

Uncertainty-aware Vision-based Metric Cross-view Geolocalization

Florian Ferver¹ Sebastian Bullinger¹ Christoph Bodenstein¹ Michael Arens¹ Rainer Stiefelhagen²
¹Fraunhofer IOSB ²Karlsruhe Institute of Technology

¹f.firstname.lastname g@iosb.fraunhofer.de ²rainer.stiefelhagen@kit.edu

Figure 1. Probability distributions for the vehicle position predicted by our model which matches the vehicle's surround camera images with an aerial image. The first and second rows show the front and back cameras in the Ford AV dataset [6]. The last row shows the aerial image with the search region in the center and driving direction pointing upwards. Blue and red color refer to low and high probability predicted by our model. Map data: Bing Maps © 2022 TomTom, © Vexcel Imaging [1].

Abstract

This paper proposes a novel method for vision-based metric cross-view geolocalization (CVGL) that matches available

the camera images captured from a ground-based vehicle with an aerial image to determine the vehicle's pose. Since aerial images are globally available at low cost, they represent a potential compromise between two established paradigms of autonomous driving, using expensive high-definition prior maps or relying entirely on the sensor data captured at runtime.

We present an end-to-end differentiable model that uses the ground and aerial images to predict a probability distribution over possible vehicle poses. We combine multiple vehicle datasets with aerial images from orthophoto providers on which we demonstrate the feasibility of our method. Since

the ground truth poses are often inaccurate w.r.t. the aerial images, we implement a pseudo-label approach to produce

more accurate ground truth poses and make them publicly

While previous works require training data from the target region to achieve reasonable localization accuracy (

same-area evaluation), our approach overcomes this limitation and outperforms previous results even in the strictly

more challenging cross-area case. We improve the previous state-of-the-art by a large margin even without ground or aerial data from the test region, which highlights the model's

potential for global-scale application. We further integrate the uncertainty-aware predictions in a tracking framework

to determine the vehicle's trajectory over time resulting in a mean position error on KITTI-360 of 0.78m.

1. Introduction

Systems for autonomous driving require both a model of the vehicle's environment as well as the location of the vehicle relative to the model. These systems either construct the full model during runtime (i.e. entirely online), or create some parts prior to runtime (i.e. partly offline). The latter methods typically construct high-definition maps of a region in advance (e.g. using lidar sensors) and localize the vehicle at runtime relative to that map [34]. While prior maps facilitate a high localization accuracy of the system, they are also expensive to construct and maintain. Online methods, on the other hand create a model of the local environment using only the live sensor readings (e.g. from lidar [52], camera [15] or both [26]). This avoids the need for expensive prior maps, but represents a more difficult task as the system has to predict both the spatial structure of the environment as well as its relative location within it.

Aerial images offer the potential to leverage the advantages of both approaches: They can be used as a prior map for localization, while also being affordable, globally available and up-to-date due to an established infrastructure of satellite and aerial orthophoto providers [1, 3]. We consider the problem of matching the sensor measurements of the vehicle against aerial images to determine the vehicle's location on the images and thereby its geo-location.

Previous research in this area focuses on methods that cover large (e.g. city-scale) search regions [23, 46], but suffer from low metric accuracy [57] insufficient for the navigation of autonomous vehicles. Since a prior pose estimate of the vehicle can be provided by global navigation satellite systems (GNSS) or by tracking the vehicle continuously, several recent methods employ smaller search regions to achieve higher metric accuracy [13, 35].

Without access to three-dimensional lidar point clouds, a purely vision-based model has to bridge the gap between ground and aerial perspectives, for example by learning the transformation in a data-centric manner. We utilize a transformer model that iteratively constructs a bird's eye view (BEV) map of the local vehicle environment by aggregating information from the ground-level perspective views (PV). The BEV refers to a nadir (i.e. orthogonal) view of the local vehicle environment. The final BEV map is matched with an aerial image to predict the relative vehicle pose with three degrees of freedom (3-DoF), i.e. a two-dimensional translation and a one-dimensional rotation.

Our model outperforms previous approaches for the metric CVGL task on the Ford AV [6] and KITTI-360 [21] datasets and even surpasses related approaches utilizing lidar sensors in addition to camera input. It predicts a soft probability distribution over possible vehicle poses (Fig. 1) rather than a single pose which specifically benefits trackers that use the model predictions to determine the vehicle's trajectory over time.

While previous works rely on the availability of training data from the target region to achieve reasonable localization accuracy, we address the strictly more challenging task of non-overlapping train and test regions. We further train and test the model on entirely different datasets that were captured with different ground-based vehicles. Our evaluation demonstrates the generalization capabilities of our model under cross-area and cross-vehicle conditions and highlights the potential for global-scale application without re-tuning on a new region or a new vehicle setup.

We collect multiple datasets from the autonomous driving sector in addition to aerial images from several orthophoto providers for our evaluation. Since the vehicle's geo-locations do not always accurately match the corresponding aerial images, we compute new geo-registered ground truth poses for all datasets used in the work and filter out invalid samples via a data-pruning approach.

We publish the source code of our method online including a common interface for the different datasets. We also make the improved ground truth for all datasets publicly available.¹

In summary, our contributions are as follows:

1. We present a novel end-to-end trainable model for metric CVGL that requires only visual input and yields uncertainty-aware predictions.
2. We collect multiple vehicle datasets and aerial images from several orthophoto providers for our evaluation. We compute improved ground truth poses using a pseudo-label approach and filter out invalid samples via data-pruning.
3. Our method outperforms previous works by a large margin even under strictly more challenging cross-area and cross-vehicle settings.

2. Related Work

Cross-view Geolocalization. CVGL refers to the task of matching camera images of a ground-based agent to geo-registered aerial images to determine the agent's geo-location. Approaches in this area typically focus on one of two problems.

Large-area CVGL methods start from a large (e.g. city-scale) search region and find a rough estimate of the agent's position. They typically use an image retrieval approach and therefore do not predict orientation or reach high metric accuracy (e.g. less than 10% of predictions reported by the state-of-the-art method TransGeo [57] are localized with less than 10m error).

Metric CVGL methods start from a rough pose estimate of the agent (e.g. up to 50m error [13]) and determine the location and orientation with higher accuracy by matching

¹Project page: <https://ffeflo.github.io/projects/vismetricvgl23>

the agent's sensor readings with an aerial image centered on the prior pose.

Zhu et al. [58] propose to differentiate the evaluation of a model into same-area and cross-area categories based on the availability of data from the test region during training. While same-area models are trained on the same aerial images used at test time and outperform cross-area models [57, 58] they require obtaining data from the target region first. This limits their scalability and contradicts our motivation for using aerial images: global availability at low cost. Furthermore, same-area models have not been shown to memorize large areas (e.g. country-scale) from training data, even when such data is available. We therefore consider only the strictly more challenging task of cross-area CVGL.

While metric CVGL utilizing lidar and camera sensors has been shown to achieve sub-meter accurate poses [13], previous purely vision-based methods have not reached comparable performance. Zhu et al. [58] treat the problem as a regression task on top of large-area CVGL. Xiao et al. [47, 48] utilize image retrieval methods from large-area CVGL and adapt them to work with smaller search regions for metric CVGL. These methods do not explicitly consider the spatial layout of both aerial and ground input data, since the images are reduced to one-dimensional feature vectors where spatial information can only be stored implicitly in the neurons' activations.

Shi et al. [35] propose the first method that digresses from the image retrieval paradigm of large-area CVGL by using an end-to-end differentiable Levenberg-Marquardt optimizer that iteratively estimates the relative pose between aerial and ground-level images. They rely on a flat-ground assumption by using a homography to project satellite features to the ground-level view. They further do not consider a multi-camera setup.

Perspective View to Bird's Eye View Cameras mounted on vehicles show a two-dimensional perspective projection of the environment typically as pinhole or panoramic images. However, the preferred representation for many tasks (navigation, object detection or localization), is a BEV, a two-dimensional orthographic projection of the ground-level features in the top-down view. The perspective view to bird's eye view transformation (PV2BEV) represents a novel research field that has recently gained attention in the research community. [27]

PV2BEV methods can be categorized based on whether they explicitly exploit the geometry of the scene to bridge the gap between PV and BEV or learn the mapping in a data-centric manner.

Geometry-based methods typically use one of two approaches. Inverse Projective Mapping [28] transforms PV features to BEV via a homography based on the camera's intrinsic and extrinsic parameters [7, 33]. These methods rely on a flat-ground assumption and are not able to properly ex-

Figure 2. Overview of the architecture. (a) Aerial and ground input images are processed by separate encoders to predict high-level feature maps. (b) A bird's eye view (BEV) representation of the local vehicle environment is constructed. The BEV map is initialized as a grid of learnable parameters and iteratively refined by cross-attending to the ground features. (c) The final BEV and aerial features are matched via cross-correlation to predict a probability distribution over possible 3-DoF vehicle poses.

exploit features above or below the ground-plane. Depth-based methods explicitly predict depth in the ground images as discrete point-clouds or probabilistic depth distributions and utilize a three-dimensional model (e.g. point-cloud or voxel-map) for the projection [31, 32, 41].

Recent learning-based methods utilize transformers to project features from PV to BEV. Here, queries defined in BEV space gather information from values defined in PV space via a cross-attention mechanism. Queries are either defined sparsely (e.g. for object detection) [10, 24, 42] or as a dense spatial grid around the vehicle (e.g. for semantic segmentation) [20, 30, 56]. Some methods further utilize deformable attention to reduce the memory consumption, such that each query attends only to a sparse set of points in the PV rather than to all PV features [20, 30].

3. Method

Our model predicts a probability distribution for the pose of a ground-based vehicle relative to an aerial image. It first builds a BEV representation of the local vehicle environment using a cross-attention mechanism. The BEV is then matched with the aerial image to determine the vehicle's pose. Fig. 2 shows a summary of the model.

3.1. Feature extraction

The ground and aerial input images are first processed by encoder networks that predict pixelwise features at stride s $2 \times N$, i.e. with $\frac{1}{s}$ of the original resolution. We use a shared

encoder for the ground images and a separate encoder for the aerial image (cf. Fig. 2a).

We choose a lightweight architecture for the network based on common design principles as follows. For a given image, we apply a pretrained vision backbone (ConvNeXt [25]) to extract a pyramid of intermediate feature maps L at strides s_L : 4; 8; 16 and 32. We use a global average for context pooling in the last feature map [12, 55]. All feature maps in L are resized to stride s_v via bilinear interpolation, summed and processed by a small multilayer perceptron (MLP) to produce the final output feature map. More details are provided in the supplementary material.

The ground image of the i -th camera is encoded into feature map F_{Gi} with $s_{G_i} = 1$. Similarly, the aerial image is encoded into feature map F_A with $s_{A_i} = 1$ since its spatial resolution is particularly important for the localization accuracy. Since vision backbones typically contain a stem that initially reduces the resolution r_g by a factor of 4, we additionally process the RGB input with two ResNet blocks [16] at stride 1 and add the result to the intermediate list of feature maps L to fully exploit the aerial image's spatial information. Our evaluation justifies this choice.

3.2. Perspective View to Bird's Eye View

Overview. The BEV of the vehicle's local environment is iteratively constructed from the PVs captured by cameras at a given point in time. The BEV is centered on the vehicle and defined as a spatial grid $\mathbb{R}^{d_B \times d_B \times c_B}$ with dimensions d_B (at q_B meters per cell) and c_B channels. It is initialized via learnable parameters and iteratively refined in n_{blocks} steps. We regard only cells B with a distance of less than $\frac{d_B}{2} \cdot q_B$ meters to the vehicle as valid and store the corresponding mask $M \in \{0, 1\}^{d_B \times d_B}$.

Each refinement step consists of two transformer blocks that apply cross-attention to the PV features as well as self-attention on B (cf. Fig. 2b). In the following, the transformer block is reviewed and contrasted with the attention mechanisms used in the cross- and self-attention blocks of our model.

The general layout of the transformer block [38, 53] is shown in Fig. 3. It operates on a set of query tokens that adaptively aggregate information from a set of value tokens T_v . Tokens are generally packed into matrix form $\mathbb{R}^{n \times c}$ with c channels per token. The transformer block consists of two consecutive residual sub-blocks: The first contains an attention mechanism that distributes information from corresponding locations. The sampled features represent the value tokens to query tokens. The second contains a MLP that processes tokens separately and is largely responsible for the representational power of the block.

The transformer blocks used in our model are based on different choices for the attention mechanism. The original design of transformers [38] uses query-key-value (QKV) into each camera coordinate system and projected onto the attention for this purpose: Tokens T_q are first projected onto

Figure 3. Transformer block [38, 53]. Queries representing the bird's eye view (BEV) map are used to attend to a set of values. Cross- and self-attention blocks differ based on the choice for the values (i.e. PV features F_{Gi} and BEV features B , respectively) and the type of attention mechanism that is used (Sec. 3.2). The attention logits are optionally forwarded via a skip connection to the next attention block of the same type. Normalization is applied in the residual block (Pre-LN) [51].

queries Q , and tokens T_v onto keys K and values V by learned linear transformations. We use the term T_q for T_q and Q , and value for T_v and V interchangeably. For each query Q_i , information is aggregated from values based on the similarity of the query Q_i w.r.t. the value's corresponding key K_j . The attention map A represents the weights assigned to each query-value pair and is defined as

$$A = \text{softmax}(A_{\text{logit}}) \quad \text{with } A_{\text{logit}} = \frac{QK^T}{c_{qk}} \quad (1)$$

where c_{qk} is the channel dimension of Q and K . The values are averaged based on the weights A and then linearly projected to produce the final output tokens.

For multi-head attention, the matrices Q , K and V are first split along the channel axis into n_{heads} equally sized blocks before computation of the weighted averages. The outputs of all heads are concatenated before the final linear projection. This enables each query to incorporate distinct aggregations of the input values for n_{heads} heads.

Cross-attention block. The cross-attention block gathers information from the PV features F_{Gi} for all cells of the BEV map B . Computing full cross-attention between queries B and values F_{Gi} such that each query attends to each PV feature leads to large memory and computational cost. Instead, for each cell in the BEV we select a small set of points in the PV and sample the features F_{Gi} at the corresponding locations. The sampled features represent the value tokens for the corresponding BEV query token. The PV points for a given query are determined as follows.

We lift the corresponding point on the BEV into a pillar of height h_{min} to h_{max} [20]. The points are then transformed from vehicle coordinate system design of transformers [38] uses query-key-value (QKV) into each camera coordinate system and projected onto the camera plane using its extrinsic and intrinsic parameters.

Points that do not fall into the camera frustum are discarded. This yields up to M points per query per camera. Since typical camera setups for surround view have only little overlap between cameras, most queries are assigned no more than one value overall.

Since the value tokens for a given query token represent only a sparse set of points on the PV, we enable more fine-grained control over the points' locations via deformable offsets [59]. Given the pillar of points for a BEV query B_{xy} , we predict offsets $p_j \in \mathbb{R}^2$ with $j \in \{1, \dots, z_g\}$ via learnable linear transformations B_{xy} . For the j -th point in the pillar that is projected to location $p_j \in \mathbb{R}^2$ in the i -th camera view, the corresponding feature is sampled via bilinear interpolation as

$$f_{ij} = F_{Gi} \left(\frac{p_{ij} + p_j}{s_G} \right) \quad (2)$$

where s_G represents the stride of the PV feature map. The offsets of each cross-attention block are further added onto the predicted offsets of the next block via a skip connection, such that each block learns to refine the existing offsets rather than predict entirely new offsets.

While queries are represented by a dense grid, values are given as a sparse set of features and cannot efficiently be packed into a dense spatial representation. We therefore implement values as a list of features $F_G \in \mathbb{R}^{n_v \times c_v}$ with c_v channels that contains all valid features concatenated along the first axis. The interaction between F_G and spatial grid B is implemented via efficient scatter and gather operations [8].

To reduce the computational complexity of the transformer block, we simplify the computation of the attention map A as follows. Instead of mapping query tokens and value tokens F_G onto queries Q and keys K respectively, we predict the attention logits A_{logit} directly from the BEV features B via a learnable linear projection. Our ablation studies demonstrate that this does not lead to reduced performance. We further add a skip connection between the attention logits of subsequent blocks such that each block learns to refine the existing weights rather than predict entirely new weights [17].

Self-attention block. The self-attention block refines the BEV representation via a self-attention operation. We choose a single block of the SegFormer architecture for this purpose [50].

In classical self-attention B is used both for query and value tokens. To avoid the large memory and computational cost of full attention, SegFormer uses spatial-reduction attention (SRA) [40]: While query tokens are given directly by the spatial resolution B is first reduced via a convolution with strides s_R for the value tokens. This reduces the number of value tokens and thereby the computational complexity by s_R^2 . The MLP component of the self-attention block is further extended with a 3×3 depthwise convolution to mimic

the use of positional encodings. Invalid features are set to zero according to mask M .

3.3. Bird's Eye View to Aerial View matching

In the last step of the model, the aerial features F_A and the BEV features B are matched to determine the relative 3-DoF pose of the vehicle α (Fig. 2c). We test different hypotheses $h \in \mathbb{H} \subseteq \text{SE}(2)$ for the vehicle pose by comparing F_A with B .

In general, we choose a more fine-grained pixel resolution $q_A = q_B$ for F_A which benefits localization accuracy. Therefore B is first upsampled to match the pixel resolution q_A via bilinear interpolation followed by a linear projection to c_A channels.

To test a hypothesis h , the upsampled BEV map is transformed into aerial coordinates by the rigid transformation yielding the transformed BEV $B^{(h)}$ with the same dimensions as F_A . The logit of h is determined as the scaled inner product of F_A and $B^{(h)}$. A softmax operation is applied to the logits of all hypotheses to produce the final probability distribution as shown in Eq. (3).

$$P(h) = \frac{\exp(k h F_A; B^{(h)})_i}{\sum_{h \in \mathbb{H}} \exp(k h F_A; B^{(h)})_i} \quad \text{with } k = \frac{1}{\sqrt{H; M; i; c_A}} \quad (3)$$

Logits are scaled with k to normalize the variance of the inner product as proposed by Vaswani et al. [38].

We choose the set of hypotheses \mathbb{H} as a two-dimensional grid of translations around the origin with pixel resolution q_A , orientation $\in \mathbb{R}$ and maximum distance $\in \mathbb{R}$. The logits of \mathbb{H} are jointly estimated by rotating B and computing the cross-correlation between F_A and the rotated B . This is repeated for a discrete set of rotations $\mathbb{S} \subseteq \mathbb{R}$ yielding the total set of evaluated hypotheses $\mathbb{H} = \mathbb{S} \times \mathbb{A} \times \mathbb{H}$. The cross-correlation is computed efficiently in the Fourier domain by utilizing the Convolution Theorem [43] which requires three Fast Fourier Transforms per rotation angle per feature channel.

3.4. Loss

Each training sample contains the camera images, the intrinsic and extrinsic parameters, a randomly chosen apriori pose and the ground truth pose of the vehicle. We define a normal distribution P_{true} centered on the vehicle pose with translation and angle standard deviations σ_t and σ_a that represents the desired model output. The loss function is defined as the cross-entropy between the predicted and target probabilities:

$$L = - \sum_{h \in \mathbb{H}} P_{\text{true}}(h) \ln P(h) \quad (4)$$

Using a soft rather than a one-hot target distribution allows training with ground truth poses that potentially lie

(a) Original ground truth poses.

(b) Our pseudo-labeled ground truth poses.

Figure 4. Comparison of original ground truth poses and pseudo-labeled poses. Projected lidar points are shown in yellow for visualization. Vehicle data: Argoverse V2 [45]. Map data: Bing Maps 2022 TomTom, © Vexcel Imaging [1].

between the discrete hypotheses. It further acts as a means for label smoothing which prevents the network from becoming over-confident [29].

4. Data

Overview. In order to build a dataset for the evaluation of CVGL methods in cross-area and cross-vehicle settings, we collect existing datasets from the autonomous driving sector (Argoverse V1 [11], Argoverse V2 [45], Ford AV [6], KITTI-360 [21], Lyft L5 [18], Nuscenes [9] and Pandaset [49]) and gather aerial images for the vehicle's geo-poses from several orthophoto providers (Google Maps [3], Bing Maps [1], DCGIS [2], MassGIS [4] and Stratmap [5]). A detailed overview of the datasets used in this work is shown in the supplementary material. In total, the combined dataset contains 1.05 million ground data-frames, each consisting of the vehicle's ground truth pose and camera images and intrinsic and extrinsic parameters. This corresponds with 2.55 million pairs of ground data-frames and corresponding aerial images when counting multiple orthophoto providers.

Since subsequent frames only have a small relative offset and multiple trajectories per dataset often follow the same route, the number of paired data-frames does not reflect the data's coverage of aerial images. We measure the coverage by grouping frames into disjoint cells of size 100m100m which results in 5.1 thousand cells containing at least one ground frame and 13.0 thousand cells when counting multiple orthophoto providers.

We group frames into cells of size 1m 1m and for each training iteration randomly sample a cell from which the next frame is chosen. This prevents areas with many ground frames from being overrepresented. We resize all ground images to a minimum size of 200 240 pixels which enables the model to run at approximately 2-3 Hz on an RTX 6000.

Pseudo-labels Similar to previous works [35, 39] we no-

Figure 5. Examples of the data-pruning method on the Ford AV dataset [6]. Blue and red color represent kept and pruned frames in the trajectory. Map data: Bing Maps 2022 TomTom © Vexcel Imaging [1].

tice that the geo-poses provided by the vehicle datasets do not accurately match the corresponding aerial images and can have large relative offsets. To address this problem, we utilize a pseudo-label approach to create ground truth poses for all datasets by using the lidar point-clouds contained in the data as follows.

We manually label a small subset of the data by aligning the vehicle's lidar point-clouds with the corresponding aerial images in top-down view. We train a variant of our model on this subset where the PV2BEV transformer is replaced with a simple geometric projection using the captured lidar point-cloud [13]. This model requires less data to train while still producing accurate localization results. We predict the poses for all pairs of ground frames and orthophoto providers. For each scene, the rigid transformations between subsequent frames provided by the dataset and the predicted poses and pose uncertainties are inserted into a pose graph and optimized using a least-squares approach [14]. The rigid transformations are factored in with high confidence and model predictions with low confidence, such that the random (non-systematic) error in the predictions averages out over sufficiently long sequences. Our evaluation in Sec. 5.3 and Sec. 5.4 justifies this approach. Fig. 4 shows a qualitative comparison between original ground truth poses and pseudo-labeled poses. We publish the improved poses online to foster future research in this area.

We do not evaluate on datasets like CVUSA [46] and CVACT [23] that are typically used in large-area CVGL since a) they do not contain accurate ground truth poses and b) our pseudo-label approach cannot be utilized since the datasets do not contain lidar point-clouds or rigid transformations between frames.

Data-pruning. The datasets contain samples that cannot be used for cross-view matching if the vehicle is traveling through tunnels or under bridges or if the data is

Table 1. Recall in percent on a subset of the first two scenes of the Ford AV dataset [6] following the evaluation protocol introduced by Shi et al. [35]. Results of previous works on Ford AV are provided by Shi et al. [35]. The first three rows represent large-area CVGL methods that are adapted for metric CVGL and have no dedicated method for predicting metric offsets. Initial pose is chosen in a 40m area around the vehicle with up to 20° of rotation noise. All methods are vision-based only.

	Cross-area	Cross-vehicle	Log1			Longitudinal			Log2			Longitudinal		
			1.0m	3.0m	5.0m	1.0m	3.0m	5.0m	1.0m	3.0m	5.0m	1.0m	3.0m	5.0m
CVM-Net [19]	7	7	9.1	25.7	41.3	4.8	13.2	21.9	9.8	28.6	47.1	4.2	11.8	20.3
SAFA [36]	7	7	9.3	28.7	48.0	4.3	11.8	20.1	11.2	34.1	53.4	5.0	13.4	22.9
DSM [37]	7	7	12.0	35.3	53.7	4.3	12.5	21.4	8.5	24.9	37.6	3.9	12.2	21.4
VIGOR [58]	7	7	20.3	52.5	70.4	6.2	16.1	25.8	20.9	54.9	75.7	6.0	16.9	27.0
HighlyAccurate [35]	7	7	46.1	70.4	72.9	5.3	16.4	26.9	31.2	66.5	78.8	4.8	15.3	25.8
Ours w/o vehicle frames	7	-	15.1	51.3	72.0	5.0	15.2	24.4	11.3	37.8	62.2	4.7	15.3	26.0
Ours	7	7	96.3	99.6	99.6	76.0	95.3	96.0	88.0	99.9	100.0	58.9	93.3	93.6
Ours	3	3	77.0	96.2	97.6	24.0	67.6	76.1	73.0	96.5	97.8	25.6	61.7	69.4

out-of-date and does not correspond with recent aerial images. We design a simple data-pruning approach to remove these samples from the datasets as follows.

We measure the difficulty of each data-frame by processing it with the pseudo-labeling model and computing the generalized variance of the predicted probability distribution (i.e. the determinant of the covariance matrix). Easy and hard samples correspond with low and high predicted variance in the lateral and longitudinal directions. We sort all data-frames by their difficulty and remove the hardest 1%. Fig. 5 shows examples of pruned frames on the Ford AV dataset. Our evaluation in Sec. 5.4 justifies this approach.

5. Evaluation

5.1. Implementation details

We use the ConvNeXt base [25] model in the encoder for both aerial and ground images. We train for 100K iterations with the RectifiedAdam optimizer [22], a batch size of 1 and a learning rate of $1 \cdot 10^{-4}$ with polynomial decay. The loss function is parametrized with $t = 0.5m$ and $\alpha = 2$.

For the cross-attention block, we use $n_{heads} = 4$ heads, sample point-pillars with $z = 16$ points from $h_{min} = 5m$ to $h_{max} = 10m$, and encode the PV features at stride = 4. We use a spatial reduction rates of $s_r = 4$ in the self-attention block. Overall, $n_{blocks} = 3$ refinement steps are applied to compute the BEV with size $d_B = 320px$ at $q_B = 2.4 \frac{m}{px}$ and $c_B = 128$ channels. The matching is performed at resolution $q_A = 0.3 \frac{m}{px}$ with an aerial image of size $d_A = 512px$ and $c_A = 8$ channels.

5.2. Per-frame evaluation

We evaluate the per-frame performance of our model on the Ford AV dataset [6] and show results in Table 1. We follow the protocol introduced by Shi et al. [35] of testing only on a subset of the first two scenes to compare our method with related approaches. The apriori pose is chosen randomly in a 40m × 40m area around the ground truth position with up to 20° of rotation noise. Since our method works with a circular search region, we choose the smallest

(1) Same-area and same-vehicle We train on the same scenes of Ford AV that are captured at a different time than the test split. Our model vastly outperforms previous approaches and successfully localizes > 90% of the frames within 3m to the ground-truth position.

(2) Cross-area and cross-vehicle We train the model on Argoverse V1, Argoverse V2, Lyft L5, Nuscenes and Pandaset, but remove data from Detroit where Ford AV was recorded. During training, the model therefore does not have access to data from either the test region (Detroit) or the test vehicle and corresponding camera setup (Ford AV vehicle). Our model still outperforms previous approaches trained in the same-area and same-vehicle setting by a large margin.

Fig. 1 shows several examples of probability distributions predicted by our model on the Ford AV dataset under cross-area and cross-vehicle conditions. With a search radius of 50m, we achieve a median position error over all six trajectories of 0.87m both when the orientation is known (up to 30° noise) and when no orientation information is given. We provide more details per trajectory in the supplementary material.

(3) Same-area and no-vehicle We additionally train our model without using information from the vehicle cameras, i.e. by setting all RGB values to zero. The model therefore only learns a prior distribution of vehicle poses w.r.t. the aerial image since the BEV map is constant over different inputs. The model shows a performance similar to the previous state-of-the-art HighlyAccurate [35] for longitudinal recall

indicating that their model might rely mainly on prior poses w.r.t. the aerial image rather than on cross-view matching (cf. Table 1). Fig. 6 shows a visualization of features learned by the corresponding model.

5.3. Tracking evaluation

We choose a tracking method using a Kalman filter to determine the trajectory of the vehicle over time [13]. Here,

Table 2. Absolute Position Error in meters on KITTI-360 scenes

Method	Camera	Lidar	00	02	03	04	05	06	07	09	10	Mean
Ferverset al. [13]	3	3	0.70	0.94	0.67	0.95	0.75	1.16	0.99	0.75	2.16	0.94
Ours	3	7	0.62	0.80	1.01	0.71	0.62	0.80	0.60	0.67	2.12	0.78

Table 3. Ablation studies tested on all scenes from Palo Alto and San Francisco. The initial pose is chosen randomly up to 30m from the vehicle with up to 0 of rotation noise. ME: Mean error in meters. RMSE: Root mean squared error in meters.

Method modification	ME	RMSE
–	1.19	3.44
No deformable offsets	1.22	3.65
No ResNet blocks at stride 1	1.22	3.51
$n_{heads} = 1$	1.23	3.59
No deformable offset skip connection	1.23	3.69
No data pruning	1.23	3.62
No MLPs in transformer blocks	1.23	3.52
$n_{blocks} = 1$	1.24	3.76
No attention skip connection	1.24	3.77
AV ! QKV attention	1.25	3.63
No Self Attention Block	1.38	4.21
No pseudo-labels	2.37	5.15
No BEV upsampling	2.63	5.87
No encoder pretraining	4.15	9.21
No vehicle images	11.95	15.20

(a) Model predictions with known orientation. Driving direction points upwards. In the first three examples the model exploits its learned knowledge of right-hand side traffic in the training data.

(b) Model predictions with unknown orientation.

Figure 6. Aerial features predicted by a model that was trained in a cross-area setting without ground cameras and therefore learns only a prior distribution of vehicle poses w.r.t. the aerial image. Features are reduced to three channels via principal component analysis and mapped onto RGB for visualization. Vehicle data: Lyft L5 dataset [18]. Map data: Bing Maps 2022 TomTom© Vexcel Imaging [1].

an inertial measurement unit is used to produce accurate short-term trajectories while the predictions of our model keep the position in alignment with aerial images in the long term. The tracker particularly benefits from the model's uncertainty estimates which are fed into the Kalman filter and propagated over time.

We test our model with this tracking method on the KITTI-360 [21] dataset and train in a cross-area and cross-vehicle setup on all other datasets. To ensure fair comparison with related works, rather than measuring the error w.r.t. our pseudo-labels we choose an evaluation approach which is typically used for odometry methods and is able to compensate for relative offsets between ground truth geo-locations and aerial images [54]. We align the predicted trajectories with the original ground truth of KITTI-360 via a 3-DoF rigid transformation and measure the relative deviation of the transformed trajectories to the ground truth. As shown in Table 2, this method has a mean position error of 0.78m over all scenes which surpasses even a recent lidar-visual based work.

The error w.r.t. our pseudo-labels (without prior alignment) is 0.85m which supports the quality of our pseudo-labeled ground truth. We further provide two videos for the tracking results of scenes from KITTI-360 and Ford AV in the supplementary material for a qualitative evaluation.

5.4. Ablation studies

For the ablation studies, we choose a smaller model with the nano variant of ConvNeXt [44] as encoder, a BEV map with size $d_B = 192px$ at $c_B = 2:0 \frac{m}{px}$ and $c_B = 32$ channels, and aerial features with size $d_A = 256px$ at $c_A = 0:5 \frac{m}{px}$. We use all scenes from Palo Alto and San Francisco as test split and train on the rest excluding KITTI-360 which does not contain full surround view with cameras.

We remove the individual components of our method listed in Table 3 and report the corresponding test scores to evaluate their effect on the localization accuracy. All components improve the performance of the model which supports their motivation in Sec. 3 and Sec. 4.

6. Conclusion

We present a novel method for vision-based cross-view geolocalization that allows localizing a vehicle on an aerial image with high metric accuracy. To evaluate the method in cross-area and cross-vehicle settings, we combine multiple vehicle datasets with aerial images from several orthophoto providers to train and test our method. We implement a pseudo-label approach to improve the inaccurate ground truth poses of these datasets, and make the improved ground truth publicly available. Our method outperforms previous approaches by a large margin even under more challenging cross-area and cross-vehicle conditions. We further show that a standard tracking framework is capable of exploiting the soft probability distributions predicted by our model to determine the vehicle's trajectory over time with sub-meter accurate poses.

References

- [1] Bing maps. <https://docs.microsoft.com/en-us/bingmaps> . 1, 2, 6, 8
- [2] Dcgis. <https://octo.dc.gov/service/dc-gis-services> . 6
- [3] Google maps <https://developers.google.com/maps/documentation> . 2, 6
- [4] Massgis. <https://www.mass.gov/orgs/massgis-bureau-of-geographic-information> . 6
- [5] Stratmap. <https://tnris.org/stratmap/> . 6
- [6] Siddharth Agarwal, Ankit Vora, Gaurav Pandey, Wayne Williams, Helen Kourous, and James McBride. Ford multi-av seasonal dataset. *International Journal of Robotics Research* 2020. 1, 2, 6, 7
- [7] Syed Ammar Abbas and Andrew Zisserman. A geometric approach to obtain a bird's eye view from an image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops* pages 0–0, 2019. 3
- [8] L Susan Blackford, Antoine Petitot, Roldan Pozo, Karin Remington, R Clint Whaley, James Demmel, Jack Dongarra, Iain Duff, Sven Hammarling, Greg Henry, et al. An updated set of basic linear algebra subprograms (blas). *ACM Transactions on Mathematical Software* 28(2):135–151, 2002. 5
- [9] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. *Conference on Computer Vision and Pattern Recognition* 2020. 6
- [10] Yigit Baran Can, Alexander Liniger, Danda Pani Paudel, and Luc Van Gool. Structured bird's-eye-view traffic scene understanding from onboard images. *Proceedings of the IEEE/CVF International Conference on Computer Vision* pages 15661–15670, 2021. 3
- [11] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, et al. Argoverse: 3d tracking and forecasting with rich maps. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pages 8748–8757, 2019. 6
- [12] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence* 40(4):834–848, 2017. 4
- [13] Florian Fervers, Sebastian Bullinger, Christoph Bodensteiner, Michael Arens, and Rainer Stiefelhagen. Continuous self-localization on aerial images using visual and lidar sensors. *arXiv preprint arXiv:2203.03334* 2022. 2, 3, 6, 7, 8
- [14] Giorgio Grisetti, Rainer Kümmerle, Hauke Strasdat, and Kurt Konolige. g2o: A general framework for (hyper) graph optimization. In *Proceedings of the IEEE international conference on robotics and automation (ICRA)*, Shanghai, China pages 9–13, 2011. 6
- [15] Adam W Harley, Zhaoyuan Fang, Jie Li, Rares Ambrus, and Katerina Fragkiadaki. A simple baseline for bev perception without lidar. *arXiv preprint arXiv:2206.07959* 2022. 2
- [16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition* pages 770–778, 2016. 4
- [17] Ruining He, Anirudh Ravula, Bhargav Kanagal, and Joshua Ainslie. Reformer: Transformer likes residual attention. *arXiv preprint arXiv:2012.11774* 2020. 5
- [18] John Houston, Guido Zuidhof, Luca Bergamini, Yawei Ye, Long Chen, Ashesh Jain, Sammy Omari, Vladimir Iglovikov, and Peter Ondruska. One thousand and one hours: Self-driving motion prediction dataset. *arXiv preprint* 2020. 6, 8
- [19] Sixing Hu, Mengdan Feng, Rang MH Nguyen, and Gim Hee Lee. Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* pages 7258–7267, 2018. 7
- [20] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. *arXiv preprint arXiv:2203.17270* 2022. 3, 4
- [21] Yiyi Liao, Jun Xie, and Andreas Geiger. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *arXiv preprint* 2021. 2, 6, 8
- [22] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. *arXiv preprint arXiv:1908.03265* 2019. 7
- [23] Liu Liu and Hongdong Li. Lending orientation to neural networks for cross-view geo-localization. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pages 5624–5633, 2019. 2, 6
- [24] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d object detection. *arXiv preprint arXiv:2203.05625* 2022. 3
- [25] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pages 11976–11986, 2022. 4, 7
- [26] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. *arXiv preprint arXiv:2205.13542* 2022. 2
- [27] Yuexin Ma, Tai Wang, Xuyang Bai, Huitong Yang, Yuenan Hou, Yaming Wang, Yu Qiao, Ruigang Yang, Dinesh Manocha, and Xinge Zhu. Vision-centric bev perception: A survey. *arXiv preprint arXiv:2208.02797* 2022. 3
- [28] Hanspeter A Mallot, Heinrich H Bülthoff, JJ Little, and Stefan Bohrer. Inverse perspective mapping simplifies optical flow computation and obstacle detection. *Biological cybernetics* 64(3):177–185, 1991. 3
- [29] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? *Advances in neural information processing systems* 32, 2019. 6

- [30] Lang Peng, Zhirong Chen, Zhangjie Fu, Pengpeng Liang, and Erkang Cheng. Bevsegformer: Bird's eye view semantic segmentation from arbitrary camera rigs. *arXiv preprint arXiv:2203.04050* 2022. 3
- [31] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *European Conference on Computer Vision* pages 194–210. Springer, 2020. 3
- [32] Cody Reading, Ali Harakeh, Julia Chae, and Steven L Waslander. Categorical depth distribution network for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pages 8555–8564, 2021. 3
- [33] Lennart Reiher, Bastian Lampe, and Lutz Eckstein. A sim2real deep learning approach for the transformation of images from multiple vehicle-mounted cameras to a semantically segmented image in bird's eye view. *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)* pages 1–7. IEEE, 2020. 3
- [34] Dávid Rozenberszki and András L Majdik. Lol: Lidar-only odometry and localization in 3d point cloud maps. *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4379–4385. IEEE, 2020. 2
- [35] Yujiao Shi and Hongdong 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* pages 17010–17020, 2022. 2, 3, 6, 7
- [36] Yujiao Shi, Liu Liu, Xin Yu, and Hongdong Li. Spatial-aware feature aggregation for image based cross-view geolocalization. *Advances in Neural Information Processing Systems* 32, 2019. 7
- [37] Yujiao Shi, Xin Yu, Dylan Campbell, and Hongdong 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* pages 4064–4072, 2020. 7
- [38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems* 30, 2017. 4, 5
- [39] Shan Wang, Yanhao Zhang, and Hongdong Li. Satellite image based cross-view localization for autonomous vehicles. *arXiv preprint arXiv:2207.13506* 2022. 6
- [40] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* pages 568–578, 2021. 5
- [41] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pages 8445–8453, 2019. 3
- [42] Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *Conference on Robot Learning* pages 180–191. PMLR, 2022. 3
- [43] Eric W. Weisstein. Convolution theorem. From MathWorld—A Wolfram Web Resource. <http://mathworld.wolfram.com/ConvolutionTheorem.html>. 5
- [44] Ross Wightman. Pytorch image models. <https://github.com/rwightman/pytorch-image-models>, 2019. 8
- [45] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. *Thirty-first Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)* 2021. 6
- [46] Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-area image geolocalization with aerial reference imagery. In *International Conference on Computer Vision* 2015. 2, 6
- [47] Zimin Xia, Olaf Booij, Marco Manfredi, and Julian FP Kooij. Cross-view matching for vehicle localization by learning geographically local representation. *IEEE Robotics and Automation Letters* 6(3):5921–5928, 2021. 3
- [48] Zimin Xia, Olaf Booij, Marco Manfredi, and Julian FP Kooij. Visual cross-view metric localization with dense uncertainty estimates. *arXiv preprint arXiv:2208.08519* 2022. 3
- [49] Pengchuan Xiao, Zhenlei Shao, Steven Hao, Zishuo Zhang, Xiaolin Chai, Judy Jiao, Zesong Li, Jian Wu, Kai Sun, Kun Jiang, et al. Pandaset: Advanced sensor suite dataset for autonomous driving. In *International Intelligent Transportation Systems Conference* 2021. 6
- [50] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems* 34, 2021. 5
- [51] Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tiejian Liu. On layer normalization in the transformer architecture. In *International Conference on Machine Learning* pages 10524–10533. PMLR, 2020. 4
- [52] Dongqiangzi Ye, Zixiang Zhou, Weijia Chen, Yufei Xie, Yu Wang, Panqu Wang, and Hassan Foroosh. Lidarmultinet: Towards a unified multi-task network for lidar perception. *arXiv preprint arXiv:2209.09385* 2022. 2
- [53] Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pages 10819–10829, 2022. 4
- [54] Zichao Zhang and Davide Scaramuzza. A tutorial on quantitative trajectory evaluation for visual(-inertial) odometry. In *IEEE/RSJ Int. Conf. Intell. Robot. Syst. (IROS)* 2018. 8
- [55] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* pages 2881–2890, 2017. 4

- [56] Brady Zhou and Philipp Kehl. Cross-view transformers for real-time map-view semantic segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13760–13769, 2022. 3
- [57] Sijie Zhu, Mubarak Shah, and Chen 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*, pages 1162–1171, 2022. 2, 3
- [58] Sijie Zhu, Taojiannan Yang, and Chen Chen. Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3640–3649, 2021. 3, 7
- [59] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 5