2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

BEV-CV: Birds-Eye-View Transform for Cross-View Geo-Localisation

Tavis Shore[1] and Simon Hadfield[1] and Oscar Mendez[1]

***Abstract*— Cross-view image matching for geo-localisation is a challenging problem due to the significant visual difference between aerial and ground-level viewpoints. The method provides localisation capabilities from geo-referenced images, eliminating the need for external devices or costly equipment. This enhances the capacity of agents to autonomously determine their position, navigate, and operate effectively in GNSS-denied environments. Current research employs a variety of techniques to reduce the domain gap such as applying polar transforms to aerial images or synthesising between perspectives. However, these approaches generally rely on having a 360° field of view, limiting real-world feasibility. We propose BEV-CV, an approach introducing two key novelties with a focus on improving the real-world viability of cross-view geo-localisation. Firstly bringing ground-level images into a semantic Birds-Eye-View before matching embeddings, allowing for direct comparison with aerial image representations. Secondly, we adapt datasets into application realistic format - limited Field-of-View images aligned to vehicle direction. BEV-CV achieves state-of-the-art recall accuracies, improving Top-1 rates of 70° crops of CVUSA and CVACT by 23% and 24% respectively. Also decreasing computational requirements by reducing floating point operations to below previous works, and decreasing embedding dimensionality by 33% - together allowing for faster localisation capabilities.**

I. INTRODUCTION

Localisation is necessary for many robotics applications - from autonomous vehicles to driverless railways the ability to localise must be ingrained. The majority of current localisation techniques rely on external sensors supplying either positional context or a calculated location. This reliance on external devices such as Global Navigation Satellite Systems (GNSS) may lead to issues caused by occlusions or sensor errors which inhibit localisation. Similarly, LIDAR-based approaches are expensive and both power and data-hungry. Vision-based localisation offers a solution as cameras are low-cost and compact, enabling robotics to discover more information about their environment from which to self-localise. Moreover, most modern vehicles are equipped with forward-facing cameras, making the adoption of limited Field-of-View (FOV) Cross-View Geo-localisation (CVGL) straightforward.

CVGL aims to match ground-level perspective images to geo-referenced aerial images. Throughout this research, we refer to images taken from car-mounted front-facing limited-FOV cameras as Point-of-View (POV) images, and satellite or aerial images as aerial images. CVGL may offer a solution for self-contained localisation as a database of aerial feature embeddings is created locally and continually queried with POV feature embeddings, shown in Figure 1.

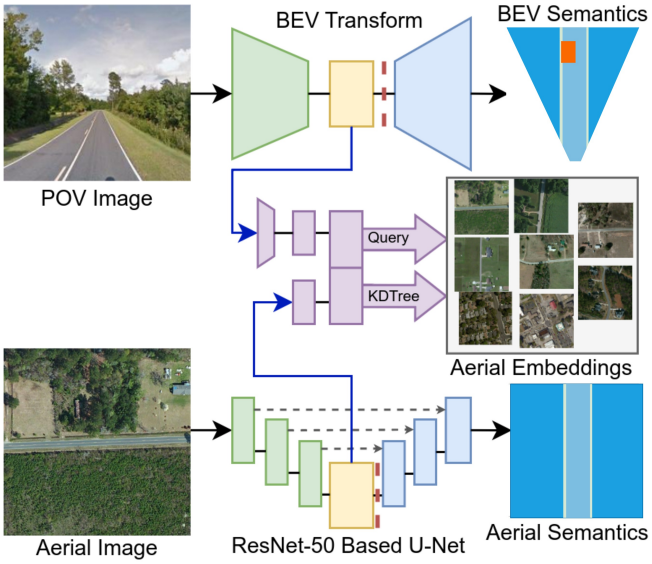

Fig. 1. General BEV-CV network structure. *POV Branch* extracts and transforms ground-level feature embeddings, *Map Branch* extracts aerial embeddings to build a KDTree. Components to the right of dotted red lines are discarded in the final BEV-CV architecture.

Our research develops a novel approach to reduce the domain disparity between POV and aerial images, improving CVGL performance while reducing computational requirements. We introduce BEV-CV, an architecture that reduces the domain gap between POV and aerial images by extracting semantic features at multiple resolutions before projecting them into a shared representation space and matching embedding pairs. We prioritise limiting computational requirements to ensure viability for mobile robotics applications.

In summary, technical contributions of our research are:

- Novel multi-branch architecture for extracting top-down representations from both viewpoints, projecting these into a shared representation space.
- Improving CVGL feasibility in two aspects: adjusting benchmark datasets to closer represent real-world application, and focusing on developing an efficient system capable of running in mobile systems - approximately reducing query times by 18% and embedding database memory requirements by 33%.
- Evaluation of our approach on CVUSA and CVACT datasets shows relative improvements upon previous state-of-the-art (SOTA) recall Top-1 rates by approximately 23% for CVUSA and 24% for CVACT with road-aligned 70° crops.

[1]Tavis Shore, Dr Simon Hadfield, and Dr Oscar Mendez are with the Centre for Vision, Speech, and Signal Processing (CVSSP), University of Surrey, Guildford, United Kingdom initial.surname@surrey.ac.uk

II. Related Works

Most existing works focus on retrieving embeddings from aerial feature databases. The field has recently started considering limited-FOV images due to their wider applicability. We describe previous works which led to the development of BEV-CV.

A. Cross-View Image Geo-Localisation

The predominant technique for CVGL is embedding retrieval. At an increasing rate, techniques are being proposed to reduce the domain gap and better match across viewpoints [1], [2], [3]. CVGL's deep learning era began in 2015 - Workman and Jacobs [4] proposed CNNs for extracting relevant features from both viewpoints for comparison, achieving good results and proving the potential of neural networks in CVGL. CNNs have since remained the predominant feature extraction mechanism. Lin et al. [5] considered each query image to have a unique identifier, using the cosine similarity distance as a similarity metric on clustered feature embeddings. Workman et al. [6] extend this, embedding representations from both viewpoints into a joint semantic representation space. Vo and Hays [7] incorporated aerial rotational information through an auxiliary loss function to observe the impact of image pair misalignment, also introducing a distance-based logistic loss to optimise performance. Limited-FOV images worsen the impact from misalignment as there is less shared information between views with which to correlate. CVM-Net [8] made a significant advancement, following a Siamese CNN with NetVLAD [9] - a network which aggregates residuals of local features to cluster centroids.

Toker et al. [10] use a multitask network that both synthesises streetview images from aerial image queries and performs image retrieval. Zhu et al. [11] employ more recent metric learning techniques and leverage activation maps to perform orientation estimation. L2LTR [12] proposes a CNN+transformer architecture, a ResNet feature extractor with vanilla ViT encoder. TransGeo [13] propose a transformer architecture that doesn't require data augmentation or alteration, using an attention-guided non-uniform cropping strategy to remove uninformative areas.

Sun et al. [14] propose a capsule network following a ResNet-based feature extractor, outperforming CVM-Net with the Vo and Hays dataset by approximately 10%. Liu and Li [15] improved the representation ability of their latent space by inserting orientation information. Shi et al. [16] apply a polar transform to aerial images and develop a spatial attention mechanism to improve feature alignment between views. Regmi et al. [17] created a conditional GAN to synthesise aerial representations of ground-level panoramas to reduce the domain gap. Shi et al. [18] propose CVFT to perform cross-view domain transfer to improve feature alignment between images. Their subsequent paper [2] proposed the new problem of applying CVGL techniques to limited FOV crops. This is important due to the ubiquity of monocular cameras compared to panoramic cameras; the ability for CVGL to perform with a limited-FOV is essential for wide-spread feasibility and adoption. To resolve the orientation ambiguity [2] computes the feature correlation between ground-level images and the corresponding polar-transformed aerial image, shifting or cropping the panorama at the strongest point before matching with their Dynamic Similarity Matching (DSM) module. However, it does so at the cost of introducing an expensive post-processing step at both training and inference time. In contrast, the BEV derived features in our own approach are discriminative enough to match uniquely over orientations with no correlation-based post-processing. In GeoDTR [19] Zhang et al. disentangle geometric information from raw features, learning spatial correlations among visual features to increase performance.

B. BEV Estimation

BEV transforms are a prominent research sub-field, mainly for autonomous road vehicles. We summarise relevant BEV papers for theoretical background as our proposed technique is the first to apply the transform to large-scale CVGL. Traditional techniques used camera intrinsics to geometrically transform POV images into the horizontal plane, but these performed poorly in some object detection cases due to extreme warping. Lu et al. [20] propose a variational encoder-decoder to encode ground-level images, decoding them into a semantic occupancy grid map. Schulter et al. [21] introduce a CNN to predict occluded areas of a scene layout, negating labelling requirements for these portions. Roddick and Cipolla [22] propose a BEV network that extracts image features at multiple resolutions before augmenting with spatial context and mapping to the BEV space. Saha et al. [23] use factorised 3D convolutions to estimate the BEV occupancy grid. Yang et al. [24] employ a cross-view transformation module, using correlations to cyclically strengthen the view transformation. Saha et al. [25] exploit the relationship between a vertical line in a POV image and a polar ray from the camera's perspective from the BEV viewpoint. Their subsequent research [26] proposes a graph neural network to learn the spatial relationship between objects in a scene and improve BEV object estimation.

To our knowledge, only one piece of research applies a BEV transform to the cross-image field - Fervers et al. [27] build a BEV representation of POV images with a cross-attention mechanism to determine a vehicle's pose. They use cross-attention to build output probabilities across a sequence of images, they could not evaluate with standard CVGL benchmark datasets as their technique requires lidar point clouds or rigid transformations between frames. BEV-CV is more applicable to wide-scale CVGL as point clouds are not required, only a single RGB perspective image.

III. Methodology

The network's objective is to minimise the domain gap apparent between aerial and POV image viewpoints to produce similar embeddings from both inputs. We bring both images into the top-down view, extracting and projecting features into a shared representation space. The network architecture

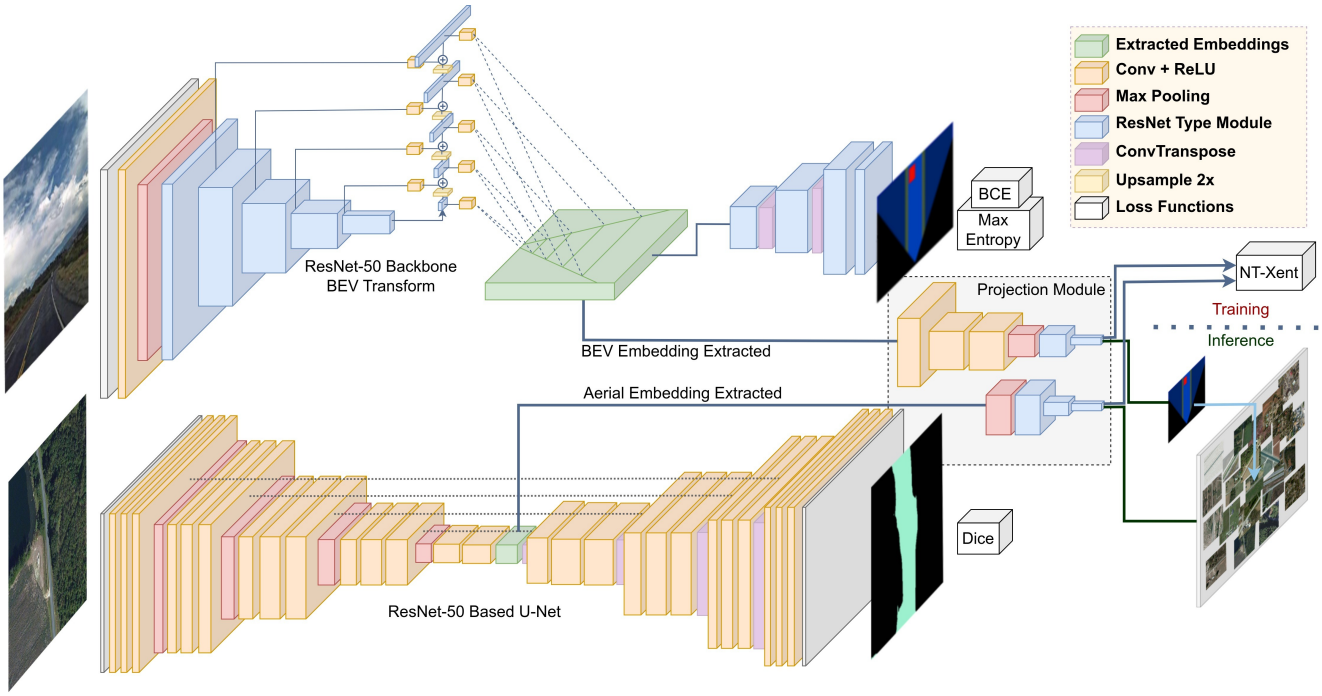

Fig. 2. BEV-CV network overview: *BEV Branch* is shown as the upper pathway, transforming from POV to BEV before extracting the embedding for projecting, the *Aerial Branch* is the lower pathway, extracting embeddings from the U-Net latent space. At training time we use an NT-Xent loss function and at inference time we build a KDTree of aerial embedding and query this with POV embeddings using descriptor cosine similarity for retrieval.

is shown in Figure 2, a two branch system with no weight sharing between branches.

A. Semantic Feature Extraction

To create a top-down view representation for the ground-level limited-FOV images we construct the *BEV Branch* of BEV-CV. This network contains 4 stages to extract and re-sample perspective information across views.

Both branches have an encoder-decoder structure to extract features that are reconstructed into semantic occupancy grids. Once trained on corresponding datasets, encoders are extracted and combined to form BEV-CV. We use semantic segmentation to compare embeddings as the classes layout within both images is similar, becoming comparable once transformed. Both POV and aerial input images are RGB: $I_t \in \mathbb{R}^{3\times W\times H}, t \in \{pov, aer\}$.

1) BEV Feature Extraction and Transform: For the BEV, the input ground-level panoramic images are cropped to the limited-FOV (examples in Figure 3) and resized to 224×224 where FOV, $\theta \in \{70°, 90°\}$, and yaw, $\Psi \in \{0°, ..., 360°\}$:

$$I_{pov} = \text{fov\_crop}\left(I_{pov}, \theta, \Psi\right) . \tag{1}$$

The images are fed through a ResNet-50 based network that extracts features at decreasing resolutions in order to retain information at different depths from the camera. These extractions are concatenated into a Feature Pyramid Network (FPN) which combines strong low-resolution semantics with weak high-resolution semantics through a top-down path and sequential lateral concatenations, shown as the *BEV Transform* in Figure 2. $\mathrm{f_i}$ denotes the i^{th} output for this n-layer FPN, concatenating ($\oplus$) outputs from the corresponding backbone layer with up-sampled lower-resolution features f_{i-1}:

$$\mathrm{f_i}(I_{pov}) = \mathrm{conv}(R_{n-i}(I_{pov})) \oplus \mathrm{u}(f_{i-1}(I_{pov})) \tag{2}$$

where R_{n-i} are ResNet layer outputs at increasing depth, producing n separate outputs as the feature pyramid.

FPN activation values at pixel locations are re-sampled into the BEV mapping using a multi-scale dense transform (MSD) with camera intrinsics, χ, to expand features in the z-dimension with Δx grid resolution, completing the conversion between the vertical to the horizontal plane. Re-sampling uses calibration intrinsics to determine the conversion of semantic information from object height to object depth. These transforms, MSD_i, condense features along the vertical dimension while maintaining horizontal resolution and expanding through depth, such that:

$$\eta_{pov} = \psi\left(\sum_{0}^{n} \mathrm{MSD}_i\left(f_i, \frac{\chi_f \Delta x}{2^{n+3}}, \chi\right)\right) \tag{3}$$

where χ_f is the focal length of the camera. In BEV-CV, the extracted orthographic features are passed through a sequence of convolutional layers ψ, to produce a compressed BEV embedding, $\eta_{pov} \in \mathbb{R}^{1\times512}$.

2) Aerial Feature Extraction: For the *Aerial Branch* of the network, we construct a ResNet-50 based U-Net. We select this structure as U-Nets [28] have become a predominant method for semantic segmentation, retaining information at decreasing resolutions to improve spatial context. The aerial branch is pre-trained with the full U-Net, using concatenation connections between the encoding and decoding stages.

After pre-training, only the encoder is used within BEV-CV. Similar to the BEV branch, aerial embeddings are extracted with a set of progressively encoded maps, $\mathrm{e}_{0\ldots n}$, with an equal number of decoders, $\mathrm{d}_{0\ldots n}$, each deconvolving from the previous decoder and concatenating extractions from the corresponding encoder.

$$\mathrm{d_i}(I_{aer}) = e_{n-i}(I_{aer}) \oplus \mathrm{deconv}(\mathrm{d}_{i-1}(I_{aer})) \tag{4}$$

where deconv represents $\mathrm{ConvTranspose2D}$. Unlike the BEV transform, this network only outputs from the final convolution module, d_n.

Detaching the encoders takes the outputs from the BEV after the multi-scale dense transforms and aerial outputs from the latent space of the U-Net. To complete the BEV feature extraction branch, this representation is compressed with further convolution layers to form practical embeddings - setting dimensionality with an optimal balance between discriminability and KDTree complexity, an important limitation when considering real-world deployment in mobile robotics. To transform both encoder outputs into a shared representation space with a standardised size we append a projection module. This contains fully connected layers, leaky ReLU functions, and batch normalisation. Aerial image embeddings $\left(\eta_{aer} \in \mathbb{R}^{1\times 512}\right)$ are taken from the output of the last U-Net encoding layer $\mathrm{e}_n\left(I_{aer}\right)$.

B. Normalised Temperature-scaled Cross Entropy Loss

The triplet loss is used throughout CVGL research to bring positive image pairs closer together and push negative pairs further apart in the representation space. Training BEV-CV with a triplet loss function yielded satisfactory results. We utilise a normalised temperature-scaled cross-entropy loss (NT-Xent) function [29] for the problem instead. NT-Xent takes the same inputs as triplet loss: perspective images along with corresponding positive and negative aerial image pairs. A variety of techniques are used for determining negative pairs, often depending on the initial L2 distance between embeddings. Hard triplet mining uses embeddings closer to the anchor than the positive for negatives, for semi-hard triplet mining the negative is not closer to the anchor than the positive is, but it still has a positive loss. We don't explicitly select triplets for training, instead using every other aerial image within a batch as negative examples, leading to each batch of size B having $B(B-1)$ negative examples.

The loss function for a batch of n CVGL embedding pairs $\left(\eta_{pov}^{0\ldots n}, \eta_{aer}^{0\ldots n}\right)$ is:

$$\mathcal{L}(\eta_{pov}^{i}, \eta_{aer}^{i}) = -log\frac{D(\eta_{pov}^{i}, \eta_{aer}^{i})}{\sum_{k=1,k\neq i}^{n} D(\eta_{pov}^{i}, \eta_{aer}^{k})}, \tag{5}$$

where D is the temperature (τ) normalised cosine similarity.

$$\mathcal{D}(i,j) = exp\left(\frac{i^T j}{\tau||i||||j||}\right) \tag{6}$$

The final loss is computed across all positive pairs in a batch.

Fig. 3. Panoramic examples of CVUSA and CVACT, heading aligned 90° FOV crops shown on the right hand side.

IV. Results

A. Datasets

The BEV transform and U-Net were pre-trained with individual datasets from each viewpoint. The BEV was trained with NuScenes [30], and the U-Net with *Massachusetts Road Dataset* [31] - consisting of aerial images and road semantic segmentation masks. Finally fine-tuning the constructed BEV-CV for CVGL.

We evaluate with CVGL benchmark datasets Crossview USA (CVUSA) [6] and CVACT [15]. These datasets contain 35,532 POV-Aerial training pairs and 8,884 testing pairs. CVUSA aerial images have a resolution of 750x750 and ground-level panoramas of 1232x224. CVACT aerial images have a resolution of 1200x1200 and ground-level panoramas of 1664x832, all north-aligned. We use the evaluation protocol implemented by [2], [15], and [18]. CVUSA contains yaw at image capture time, allowing for front-facing limited-FOV crops - the expected input for a real-world application of CVGL for autonomous vehicles. Unlike previous works, we do not perform pre-processing such as aerial polar transforms or panoramic shifting crops to match orientation between the viewpoints. We use a more realistic evaluation protocol where images are aligned to the heading of the car - achievable simply using a cheap compass sensor and the vehicle's yaw. Thus we avoid the expensive pre-processing steps used by DSM [2] and GAL [32] to observe orientation-aware performance. We evaluate previous works with original publicly available source code, and our direct image alignment protocol. CVACT does not contain the vehicle's yaw at image capture time. To evaluate with CVACT we first estimate the heading for each image pair using a semantic segmentation network [33] on the panoramic images. The yaw inaccuracies from this adversely affect results evaluated under the aligned protocol, however this would not be apparent in real-world application where vehicle-mounted cameras would be inherently aligned.

Config (BEV-CV minus)	R@1	R@5	R@10	R@1%
- BEV	0.80	3.11	5.53	21.35
- Projection	9.83	24.15	33.26	73.70
- U-Net	12.95	31.43	42.55	81.57
Full BEV-CV	**14.03**	**32.32**	**43.25**	81.48

TABLE I

ABLATION STUDY WITH CVUSA 70° CROPS.

Offset	R@1	R@5	R@10	R@1%
0°	**14.03**	**32.32**	**43.25**	**81.48**
±5°	11.89	26.74	36.37	74.41
±15°	9.79	23.31	31.86	71.07
±25°	8.81	20.14	28.13	66.19
±35°	2.80	10.82	17.54	51.77
±45°	1.11	5.36	9.70	39.18

TABLE II

RECALL ACCURACIES WITH OFFSET 70° POV CROPS

Fig. 4. BEV-CV CVUSA Top-5 recall examples. Outlines: Purple - query POV image, green - correct aerial image, red - incorrect aerial image

B. Training Details

Training occurs in two stages: pre-training the BEV & U-Net, and training the combined two-branch network. The BEV network is trained with NuScenes to take ground-level monocular images as input - outputting semantic map representations. The aerial branch's U-Net is trained with the Massachusetts dataset to extract road semantic features. The encoders from both networks are extracted, combining the branches to form a network trained with the evaluation datasets described in Section IV-A. BEV-CV is trained with triplets for 80 epochs using an Adam optimiser with an initial learning rate of 1e-4 and a *ReduceLROnPlateau* scheduler.

C. Implementation Details

We implement the architecture using PyTorch [34] with the Lightning framework [35], and style our BEV network from [22]. Once trained, we extract the encoder consisting of a ResNet-50 backbone whose outputs at multiple resolutions are concatenated into a 5-layer Feature Pyramid Network (FPN) before entering multi-scale dense transforms as BEV features with shape $[64, 100, 100]$. We append 3 *Conv-BatchNorm-LeakyReLU* sequences to compress feature extractions to shape $[512, 7, 7]$. We extract the U-Net encoder output, disregarding prior concatenation operations. The U-Net is a ResNet-50 based architecture with outputs features of shape $[2048, 7, 7]$. Feature extractions are projected into a shared representation space using a module that first applies a max pooling layer to flatten inputs. For the aerial branch, this is followed by a fully-connected layer to reduce the output size to 512. For both branches, the network ends with a single module of *BatchNorm-LeakyReLU-FC* which leaves the dimensions unchanged but in the shared latent space.

D. Evaluation

We use Top-K recall accuracy to evaluate, similar to previous works [2], [8], [18], [32]. Constructing a KDTree of aerial image embeddings, we retrieve the Top-K of these for a queried POV embedding determined by the cosine similarity between descriptors. A query is deemed successfully localised if the correct aerial image is within the top K retrievals. Top-K uses the absolute value of K for retrievals whereas Top-K% uses the K% length of the total dataset.

E. Ablation Study

To demonstrate the BEV module's effectiveness we ran an ablation study - replacing each branch with a vanilla ResNet-50, also removing the projection module, results are shown in Table I. This concluded that each constituent module within BEV-CV has a positive impact on its performance, with the BEV transform causing the largest increase in Top-K accuracy. To display the sensitivity of BEV-CV to POV road-alignment, we add an offset angle to the yaw before cropping panoramas, the experimental outcome is shown in Table II. Although BEV-CV performs well with misalignment, to operate optimally - POV images should be road-aligned.

F. Triplet vs NT-Xent Loss

Triplet loss has been widely used within computer vision research, the addition of a temperature parameter (τ) in NT-Xent to scale cosine similarities has been found to improve learning from hard negative examples [29]. Utilising the NT-Xent loss function yielded improvements across all Top-K recall accuracies, on average by 5%. With CVUSA 70° unaligned images, Top-1 and Top-1% increased from 11.30% to 14.03% and from 74.4% to 81.48%, respectively.

G. Prior Work Comparison

Recall rates are compared to previous SOTA techniques shown in Table III. We focus on orientation-aware evaluation, where images are yaw aligned, as it is the most applicable design for real-world deployment - including the orientation-unaware evaluations for comparison and to demonstrate how previous works are effected by the change in data representation. We achieve a 23% improvement in Top-1 rates with 70° CVUSA crops, and 24% improvement in Top-1 rates with 70° CVACT crops, proving the utility of BEV transforms for reducing the CVGL domain disparity.

Model	Orientation Aware	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%
		CVUSA 90°				CVUSA 70°			
CVM [8]	✗	2.76	10.11	16.74	55.49	2.62	9.30	15.06	21.77
CVFT [18]	✗	4.80	14.84	23.18	61.23	3.79	12.44	19.33	55.56
DSM [2]	✗	16.19	31.44	39.85	71.13	8.78	19.90	27.30	61.20
L2LTR [12]	✗	26.92	50.49	60.41	86.88	13.95	33.07	43.86	77.65
TransGeo [13]	✗	30.12	54.18	63.96	89.18	16.43	37.28	48.02	80.75
GeoDTR [19]	✗	18.81	43.36	57.94	88.14	14.84	38.03	51.27	88.17
BEV-CV	✗	15.17	33.91	45.33	82.53	14.03	32.32	43.25	81.48
GAL [32]	≈	22.54	44.36	54.17	84.59	15.20	32.86	42.06	75.21
DSM [2]	✓	**33.66**	51.70	59.68	82.46	20.88	36.99	44.70	71.10
L2LTR [12]	✓	25.21	*51.90*	*63.54*	*91.16*	*22.20*	*46.71*	*58.99*	*89.37*
TransGeo [13]	✓	21.96	45.35	56.49	86.80	17.27	38.95	49.44	81.34
GeoDTR [19]	✓	15.21	39.32	52.27	88.72	14.00	35.28	47.77	86.39
BEV-CV	✓	*32.11*	**58.36**	**69.06**	**92.99**	**27.40**	**52.94**	**64.47**	**90.94**
		CVACT 90°				CVACT 70°			
CVM [8]	✗	1.47	5.70	9.64	38.05	1.24	4.98	8.42	34.74
CVFT [18]	✗	1.85	6.28	10.54	39.25	1.49	5.13	8.19	34.59
DSM [2]	✗	18.11	33.34	40.94	68.65	8.29	20.72	27.13	57.08
L2LTR [12]	✗	13.07	30.38	41.00	76.07	6.67	15.94	23.45	49.37
TransGeo [13]	✗	10.75	28.22	37.51	70.15	7.01	19.44	27.50	62.19
GeoDTR [19]	✗	26.53	53.26	64.59	91.13	16.87	40.22	53.13	87.92
BEV-CV	✗	4.14	14.46	22.64	61.18	3.92	13.50	20.53	59.34
GAL [32]	≈	26.05	49.23	59.26	*85.60*	14.17	32.96	43.24	77.49
DSM [2]	✓	31.17	51.44	60.05	82.90	18.44	35.87	44.39	71.97
L2LTR [12]	✓	*33.62*	46.28	58.21	78.62	*28.65*	*53.59*	*65.02*	*90.48*
TransGeo [13]	✓	28.16	34.44	41.54	67.15	24.05	42.68	55.47	80.72
GeoDTR [19]	✓	26.76	*53.65*	*65.35*	*92.12*	15.38	37.09	49.40	86.38
BEV-CV	✓	**45.79**	**75.85**	**83.97**	**96.76**	**37.85**	**69.00**	**78.52**	**95.03**

TABLE III

BEV-CV EVALUATION AGAINST PREVIOUS WORKS - FOCUSING ON ROAD-ALIGNED ORIENTATION-AWARE ANALYSIS, BEST RESULTS ARE SHOWN IN **BOLD**, WITH SECOND BEST IN *Italic*. ≈ DENOTES WHERE CODE WAS UNAVAILABLE AND ORIENTATION WAS PARTLY UTILISED.

Model	Backbone	Dims ↓	Params (M) ↓	FLOPs (G) ↓
CVM [8]	VGG16	4096	160.3	-
CVFT [18]	VGG16	4096	26.8	-
DSM [2]	VGG16	4096	**14.5**	39.3
GAL [32]	ResNet-18	-	-	-
L2LTR [12]	HybridViT	768	195.9	57.1
TransGeo [13]	DeiT-S/16	1000	44.9	12.3
GeoDTR [19]	ResNet-34	4096	48.5	39.89
BEV-CV	ResNet-50	**512**	65.2	**11.5**

TABLE IV

COMPLEXITY COMPARISON FOR REAL-TIME APPLICATION FEASIBILITY.

H. Computational Efficiency

Comparing against these works, we also demonstrate computation improvements, shown in Table IV. Reducing dimensionality dramatically improves real-world feasibility as querying for Top-K retrievals in a KDTree takes at best $O(\sqrt{D}+k)$ time, where k is the number of retrievals, and D the dimensionality of the KDTree. Therefore lowering the dimensionality from 768-dims to 512-dims reduces Top-1 query retrieval time complexity and KDTree memory requirements by 18% and 33% respectively. We reduce floating point operations (FLOPs) by 6.5% compared with the next best previous work - further increasing feasibility for use in mobile robotics where computational capacity is limited and expensive.

V. CONCLUSION & FUTURE WORK

This paper introduced a novel technique to reduce the domain gap within limited-FOV CVGL, establishing the validity of BEV transforms in CVGL as a route to increase real-world feasibility. Evaluating against previous feature extraction approaches with road-aligned image queries, we improve upon nearly all recall accuracies for both CVUSA and CVACT, demonstrating strong potential for the field. BEV-CV achieves this while reducing the computational requirements for practical implementation - lowering both image embedding dimensionalities and retrieval time complexities. Our approach has some limitations; for instance, BEV transform specifications are set during training, determining parameters and transform shape within the BEV transform module according to the camera intrinsics. This limits the generalisability of the network for unseen image intrinsics at inference. Recent advances in BEV transforms have attempted to remove the need for explicit intrinsics, this will be a valuable addition to future BEV-CV techniques. However, we have shown that CNN-based BEV networks can be used as a drop-in replacement for the backbones commonly used in CVGL. Further work should aim to generalise BEV-CV across a wider varieties of regions, light, and weather conditions as current datasets were collected during daytime with clear weather from a small region.

VI. ACKNOWLEDGEMENTS

This work was partially funded by the EPSRC under grant agreement EP/S035761/1 and FlexBot - InnovateUK project 10067785.

References

[1] Yujiao Shi and Hongdong Li. "Beyond Cross-view Image Retrieval: Highly Accurate Vehicle Localization Using Satellite Image". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2022.

[2] Yujiao Shi et al. "Where Am I Looking At? Joint Location and Orientation Estimation by Cross-View Matching". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), pp. 4063–4071.

[3] Yujiao Shi et al. *CVLNet: Cross-View Semantic Correspondence Learning for Video-based Camera Localization*. 2022.

[4] Scott Workman and Nathan Jacobs. "On the location dependence of convolutional neural network features". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2015, pp. 70–78.

[5] Tsung-Yi Lin et al. "Learning deep representations for ground-to-aerial geolocalization". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 5007–5015.

[6] Scott Workman, Richard Souvenir, and Nathan Jacobs. "Wide-Area Image Geolocalization with Aerial Reference Imagery". In: *2015 IEEE International Conference on Computer Vision (ICCV)*. 2015, pp. 3961–3969.

[7] Nam N. Vo and James Hays. "Localizing and Orienting Street Views Using Overhead Imagery". In: *European Conference on Computer Vision*. 2016.

[8] Sixing Hu et al. "CVM-Net: Cross-View Matching Network for Image-Based Ground-to-Aerial Geo-Localization". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018, pp. 7258–7267.

[9] Relja Arandjelović et al. "NetVLAD: CNN Architecture for Weakly Supervised Place Recognition". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40 (2015), pp. 1437–1451.

[10] Aysim Toker et al. "Coming Down to Earth: Satellite-to-Street View Synthesis for Geo-Localization". In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 6484–6493.

[11] Sijie Zhu, Taojiannan Yang, and Chen Chen. "Revisiting Street-to-Aerial View Image Geo-localization and Orientation Estimation". In: *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)* (2020), pp. 756–765.

[12] Hongji Yang, Xiufan Lu, and Ying J. Zhu. "Cross-view Geo-localization with Layer-to-Layer Transformer". In: *Neural Information Processing Systems*. 2021.

[13] Sijie Zhu, Mubarak Shah, and Chen Chen. "TransGeo: Transformer Is All You Need for Cross-view Image Geo-localization". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 1152–1161.

[14] Bin Sun et al. "GEOCAPSNET: Ground to Aerial View Image Geo-Localization using Capsule Network". In: *2019 IEEE International Conference on Multimedia and Expo (ICME)* (2019), pp. 742–747.

[15] Liu Liu and Hongdong Li. "Lending Orientation to Neural Networks for Cross-View Geo-Localization". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 5617–5626.

[16] Yujiao Shi et al. "Spatial-Aware Feature Aggregation for Image based Cross-View Geo-Localization". In: *Neural Information Processing Systems*. 2019.

[17] Krishna Regmi and Mubarak Shah. "Bridging the Domain Gap for Ground-to-Aerial Image Matching". In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), pp. 470–479.

[18] Yujiao Shi et al. "Optimal Feature Transport for Cross-View Image Geo-Localization". In: *ArXiv* abs/1907.05021 (2019).

[19] Xiaohan Zhang et al. *Cross-view Geo-localization via Learning Disentangled Geometric Layout Correspondence*. 2023. arXiv: 2212.04074 [cs.CV].

[20] Chenyang Lu, M. J. G. van de Molengraft, and Gijs Dubbelman. "Monocular Semantic Occupancy Grid Mapping With Convolutional Variational Encoder–Decoder Networks". In: *IEEE Robotics and Automation Letters* 4 (2018), pp. 445–452.

[21] Samuel Schulter et al. "Learning to Look around Objects for Top-View Representations of Outdoor Scenes". In: *European Conference on Computer Vision*. 2018.

[22] Thomas Roddick and Roberto Cipolla. "Predicting Semantic Map Representations From Images Using Pyramid Occupancy Networks". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), pp. 11135–11144.

[23] Avishkar Saha et al. "Enabling spatio-temporal aggregation in Birds-Eye-View Vehicle Estimation". In: *2021 IEEE International Conference on Robotics and Automation (ICRA)* (2021), pp. 5133–5139.

[24] Weixiang Yang et al. "Projecting Your View Attentively: Monocular Road Scene Layout Estimation via Cross-view Transformation". In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 15531–15540.

[25] Avishkar Saha et al. "Translating Images into Maps". In: *2022 International Conference on Robotics and Automation (ICRA)* (2021), pp. 9200–9206.

[26] Avishkar Saha et al. "'The Pedestrian next to the Lamppost" Adaptive Object Graphs for Better Instantaneous Mapping". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 19506–19515.

[27] Florian Fervers et al. “Uncertainty-aware Vision-based Metric Cross-view Geolocalization”. In: *ArXiv* abs/2211.12145 (2022).

[28] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-Net: Convolutional Networks for Biomedical Image Segmentation”. In: *ArXiv* abs/1505.04597 (2015).

[29] Ting Chen et al. “A Simple Framework for Contrastive Learning of Visual Representations”. In: *arXiv preprint arXiv:2002.05709* (2020).

[30] Holger Caesar et al. “nuScenes: A multimodal dataset for autonomous driving”. In: *CVPR*. 2020.

[31] Volodymyr Mnih. “Machine Learning for Aerial Image Labeling”. PhD thesis. University of Toronto, 2013.

[32] Royston Rodrigues and Masahiro Tani. “Global Assists Local: Effective Aerial Representations for Field of View Constrained Image Geo-Localization”. In: *2022 IEEE Winter Conference on Applications of Computer Vision (WACV)* (2022).

[33] Bolei Zhou et al. “Semantic understanding of scenes through the ade20k dataset”. In: *International Journal on Computer Vision* (2018).

[34] Adam Paszke et al. “PyTorch: An Imperative Style, High-Performance Deep Learning Library”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019.

[35] William Falcon and The PyTorch Lightning team. *PyTorch Lightning*. Version 1.4. 2019.