSV-5M	YFCC	iNat21				
	NLL ↓	NLL ↓	NLL ↓	precision ↑	recall ↑	density ↑	coverage ↑
Uniform	1.22	1.22	1.22	0.58	**0.98**	0.38	0.22
vMF Regression	10.13	0.01	1.99	0.52	**0.98**	0.37	0.24
vMFMix	0.06	-0.04	-0.23	0.63	**0.98**	0.47	0.29
RFlowMatch $\mathcal{S}_2$ (ours)	**-1.51**	**-3.71**	**-1.94**	**0.88**	0.95	**0.78**	**0.59**
Diffusion $\mathbb{R}^3$ (ours)	0.58	0.63	0.68	0.76	**0.98**	0.60	0.44
FlowMatch $\mathbb{R}^3$ (ours)	**-5.01**	**-7.15**	**-4.00**	0.76	0.97	0.61	0.47

al. [2]). Longer training helps. Retrieval methods remain better at very fine scales. Among generative models, Flow Matching (FM) improves over Diffusion (Diff), while Riemannian FM ($\mathcal{S}_2$) outperforms Euclidean FM ($\mathbb{R}^3$), highlighting benefits of modeling Earth's geometry. Single vMF is on par with regression while vMFMix overfits.

4.2. Probabilistic Visual Geolocation

Beyond predicting a single location, our model estimates a distribution $p(y \mid c)$ over locations $y \in \mathcal{S}^2$, capturing geolocation uncertainty.

Metrics. We evaluate distribution quality using: **Negative Log-Likelihood (NLL)**, the average NLL per-dimension of true locations x_i under predicted distributions $p(y \mid c_i)$, lower is better: $\text{NLL} = -\frac{1}{3N}\sum_{i=1}^{N}\log_2 p(x_i \mid c_i)$; **Localizability**, the negative entropy of $p(y \mid c)$, estimated via Monte Carlo, higher is more confident: $\text{Localizability}(c) = \int_{\mathcal{S}^2} p(y \mid c)\log_2 p(y \mid c)dy$; and **Generative Metrics** (Precision, Recall, Density, Coverage).

Results. Table 2 shows our models achieve lower NLL than baselines, indicating better distribution alignment. FM $\mathbb{R}^3$ yields better NLL than Diffusion. vMFMix improves over single vMF, suggesting better ambiguity handling despite lower geolocation accuracy. RFM $\mathcal{S}^2$ excels on generative metrics, likely because it operates directly on the sphere, avoiding projection errors inherent in $\mathbb{R}^3$ models.

Localizability. Figure 3 demonstrates that negative entropy correlates with perceived image localizability: high scores for distinct landmarks (c, Eiffel Tower), medium for broader regions (f, NFL stadiums), and low for ambiguous scenes (i, featureless beach).

5. Conclusion

We presented a generative visual geolocation method using diffusion and Riemannian flow matching on the sphere, capturing inherent spatial ambiguity often ignored by deterministic approaches. Our method achieves state-of-the-art performance on standard benchmarks. We also introduced probabilistic visual geolocation, demonstrating our model's ability to predict accurate probability distributions

References

[1] YFCC100m. https://gitlab.com/jfolz/yfcc100m, accessed: 2023-10-10 1, 3, 4

[2] Astruc, G., Dufour, N., Siglidis, I., Aronssohn, C., Bouia, N., Fu, S., Loiseau, R., Nguyen, V.N., Raude, C., Vincent, E., et al.: OpenStreetView-5M: The many roads to global visual geolocation. In: CVPR (2024) 1, 2, 3, 4

[3] Bamigbade, O., Sheppard, J., Scanlon, M.: Computer vision for multimedia geolocation in human trafficking investigation: A systematic literature review. In: arXiv preprint arXiv:2402.15448 (2024) 1

[4] Berry, L., Brando, A., Meger, D.: Shedding light on large generative networks: Estimating epistemic uncertainty in diffusion models. In: UAI (2024) 2

[5] Chen, R.T., Lipman, Y.: Riemannian flow matching on general geometries. In: ICLR (2024) 1, 2

[6] Chen, R.T., Lipman, Y.: Riemannian flow matching on general geometries. In: ICLR (2024) 2

[7] Clark, B., Kerrigan, A., Kulkarni, P.P., Cepeda, V.V., Shah, M.: Where we are and what we're looking at: Query based worldwide image geo-localization using hierarchies and scenes. In: CVPR (2023) 2, 3

[8] Courant, R., Dufour, N., Wang, X., Christie, M., Kalogeiton, V.: ET the exceptional trajectories: Text-to-camera-trajectory generation with character awareness. In: ECCV (2024) 2

[9] Daoud, M., Huang, J.X.: Mining query-driven contexts for geographic and temporal search. International Journal of Geographical Information Science (2013) 1

[10] Darcet, T., Oquab, M., Mairal, J., Bojanowski, P.: Vision transformers need registers. ICLR (2024) 3

[11] Dellaert, F., Fox, D., Burgard, W., Thrun, S.: Monte Carlo localization for mobile robots. In: ICRA (1999) 2

[12] Deng, H., Bui, M., Navab, N., Guibas, L., Ilic, S., Birdal, T.: Deep Bingham networks: Dealing with uncertainty and ambiguity in pose estimation. International Journal of Computer Vision (2022) 1, 2

[13] Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. ICLR (2021) 3

[14] Dufour, N., Besnier, V., Kalogeiton, V., Picard, D.: Don't drop your samples! Coherence-aware training benefits conditional diffusion. In: CVPR (2024) 2

[15] Flatow, D., Naaman, M., Xie, K.E., Volkovich, Y., Kanza, Y.: On the accuracy of hyper-local geotagging of social media content. In: International Conference on Web Search and Data Mining (2015) 1

[16] Grathwohl, W., Chen, R.T., Bettencourt, J., Sutskever, I., Duvenaud, D.: FFJORD: Free-form continuous dynamics for scalable reversible generative models. In: ICLR (2019) 2

[17] Haas, L., Alberti, S., Skreta, M.: Learning generalized zero-shot learners for open-domain image geolocalization. In: arXiv preprint arXiv:2302.00275 (2023) 3

[18] Haas, L., Alberti, S., Skreta, M.: PIGEON: Predicting image geolocations. In: CVPR (2023) 2, 3

[19] Hays, J., Efros, A.A.: Im2GPSs: Estimating geographic information from a single image. In: CVPR (2008) 1, 2

[20] Hays, J., Efros, A.A.: Large-scale image geolocalization. Multimodal location estimation of videos and images (2015) 2

[21] Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., et al.: Imagen video: High definition video generation with diffusion models. arXiv (2022) 2

[22] Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS (2020) 1, 2

[23] Ho, J., Salimans, T.: Classifier-free diffusion guidance. In: NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications (2021) 3

[24] Huang, B., Yu, W., Xie, R., Xiao, J., Huang, J.: Two-stage denoising diffusion model for source localization in graph inverse problems. In: ECML-PKDD. Springer (2023) 2

[25] Izbicki, M., Papalexakis, E.E., Tsotras, V.J.: Exploiting the Earth's spherical geometry to geolocate images. In: MLKDD (2020) 1, 2, 3

[26] Kendall, A., Cipolla, R.: Modelling uncertainty in deep learning for camera relocalization. In: ICRA (2016) 2

[27] Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? In: NeurIPS (2017) 2

[28] Kordopatis-Zilos, G., Galopoulos, P., Papadopoulos, S., Kompatsiaris, I.: Leveraging EfficientNet and contrastive learning for accurate global-scale location estimation. In: International Conference on Multimedia Retrieval (2021) 2

[29] Levinson, J., Thrun, S.: Robust vehicle localization in urban environments using probabilistic maps. In: ICRA (2010) 2

[30] Li, A.C., Prabhudesai, M., Duggal, S., Brown, E., Pathak, D.: Your diffusion model is secretly a zero-shot classifier. In: CVPR (2023) 2

[31] Li, W., Yang, Y., Yu, S., Hu, G., Wen, C., Cheng, M., Wang, C.: Diffloc: Diffusion model for outdoor lidar localization. In: CVPR (2024) 2

[32] Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: ICLR (2023) 2

[33] Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In:

The Eleventh International Conference on Learning Representations (2024) 1, 2, 3
[34] Mackowiak, R., Ardizzone, L., Kothe, U., Rother, C.: Generative classifiers as a basis for trustworthy image classification. In: CVPR (2021) 2
[35] Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: ICCV (2001) 2
[36] Merrill, N., Guo, Y., Zuo, X., Huang, X., Leutenegger, S., Peng, X., Ren, L., Huang, G.: Symmetry and uncertainty-aware object SLAM for 6DOF object pose estimation. In: CVPR (2022) 1, 2
[37] Mullane, J., Vo, B.N., Adams, M.D., Vo, B.T.: A random-finite-set approach to Bayesian SLAM. IEEE transactions on robotics (2011) 2
[38] Muller-Budack, E., Pustu-Iren, K., Ewerth, R.: Geolocation estimation of photos using a hierarchical model and scene classification. In: ECCV (2018) 3
[39] Nicolas Dufour, David Picard, V.K.: SCAM! Transferring humans between images with semantic cross attention modulation. In: ECCV (2022) 2
[40] Nikolaidou, K., Seuret, M., Mokayed, H., Liwicki, M.: A survey of historical document image datasets. International Journal on Document Analysis and Recognition (2022) 1
[41] Oliva, A., Torralba, A.: Building the gist of a scene: The role of global image features in recognition. Progress in brain research (2006) 1, 2
[42] Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., et al.: DINOv2: Learning robust visual features without supervision. TMLR 3
[43] Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K.: Coarse-to-fine volumetric prediction for single-image 3D human pose. In: CVPR (2017) 2
[44] Petrovich, M., Litany, O., Iqbal, U., Black, M.J., Varol, G., Peng, X.B., Rempe, D.: Multi-track timeline control for text-driven 3D human motion generation. In: CVPR Workshop on Human Motion Generation (2024) 2
[45] Polyak, A., Zohar, A., Brown, A., Tjandra, A., Sinha, A., Lee, A., Vyas, A., Shi, B., Ma, C.Y., Chuang, C.Y., et al.: Movie Gen: A cast of media foundation models. arXiv (2024) 2
[46] Pramanick, S., Nowara, E.M., Gleason, J., Castillo, C.D., Chellappa, R.: Where in the world is this image? Transformer-based geo-localization in the wild. In: ECCV (2022) 2, 3
[47] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022) 2
[48] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. In: NeurIPS (2022) 2
[49] Seo, P.H., Weyand, T., Sim, J., Han, B.: Cplanet: Enhancing image geolocalization by combinatorial partitioning of maps. In: ECCV (2018) 3
[50] Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2021) 2
[51] Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2021) 2
[52] Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: ICLR (2021) 2
[53] Theiner, J., Müller-Budack, E., Ewerth, R.: Interpretable semantic photo geolocation. In: WACV (2022) 2
[54] Tompson, J.J., Jain, A., LeCun, Y., Bregler, C.: Joint training of a convolutional network and a graphical model for human pose estimation. In: NeurIPS (2014) 2
[55] Van Horn, G., Cole, E., Beery, S., Wilber, K., Belongie, S., Mac Aodha, O.: Benchmarking representation learning for natural world image collections. In: CVPR (2021) 1, 3, 4
[56] Vo, N., Jacobs, N., Hays, J.: Revisiting IMG2GPS in the deep learning era. In: ICCV (2017) 1, 2, 3
[57] Weyand, T., Kostrikov, I., Philbin, J.: Planet-photo geolocation with convolutional neural networks. In: ECCV (2016) 1, 2, 3
[58] Wolleb, J., Sandkühler, R., Bieder, F., Valmaggia, P., Cattin, P.C.: Diffusion models for implicit image segmentation ensembles. In: International Conference on Medical Imaging with Deep Learning. PMLR (2022) 2
[59] Xu, L., Qu, H., Cai, Y., Liu, J.: 6D-diff: A keypoint diffusion framework for 6d object pose estimation. In: CVPR (2024) 1
[60] Zangeneh, F., Bruns, L., Dekel, A., Pieropan, A., Jensfelt, P.: A probabilistic framework for visual localization in ambiguous scenes. In: ICRA (2023) 2