RGB2LIDAR: Towards Solving Large-Scale Cross-Modal Visual Localization

Niluthpol Chowdhury Mithun, Karan Sikka, Han-Pang Chiu, Supun Samarasekera, Rakesh Kumar
Center for Vision Technologies, SRI International, Princeton, NJ
{niluthpol.mithun,karan.sikka,han-pang.chiu,supun.samarasekera,rakesh.kumar}@sri.com

ABSTRACT

We study an important, yet largely unexplored problem of large-scale cross-modal visual localization by matching ground RGB images to a geo-referenced aerial LIDAR 3D point cloud (rendered as depth images). Prior works were demonstrated on small datasets and did not lend themselves to scaling up for large-scale applications. To enable large-scale evaluation, we introduce a new dataset containing over 550K pairs (covering 143km^2 area) of RGB and aerial LIDAR depth images. We propose a novel joint embedding based method that effectively combines the appearance and semantic cues from both modalities to handle drastic cross-modal variations. Experiments on the proposed dataset show that our model achieves a strong result of a median rank of 5 in matching across a large test set of 50K location pairs collected from a 14km^2 area. This represents a significant advancement over prior works in performance and scale. We conclude with qualitative results to highlight the challenging nature of this task and the benefits of the proposed model. Our work provides a foundation for further research in cross-modal visual localization.

CCS CONCEPTS

• **Information systems** → **Multimedia and multimodal retrieval**; • **Computing methodologies** → *Matching*.

KEYWORDS

Large-scale Visual Localization, Cross-Modal Matching, Joint Embedding, RGB2LIDAR, Weak Cross-Modal Supervision

ACM Reference Format:
Niluthpol Chowdhury Mithun, Karan Sikka, Han-Pang Chiu, Supun Samarasekera, Rakesh Kumar. 2020. RGB2LIDAR: Towards Solving Large-Scale Cross-Modal Visual Localization. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3394171.3413647

1 INTRODUCTION

The real-world environment can be represented in many data modalities that are sensed by disparate sensing devices. For example, the same scene can be captured as an RGB or IR image, a depth map, or a set of 3D point clouds. Due to significant growth in the availability and diversity of sensors, the problem of matching data across heterogeneous modalities has gained noticeable momentum in recent years. Moreover, being able to match an image to different data modalities, whose acquisition is simpler/faster, opens more opportunities for visual localization (**VL**) across different autonomous platforms, *e.g.*, unmanned-ground vehicles.

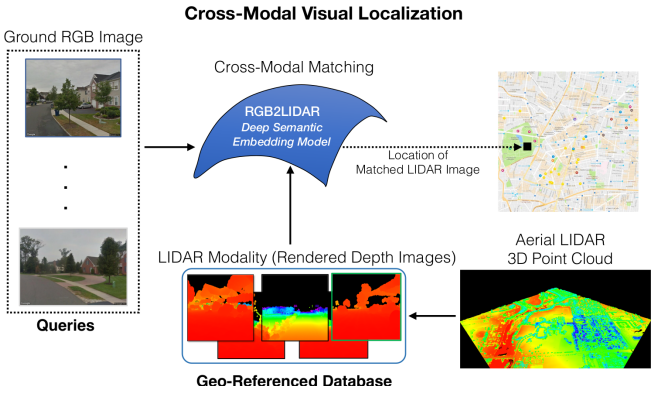

Figure 1: A brief illustration of our cross-modal geo-localization approach by matching a query ground RGB image to a geo-referenced aerial LIDAR 3D point cloud rendered as depth images.

In this paper, we target the problem of matching instances across diverse data modalities in a large-scale setting. Specifically, our goal is to localize a given street-level RGB image by querying a geo-referenced database from a different data modality, *i.e.*, LIDAR (Fig. 1). Prior works [8, 9, 50] are limited in tackling this task since they (1) were evaluated on small benchmark datasets and reported low accuracies, (2) heavily relied on hand-crafted features resulting in limited robustness to cross-modal appearance variations, and (3) operated mostly in specific settings such as urban areas. As a result, it has been difficult to establish the effectiveness of such works for real-world applications.

We address this problem by contributing a learning-based approach **RGB2LIDAR** and creating a large-scale dataset , referred to as **GRAL**, to evaluate large-scale cross-modal retrieval. GRAL consists of 550K location-coupled cross-modal pairs (covering 143km^2 area) collected automatically from Google Street View and USGS LIDAR data based on location coordinates. We choose 3D point cloud data from a LIDAR sensor for constructing the geo-referenced database. LIDAR, that directly senses the world in 3D, has recently become a popular data source for localization [33, 80, 87]. We particularly examine data acquired from an airborne LIDAR sensor. This problem is challenging compared to matching images across time (ground-to-ground) [12, 51, 55, 56, 74], or viewpoint changes in the

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413647

same modality (ground-to-aerial) [30, 46, 63, 73, 81]. Additionally, since LIDAR data points on vertical surfaces are often missing from aerial collections [50], the matching problem becomes more difficult. The benefit of using an aerial LIDAR collection system is that it covers a larger area much faster than the traditional time-consuming collections from the ground. Thus, matching technologies between ground RGB images and aerial LIDAR can benefit many VL applications where the geographical tags for images are unavailable, including historical images, self-driving cars, and images taken from GPS-denied or GPS-challenged environments. Instead of directly using 3D point clouds, we use LIDAR depth images generated from the projections of the point cloud data at a uniform grid of 2D locations on the ground plane. LIDAR data remains somewhat invariant to the projections in 2D and hence allows matching across modalities [50]. Moreover, it is easier to tackle large-scale visual localization in 2D as compared to 3D, since it is hard to work with 3D maps for large areas [66, 68].

The proposed approach (RGB2LIDAR VL) will enable autonomous vehicle-robot navigation. The traditional solution to this problem has been to use costly LIDAR sensors to localize the robot using geo-referenced LIDAR data in 3D world coordinates. The new trend is to use low-cost cameras (cf. LIDAR) to match with the geo-referenced 3D database, which requires 2.5D rendering from the 3D database due to the difficulty of direct 2D-3D matching [4, 14, 29]. In GPS-challenged environments, it typically involves two steps: (1) coarse search (or, geo-tagging [50]) of the 2D input image to find a set of candidate matches of 2.5D maps from the database, (2) fine alignment performs 2D-3D match verification for each candidate and returns the estimated 3D (6-DoF) geo-pose. As long as the candidate list includes the correct match, the fine alignment step can estimate accurate 3D geo-pose. Our work is the first to provide a large-scale cross-modal RGB to aerial LIDAR based solution for the first step (i.e., coarse search). The potential benefit from this application (such as 3D alignment) cannot be enabled from image-to-image (either ground-to-ground or ground-to-aerial) localization.

To the best of our knowledge, RGB2LIDAR is the first deep learning-based approach for large-scale cross-modal VL. The proposed method is based on utilizing multimodal deep convolutional neural networks (CNN) to learn joint representations for ground-level RGB images and aerial LIDAR depth images. This formulates cross-modal VL as a retrieval problem, that matches query RGB images to geo-referenced LIDAR depth images in the database (Fig. 1). Since the data from the two modalities exhibit large variations in appearances (Fig. 2), appearance information alone is not sufficient for identifying accurate correspondences. Semantic information can help significantly in this regard, as it focuses on the overall scene layout, and are generally more consistent as compared to appearance cues across heterogeneous data modalities (Fig. 3) [2, 61, 67, 68]. We use the representations generated from segmentation networks for LIDAR depth maps as semantic cues. However, it is difficult to obtain labels for training these models. We thus study and compare two complementary methods to tackle this issue. The first method is based on training a LIDAR segmentation network using cross-modal weak supervision from the segmentation maps of the paired RGB images from GRAL. The second involves training a segmentation network on a recently introduced DublinCity dataset that is the first semantic labeled dataset for dense aerial laser scanning [95]. To effectively fuse the appearance and semantic cues, we train multiple joint embedding models using different combinations of cues, and perform a weighted fusion.

Our experiments show that the proposed RGB2LIDAR model performs significantly better than prior works and baselines (Table 1). Our appearance-only joint embedding model achieves a median rank of 19 in matching across a large test set, whereas the random baseline is 24,921 (Table 2). The proposed fusion with semantic cues achieves median rank 5, a relative improvement of 73.6% over the appearance-only model. We also highlight the benefits of using weak cross-modal supervision for obtaining semantic cues. We believe that the underlying ideas and dataset developed in this work will open up avenues for further work in cross-modal VL.

Contributions: The main contributions of this work can be summarized as follows.

(1) We study an important, yet largely unexplored problem of large-scale cross-modal visual localization. Prior works were demonstrated on small datasets and did not lend themselves to scaling up for large-scale applications. We hope our work will encourage future work in cross-modal VL.
(2) We propose, to the best of our knowledge, the first deep learning-based method for cross-modal VL. The proposed RGB2LIDAR method is based on training joint representations and simultaneously utilizing appearance and semantic cues from cross-modal pairs.
(3) To enable large-scale evaluation for the task of cross-modal VL, we introduce a new large-scale dataset containing $550K$ location-coupled cross-modal pairs of ground RGB images and rendered depth images from aerial LIDAR point cloud covering around $143km^2$ area (cf. $5km^2$ from [50]).
(4) We compare two complementary approaches for training the semantic segmentation network (used to obtain semantic cues) for LIDAR depth images– first based on weakly supervised training of a LIDAR depth segmentation network. The second is based on training with full supervision from a dataset of limited diversity.
(5) We perform extensive experimental studies to establish the challenging nature of this problem and show the advantages of the proposed model compared to prior works. The proposed model achieves a strong result of a median rank 5 in matching across a large test set of $50K$ location pairs collected from a $14km^2$ area.

2 RELATED WORKS

Visual Localization: Vision-based methods generally localize [5, 28, 43, 55, 65, 67] or categorize [60, 64, 83] a query image based on a database of geo-referenced images or video streams [3, 6, 42]. Most prior works consider the problem of visual localization by matching location-coupled image database collected from the same viewpoint and sensor modality (Electrical-Optical Camera) as the query image. Significant work has been done in recent years in this direction utilizing hand-designed feature descriptors [18, 41], pre-trained CNN based features [13, 71, 72], and feature learning based approaches [55, 56, 67, 68, 94]. Although these methods show

good localization performance, their applications are limited by the difficulty in collecting reference ground images covering a large area. We refer our readers to [57], [47] and [24] for a comprehensive reviews on state-of-the-art approaches on vision-based localization.

Cross-View Visual Localization: To overcome the limitation of ground collections, several recent works have tried to localize ground-level image by matching against reference aerial imagery, which are easier to obtain [30, 45, 46, 63, 69, 73, 81]. However, these cross-view localization approaches suffer from significant perspective distortion due to the drastic viewpoint changes. The performance of these methods are also quite low when matching to single-view query images, and requires panoramic ground-view images as queries to achieve good accuracy.

Cross-Modal Visual Localization: Over the last decade, we have seen significant growth in the availability and diversity of sensors. Matching an image to different data modalities, that are simpler to be collected, opens more opportunities and possibilities for large-scale visual localization. However, this problem has been rarely investigated and has relied heavily on hand-crafted features [9, 50, 57, 76]. In contrast to prior works, we propose the first deep learning method to tackle this problem and demonstrate its feasibility for large-scale visual localization.

Image-to-LIDAR Visual Localization: Our data choice for cross-modal visual localization is matching ground RGB images to aerial geo-referenced LIDAR depth data [8, 50]. In terms of the experimental setup, our work is most closely related to [50]. However, [50] needs manually annotated building outlines for the query image during matching, which is not practical for large-scale applications. [50] also depends on local feature matching and urban scene to get distinctive features and is thus unlikely to work when paired images are not strongly aligned. A closely related work uses geometric point-ray features to compute candidate query poses without any appearance matching [8]. However, the method is also limited to urban settings and depends heavily on the availability of building corners in the image. Moreover, utilizing hand-crafted features limits the performance of these approaches [8, 50]. Additionally, they are evaluated on very few queries and the reported accuracy is also quite low.

Joint Embedding: Joint embedding models have shown excellent performance on several multimedia tasks, *e.g.*, cross-modal retrieval [10, 19, 31, 38, 53, 54, 75, 77, 82, 84], image captioning [35, 49], image classification [23, 25, 32] video summarization [15, 58], cross-view matching [81]. Cross-modal retrieval methods require computing similarity between two different modalities, *e.g.*, RGB and depth. Learning a joint representation naturally fits our task of cross-modal RGB-LIDAR retrieval since it is possible to directly compare RGB images and LIDAR depth images in such a joint space.

3 APPROACH

In this section, we describe the proposed GRAL dataset and the RGB2LIDAR approach.

3.1 GRAL Dataset

A major obstacle in exploring the large-scale cross-modal visual localization task is that none of the existing datasets are suitable for evaluation. To mitigate this issue, we create a new dataset which contains over $550K$ location-coupled pairs of ground RGB images and depth images collected from aerial LIDAR point clouds (Fig. 2). We shall refer to this dataset as the Ground RGB to Aerial LIDAR (GRAL) dataset. Although the primary purpose of this dataset is to evaluate cross-modal localization, it also allows us to evaluate matching under challenging cross-view setting. See the project page [1] for additional details.

3.1.1 ***Data Collection.*** We select a $143km^2$ area around Princeton, NJ, USA for data collection. We choose the area as it contains diverse urban, suburban and rural terrain characteristics. Our dataset contains a wide variety of scenes including forest, mountain, open country, highway, city interior, building, street etc. Note that, the LIDAR to image geo-localization approach in [50] also collected data from New Jersey but within a $5km^2$ area.

To ensure that each ground RGB image is paired with a single distinct depth image from aerial LIDAR, we create our dataset in two phases. First, we download available ground RGB images in the selected area for different GPS locations (latitude, longitude) using the Google Street View API. Second, we use LIDAR scan of the area from USGS to create a Digital Elevation Model (DEM) and from the DEM, location-coupled LIDAR depth images are collected for each street view images. For each location, we used 12 heading (0° to 360° at 30° intervals) for data collection.

Harvesting Ground RGB Images: We collect ground RGB images by densely sampling GPS locations. As the images are only available on streets from Google Street View, RGB imagery is not available in many locations in the selected area and Google returns a generic image for these locations. We use image metadata from the street view API to filter these images. We ensure that selected locations are 5 meters apart as we empirically found that spacing the samples around 5 meters apart tended to get a new image. We ultimately list about $60K$ GPS coordinates on the streets for data collection. We hard set the image pixel size (640×480), the horizontal field of view as 60 and pitch as 0 in the API.

Harvesting LIDAR Depth Images: We collect aerial LIDAR point-cloud of the selected area to create a DEM which is rendered exhaustively from multiple locations and viewpoints. For each GPS location containing RGB images, we render the LIDAR depth images from $1.7m$ above the ground. A digital surface model is used for height correction above sea level. We remove the depth images with no height correction and corresponding RGB images as we empirically found the viewpoint of these depth images are different from paired RGB images in most cases. We also remove the pairs where more than 60% pixels are black in the depth image.

Insufficient accuracy of instruments (GPS and IMU) and calibration issues may add some noise in data collection. Hence, there may exist some mis-alignment between location-coupled RGB and LIDAR depth images. To ensure good alignment between RGB and LIDAR depth images, we manually select a set of 100 locations on the LIDAR point cloud and corresponding points from Google Earth and then calculated the best fitting offset to bring the LIDAR point cloud closest to the points on Google Earth. The offset is ($2.77m$ easting, $0.18m$ northing) in UTM coordinates. We consider this offset when rendering LIDAR depth images.

[1] https://github.com/niluthpol/RGB2LIDAR

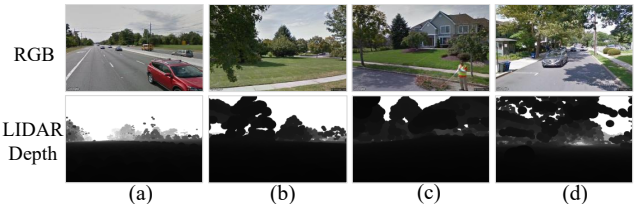

Figure 2: Example pairs of ground RGB images and aerial LIDAR depth images from the GRAL Dataset. RGB images are collected from Google Street View and depth images are collected by rendering aerial LIDAR point cloud from USGS.

3.1.2 ***Dataset Summary.*** Collecting readily available data from Google and USGS allowed us to create a large dataset without any manual effort. However, as the data collection process may introduce some mis-alignment in the dataset, we attempted to limit the issue by finding an offset based on a few manually selected points. The dataset still may encounter small horizontal and vertical mis-alignment in some cases (*e.g.*, (b) in Fig. 2). Moreover, as the LIDAR data collection is performed from an airborne sensor, the rendered depth images may miss many pixels (*e.g.*, (b) in Fig. 2). A change in the scene is also observed in some cases due to the time difference in the data collection of Street View and LIDAR. The dataset also exhibits challenges common in most visual localization datasets, such as dynamic aspects of environment (*e.g.*, (a) in Fig. 2).

To evaluate how our approach generalizes to unseen areas, we hold out a part of images for testing and another part for validation. Dataset split was done spatially. We use approximately 20% of the area for collecting validation images, 10% of the area for collecting testing images, and the rest for training. The collected dataset finally contains 557, 627 location-coupled pairs with 417, 998 pairs for training, 89, 787 pairs for validation and 49, 842 pairs for the testing. For each unique location, we initially collected 12 pairs based on different heading. However, after cleaning, the number of pairs for some locations may be less.

3.2 RGB2LIDAR

We consider that during training we are provided with paired examples of RGB images and LIDAR depth images captured from the same geo-location. During testing, we perform cross-modal retrieval to find the geo-location of a query RGB image by matching it with a geo-referenced database of LIDAR depth images (Fig. 1).

We first describe a general framework for cross-modal matching based on learning a joint multimodal embedding using appearance information (Sec. 3.2.1). This task is challenging since it involves matching examples across modalities exhibiting large disparities in appearance characteristics. As a result, appearance information alone is not sufficient for yielding high-quality matches. In comparison to appearance information, higher-level scene information is generally better preserved across inputs, from different visual sensors, capturing the same scene [2, 61, 67, 68]. We thus propose to use semantic information from intermediate feature maps generated from semantic segmentation networks for both modalities. We then discuss the proposed overall approach for fusing joint embeddings learned on different combination of appearance and semantic cues for improved cross-modal matching (Sec. 3.2.2). A key issue with using semantic information for LIDAR is that it is difficult to obtain or annotate semantic segmentation maps for LIDAR depth images. To tackle this issue, we explore two approaches in Sec. 3.2.3. The first is based on using weak cross-modal supervision from the RGB modality to train a segmentation network for the LIDAR depth images. The second approach is based on using representations from a segmentation network for LIDAR depth images trained on a different (smaller) dataset with ground-truth labels. The two approaches are complementary in that the first approach utilizes noisy annotations but has access to larger in-domain data. While the second approach uses a cleaner data from a small possibly out-of-domain dataset.

3.2.1 ***Training Joint Multimodal Embedding.*** We use a triplet ranking loss to embed both modalities in a common embedding space using the training pairs [21, 34, 37, 75, 91]. We denote the feature vector of RGB image and LIDAR depth image as $\boldsymbol{f_r}$ ($\in \mathbb{R}^I$) and $\boldsymbol{f_d}$ ($\in \mathbb{R}^L$) respectively. We use a linear projection to project both modalities in a common space– $\boldsymbol{r}_p = W^{(r)}\boldsymbol{f_r}$ ($\boldsymbol{r}_p \in \mathbb{R}^J$) and $\boldsymbol{d}_p = W^{(d)}\boldsymbol{f_d}$ ($\boldsymbol{d}_p \in \mathbb{R}^J$). Here, $W^{(r)} \in \mathbb{R}^{J\times I}$ and $W^{(d)} \in \mathbb{R}^{J\times L}$ project RGB and LIDAR depth maps to the joint space respectively. Using pairs of feature representation of RGB images and corresponding depth images, the goal is to learn a joint embedding such that the pairs from similar geo-locations, *i.e.*, positive pairs are closer that than negative pairs in the feature space. We achieve this by using a bi-directional triplet ranking loss as shown below:

$$\begin{aligned}\mathcal{L}_p = &\max_{\hat{\boldsymbol{d}}_p}[\Delta - S(\boldsymbol{r}_p, \boldsymbol{d}_p) + S(\boldsymbol{r}_p, \hat{\boldsymbol{d}}_p)]_+ \\ &+ \max_{\hat{\boldsymbol{r}}_p}[\Delta - S(\boldsymbol{r}_p, \boldsymbol{d}_p) + S(\hat{\boldsymbol{r}}_p, \boldsymbol{d}_p)]_+\end{aligned} \tag{1}$$

where $[x]_+ = \max(x, 0)$, $\mathcal{L}_p$ is the loss for a positive pair $(\boldsymbol{r}_p, \boldsymbol{d}_p)$, $\hat{\boldsymbol{r}}_p$ and $\hat{\boldsymbol{d}}_p$ are the negative samples for $\boldsymbol{r}_p$ and $\boldsymbol{d}_p$ respectively. Δ is the margin value for the loss. We use cosine similarity as the scoring function $S(\boldsymbol{r}_p, \boldsymbol{d}_p)$. We sample the negatives in a stochastic manner from each minibatch.

We utilize this framework to learn multimodal embeddings for different combinations of appearance and semantic features from the two modalities. We rely on using features from the semantic segmentation networks trained for both RGB and LIDAR depth images for semantic information. Although there are prior works on segmenting point clouds [27, 40, 79], they are not directly applicable to aerial LIDAR depth images in our setting. We later discuss two approaches for obtaining segmentations from LIDAR depth maps.

3.2.2 ***Combining Appearance and Semantic Cues.*** In order to exploit both the appearance and semantic information effectively, we construct our retrieval system based on a model ensemble [22, 59], where multiple expert models are generated strategically and combined to obtain a high-quality predictor. As the success of ensemble approaches depends significantly on the diversity of models inside the ensemble [59], it is important for us to choose a diverse set of joint embeddings. We propose to use both appearance and semantic features by training four joint embedding models utilizing different combinations of these features from the RGB and the LIDAR depth images. We train these models using the multimodal embedding framework discussed in Sec. 3.2.1.

At the time of retrieval, given a query ground image $\boldsymbol{r}$, the similarity score of the query is computed in each of the joint embedding spaces with LIDAR images $\boldsymbol{d}$ from the dataset. Our mixture of expert model uses a weighted fusion of scores for the final ranking.

$$S(\boldsymbol{r},\boldsymbol{d}) = \boldsymbol{w_1} S_{App-App}(\boldsymbol{r},\boldsymbol{d}) + \boldsymbol{w_2} S_{App-Sem}(\boldsymbol{r},\boldsymbol{d}) + \boldsymbol{w_3} S_{Sem-App}(\boldsymbol{r},\boldsymbol{d}) + \boldsymbol{w_4} S_{Sem-Sem}(\boldsymbol{r},\boldsymbol{d}) \quad (2)$$

where hyphen (-) symbol in subscript below S separates features from RGB and LIDAR depth used in learning joint the representations. For example, $S_{App-Sem}$ refers to the similarity score calculated in the joint space trained with appearance feature from RGB and semantic feature from LIDAR depth. The values of $\boldsymbol{w_1}, \boldsymbol{w_2}, \boldsymbol{w_3}, \boldsymbol{w_4}$ are chosen empirically based on the validation set.

In this work, we adopt a simple fusion strategy since our focus is to investigate the importance of utilizing multiple cues together for the task and whether it leads to a significant improvement over using only one of the cues. We later show in experiments that such a strategy helps in improving results highlighting the complementarity of the embedding models. We believe more sophisticated fusion strategies (*e.g.*, [48, 88]) will further improve the performance of our retrieval model. We leave this as future work.

3.2.3 ***Semantic Cues for LIDAR Depth Images.*** We use feature maps generated from pre-trained segmentation networks from each input modality as the semantic cues. For RGB modality, we use a pre-trained segmentation network [7], which works reasonably well. However, to the best of our knowledge, there exists no pre-trained segmentation network for aerial LIDAR point-clouds. Moreover, it is extremely difficult to obtain manual annotations for training segmentation networks on LIDAR depth images. We also observe that directly utilizing features from networks trained on RGB images for extracting semantic cues from LIDAR images and then training joint embeddings results in poor performance. We explore two complementary approaches to tackle this issue.

Weak Cross-Modal Supervision: The first method is inspired by the success of supervision transfer from paired multi-modal data in several vision and multimedia tasks [16, 26, 85]. Our method is motivated by two simple intuitions- (1) although there exist no pre-trained segmentation networks for aerial LIDAR depth images, there exist powerful segmentation networks for RGB images, and (2) the RGB and the LIDAR depth images are weakly aligned, and thus the RGB segmentation maps contain useful information about the general layout of the captured scene. We utilize the segmentation maps extracted from the paired RGB image as the ground-truth maps to train the segmentation network for LIDAR depth maps. We believe that due to the weak alignment between the modalities, the RGB segmentation maps contain sufficiently rich signals to train a reasonable segmentation network for the LIDAR depth images. Further, deep networks have been shown to be robust to noisy ground-truth and having some ability to self-correct their estimate of the noisy ground-truth in tasks such as weakly supervised segmentation and unsupervised machine translations [36, 39].

Full supervision from a smaller dataset: A possible drawback with the first approach is the use of noisy annotations from RGB modality, which are only weakly aligned with the LIDAR modality. This could result in a weak model that may not generalize well. We also explore a second approach where we use features from a segmentation network trained on a different dataset. We use the DublinCity dataset which contains labeled LIDAR point clouds collected around Dublin city centre and it is the first labeled dataset for a dense Aerial Laser Scanning [95]. Although the dataset provides clean labels for training, the coverage area is small ($2km^2$) and thus allows to collect a limited diversity of scenes. Hence, despite using cleaner labels the segmentation network is only trained with a smaller diversity of images and could contain out-of-domain examples, which might not transfer well.

We later show these complementary approaches are able to provide useful semantic cues for retrieval, highlighting the need for using alternate forms of supervision for cross-modal localization.

4 EXPERIMENTS

We now evaluate our method on the task of cross-modal VL by retrieving aerial LIDAR depth images by using the ground RGB images as queries. We first describe the evaluation metric and the implementation details. Then, we establish useful baselines on the proposed dataset and highlight the challenging nature of the task and the dataset. We position our work relative to the prior works on cross-modal VL through a discussion due to the difficulty of an empirical comparison since these methods used nonstandard datasets and pipelines. Next, we perform an analysis of our model to establish the benefits of the proposed fusion of appearance and semantic cues for cross-modal retrieval. We also evaluate different approaches for training the segmentation networks for LIDAR depth images and provide useful insights for future works. We finally present some qualitative results.

Evaluation Metric: We use standard evaluation metrics used in prior cross-modal retrieval tasks [20, 37, 52]. We report R@K (Recall at K) that calculates the percentage of queries for which the ground truth (GT) results are found within the top-K retrievals (higher is better). 5m R@1 calculates the percentage of queries for which the best matching sample is found within 5-meter distance to the GT. We also report MedR which calculates the median rank of the GT results in the retrieval (lower is better).

Implementation Details: The joint embedding models are trained using a two-branched neural network consisting of expert networks for encoding each modality and corresponding fully-connected layers to transform their outputs to a joint (aligned) representation ($J = 1024$) [37, 75]. We initialize appearance and semantic CNN models for both modalities using networks pre-trained for image classification and semantic segmentation respectively. We use an 18 layer wide-ResNet model [89] pre-trained on Places365 [92] as the RGB appearance CNN. We initialize the depth appearance model with pre-trained weights from the same appearance CNN, which we find works reasonably well in experiments. We use PSPNet model trained on ADE20K [93] as the semantic segmentation network for RGB [90]. We also attempted initializing depth semantic CNN with pre-trained RGB semantic CNN models. However, as the retrieval performance was quite low, we train LIDAR depth semantic segmentation models based on the approaches described in Sec. 3.2.3. We use SegNet model [7] trained on Cityscapes [17] to initialize the LIDAR depth segmentation network.

Table 1: This table compares our RGB2LIDAR model with hand-crafted and pre-trained CNN features based matching for cross-modal localization on the GRAL dataset. The results highlight the difficulty of the task, where hand-crafted features and deep features only perform slightly better than chance.

Method	R@1	R@5	R@10	MedR	MeanR
Chance	0.002	0.010	0.020	24921	24921
GIST	0.002	0.016	0.026	21101	22479
wide-ResNet18	0.014	0.059	0.114	19028	20131
ResNet50	0.007	0.031	0.067	19887	20328
MegaDepth	0.9	3.2	5.1	1735	5628
RGB2LIDAR	**27.6**	**51.1**	**57.9**	**5**	**34.5**

4.1 Baselines on GRAL dataset

Previous methods for this problem typically relied on hand-crafted features [8, 50]. On the other hand, pre-trained CNN features has been shown to be a strong baseline in many vision and multimedia tasks [62]. In Table 1, we compare RGB2LIDAR with hand-crafted GIST, pre-trained wide-ResNet18, and ResNet50 feature-based matching. Although directly utilizing hand-crafted or deep features for matching is unlikely to be effective in our setting due to viewpoint and modality differences, we provide these comparisons to understand the difficulty of the cross-modal (and cross-view) localization task on the proposed dataset. From Table 1, we observe that GIST (R@1 of 0.002%), wide-ResNet18 (0.014%) and ResNet50 (0.007%) perform slightly better than chance (0.002%), whereas the proposed RGB2LIDAR approach shows strong results (27.6%).

We also consider predicting depth cues from RGB images, and then performing retrieval based on the predicted depth. Specifically, we predict depth cues from RGB images using a single-view depth prediction model– MegaDepth [44] and train joint spaces with appearance cues from LIDAR depth images. This achieves a significantly lower performance compared to our RGB2LIDAR model (0.9% vs. 27.6% of ours in R@1).

4.2 Comparison with Prior Work

Ground RGB to Aerial Lidar based Retrieval: It is difficult to directly compare our approach with prior works on cross-modal localization [8, 50] since the query images and the exact location of LIDAR data are not available. Moreover, these methods use specific pre-processing which makes them unfit for large-scale applications as targeted in our work. For example, they are limited to an urban setting and localization performance depends on the availability of the buildings in the image. Moreover, [50] also requires manually annotating the building outlines of query images.

Bansal *et al.*[8] evaluated their approach on 50 Google Street View image queries and reported only 20% accuracy in $5m$ localization in the top-1000 ranks in $1Km \times 0.5Km$ area. On the other hand, our method shows 34% accuracy in $5m$ localization in top-1 rank based on testing across $50K$ pairs. Matei *et al.*[50] evaluated their approach on 14 queries in $5km^2$ area and reported R@1 of 7%, whereas our method shows R@1 of 27.6% based on $50K$ queries in $14km^2$ area. Hence, our method is likely to be more effective and generalizable than these methods [8, 50].

Table 2: Performance of the proposed RGB2LIDAR fusion of appearance and semantic cues for the cross-modal VL task on the GRAL dataset. We divide the table into two blocks to compare different RGB2LIDAR methods, i.e., joint embedding with different RGB and LIDAR features (2.1) and fusion of embeddings (2.2). A_R and S_R refers to appearance and semantic features from RGB images respectively. The (+) symbol denotes the ensemble of embeddings. The (-) symbol separates features from RGB images and LIDAR depth images, *e.g.*, A_R-S_L method refers to the embedding learned using appearance feature from RGB images and semantic feature from LIDAR depth images.

	Method	Evaluation Metric				
		R@1	R@5	R@10	MedR	5m R@1
	Chance	0.002	0.01	0.02	24921	0.003
2.1	A_R-A_L	20.3	39.0	45.1	19	26.4
	S_R-S_L	10.6	26.8	34.8	38	14.1
	A_R-S_L	9.5	24.3	32.0	48	12.6
	S_R-A_L	18.6	37.2	43.6	22	24.4
2.2	A_R-A_L+ S_R-A_L	24.8	45.5	51.8	9	31.5
	A_R-A_L+A_R-S_L	22.9	44.7	52.0	9	29.2
	A_R-A_L+S_R-A_L+A_R-S_L	27.0	49.5	56.2	6	34.0
	A_R-A_L+S_R-A_L+A_R-S_L+S_R-S_L (Proposed)	**27.6**	**51.1**	**57.9**	**5**	**34.5**

Ground-Aerial RGB based Retrieval: One of the motivations behind this work is that the performance of present ground-aerial methods is quite low in practice due to significant perspective distortions from viewpoint changes. Moreover, state-of-the-art ground-aerial methods require panorama image as queries for decent performance [30], whereas we use images with a 60 degree field of view in our evaluation. We perform comparison with a prominent cross-view localization model CVM-Net-I [30], by collecting ground panoramas and aerial satellite images for locations in GRAL. We follow ground-aerial image dataset CVUSA [81] data collection protocol, which was used to train CVM-Net-I. We find that CVM-Net-I model achieves low accuracy (i.e., R@1 =0.7%, R@5 =2.9%, R@10 = 5.1%) in RGB→Aerial-image based localization, whereas our RGB2LIDAR model achieves significantly better performance (i.e., R@1 =27.6%, R@5 =51.1%, R@10 = 57.9%) in RGB→LIDAR as shown in Table 1. Hence, we believe the proposed method is promising as a large-scale visual localization solution.

4.3 Analysis of the Proposed Method– RGB2LIDAR

We analyze the performance of our approach in Table 2 to show the benefits of the proposed ideas- (i) joint embeddings, (ii) semantic information, and (iii) overall approach using weighted ensemble-based fusion of multiple embeddings (RGB2LIDAR), for cross-modal retrieval. We use semantic features from the best performing segmentation model trained with weak cross-modal supervision (Table 3). We divide the table into two blocks (2.1-2.2) to aid our study.

Different Image Features for Joint Embedding: We first analyze the different combinations of appearance and semantic features from RGB and LIDAR depth images in training the joint embeddings in block-2.1. We observe that all four combinations perform reasonably well compared to the chance baseline. Utilizing the appearance features from both modalities seems to be performing the best. We

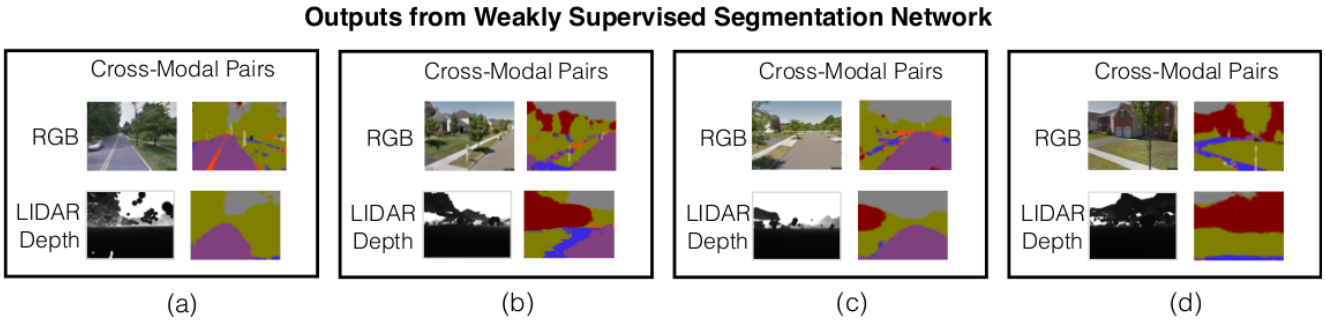

Figure 3: Sample semantic segmentation output for four cross-modal pairs. Here, pre-trained SegNet model is used for RGB image segmentation and segmentation network trained with weak cross-modal supervision is used for LIDAR depth image segmentation. The palette for semantic segmentation is as follows: sky, building, pole, road marking, road, pavement, tree. Best viewed in color.

note a drop in performance when using the semantic features, instead of the appearance features, in learning the joint embedding (R@1 is 20.3% for A_R-A_L, 9.5% for A_R-$S_L$and 18.6% for S_R-A_L). We expect a higher drop for the case when using semantic features for the LIDAR depth images since the semantic segmentation network for the LIDAR modality is trained with noisy labels. Moreover, the rendered depth images are generally misaligned relative to their RGB counterparts leading to further noise in the generated segmentations (Fig. 2). On the contrary, the drop in performance while using semantic features for the RGB modality is not significant (−1.7% absolute in R@1 for S_R-A_L).

Fusion of Appearance and Semantics: We propose a weighted fusion-based strategy to effectively combine the multiple joint embeddings (Sec. 3.2.2). The fusion weights (Eq. 2) were $\boldsymbol{w_1} = 0.33$, $\boldsymbol{w_2} = 0.16, \boldsymbol{w_3} = 0.32, \boldsymbol{w_4} = 0.19$ and chosen based on a validation set. The results from our weighted fusion-based approach, as shown in block-2.2, show large improvements over single joint embedding models (block-2.1). The proposed retrieval model utilizing both appearance and semantic cues achieves a 7.3% absolute improvement in R@1 compared to the best performing single cue embedding model A_R-A_L. We also observe improvements in fusion with the two joint embedding models trained using semantic cues from the LIDAR depth images (*i.e.*, A_R-S_L, S_R-S_L). For example, A_R-A_L+ A_R-$S_L$shows an absolute R@1 improvement of 2.6% over A_R-A_L. We also observe improvements when comparing A_R-A_L+ S_R-A_L to the proposed model (+2.8% absolute in R@1). We also verify that the improvements do not occur by chance by averaging performance across 12 randomized runs (R@1 of 27.37 ± 0.24 for proposed vs. 24.53 ± 0.19 when not using LIDAR semantic cues).

We also performed experiments with two typical early fusion techniques (i.e., concatenation, max). We find the performance of these baseline approaches are significantly lower (R@1 is 20.4 and 21.7 respectively) than our approach. We believe this happens since these early fusion strategies are unable to preserve the intra-modal invariances captured by appearance and semantic cues in the final (fused) representation.

4.4 Analysis of the LIDAR Depth Image Semantic Segmentation Network

We use feature maps from the outputs of LIDAR depth segmentation network for semantic features. Since there are no ground-truth annotations, we first followed the standard practice of initialization with a pre-trained segmentation network on RGB images [7, 78]. We observe in Table 3 that the retrieval performance of this model is quite low (1.9% in R@1) highlighting that the representations are not directly transferable owing to the large domain (modality) gap. We now empirically discuss the two complementary approaches introduced in Sec 3.2.3. The first approach (wSegNet-Depth) is trained using weak supervision from paired RGB images, while the second (SegNet-Depth) is trained using a labeled LIDAR depth point cloud dataset covering a smaller area.

Table 3: Cross-modal retrieval performance with different semantic feature extracted from different segmentation networks for LIDAR depth images in learning joint embeddings. The results show that the semantic segmentation network trained using weak cross-modal supervision performs better.

LIDAR Semantic Feature (Supervision - Dataset)	RGB Image Feature	Evaluation Metric			
		R@1	R@5	R@10	MedR
SegNet-RGB (Full - Cityscapes)	Appearance	1.9	3.9	5.8	1958
	Semantic	1.7	4.1	6.3	1716
SegNet-Depth (Full - DublinCity)	Appearance	8.8	22.0	30.7	52
	Semantic	9.0	23.4	30.9	53
wSegNet-Depth (Weak - GRAL)	Appearance	**9.5**	**24.3**	**32.0**	**48**
	Semantic	**10.6**	**26.8**	**34.8**	**38**

Table 3 shows that both the SegNet-Depth and wSegNet-Depth model significantly outperform the segmentation network initialized from pre-trained RGB segmentation network. We also observe that the SegNet-Depth model, despite being trained with noisy supervision, performs slightly better than the SegNet-Depth trained on DublinCity (*e.g.*10.6% vs. 9.0% R@1 with semantic features from RGB). We believe this happens since the SegNet-Depth is trained on out-of-domain images collected in a smaller area and is thus not able to generalize as well as the wSegNet-Depth, which is trained on in-domain examples covering a larger area. However, the performance gap is not huge given that SegNet-Depth is trained on a smaller area ($2km^2$ vs $101km^2$ in GRAL), which shows the advantages of using clean labeled data. Prior works on learning from noisy supervision have also shown that large quantities of noisy labels are required to match the performance of clean labels [1, 11, 70, 86]. We

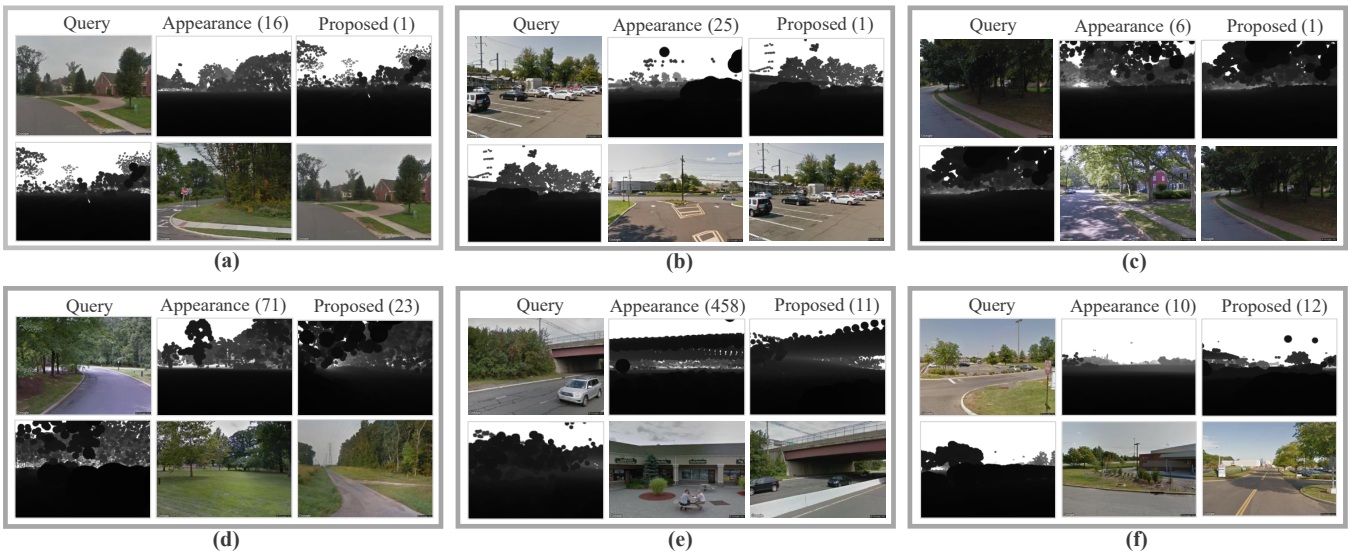

Figure 4: Showing six query RGB images and top-1 retrieved LIDAR depth image obtained using appearance-only embeddings and the proposed method (fusing appearance and semantic cues). We also show the corresponding cross-modal pairs below each image and the rank of the ground-truth depth image in brackets for each method. We observe both the proposed RGB2LIDAR method works well in retrieving the ground-truth match from a large set of about $50K$ candidates with good accuracy.

obtain further improvements by fine-tuning the weakly supervised model with clean labeled data (*e.g.*, R@1 of 28.2% compared 27.6% reported in Table 2, refer to supplementary material). We believe that although the performance of cross-modal VL will improve as more labeled LIDAR point clouds become available, there is potential in combining weak and full supervision for this task. We also hope that our findings will motivate future works in the exploration of freely available weak cross-modal supervision for handling multiple modalities for this task. We show segmentation outputs for four LIDAR depth images generated using the wSegNet-Depth segmentation network in Fig. 3.

4.5 Qualitative Results

In Fig. 4, we show six cases of query ground RGB images along with the top-1 retrieved LIDAR depth images using RGB2LIDAR and the appearance-only embedding baseline. These results highlight the challenging nature of this task especially the high visual similarity between different urban scenes, which necessitates the need for representations that are discriminative yet invariant to cross-modal appearance variations [67]. Overall, we observe both the proposed model and the appearance-only embedding baseline works well in retrieving the ground-truth match from a large set of about $50K$ candidates with good accuracy. However, our proposed fusion is able to consistently perform better in most cases. We expect this since cross-modal matching needs to deal with major changes in appearance across modalities, and semantic properties of a scene are generally more invariant to such factors. For example, in (e) the appearance-only model performs poorly in retrieving the ground truth as the appearance cues seem to confuse the bridge with the building outline, while the combined model is able to resolve this ambiguity. Since the segmentation of LIDAR depth images is not perfect, the proposed model could sometimes degrade retrieval performance, *e.g.*, (f).

5 CONCLUSION

We propose a new approach for large-scale cross-modal visual localization by matching query ground RGB images to a geo-referenced depth image database constructed from aerial LIDAR 3D point cloud. The proposed RGB2LIDAR approach is based upon a deep embedding based framework that capitalizes only on automatically collected cross-modal location-coupled pairs in training and simultaneously utilizes appearance and semantic information by using an ensemble approach for efficient retrieval. We introduce a new dataset GRAL to evaluate the large-scale cross-modal VL task. We also compare two complementary approaches, that do not require any labels from our dataset, for training the segmentation network (used to obtain semantic cues) for LIDAR depth images. The proposed approach is expected to generalize well to unseen locations based on the strong results of a median rank 5 in matching across a test set of about $50K$ location candidates covering an area of $14km^2$. The underlying ideas developed in this work can be applied to matching RGB images to other data choices for localization, *e.g.*, 3D point clouds from other sensors, CAD model. In the future, cross-modal visual localization could be a useful primitive across autonomous platforms. The proposed approach takes a crucial step towards realizing that vision.

ACKNOWLEDGMENTS

We would like to thank Ajay Divakaran for proof-reading the manuscript and providing many helpful comments. We would also like to acknowledge Zachary Seymour and Avi Ziskind for providing valuable suggestions and helping in preparing the dataset.

REFERENCES

[1] Jiwoon Ahn and Suha Kwak. 2018. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*. 4981–4990.

[2] Relja Arandjelović and Andrew Zisserman. 2014. Visual vocabulary with a semantic twist. In *Asian Conference on Computer Vision*. 178–195.

[3] Roberto Arroyo, Pablo F Alcantarilla, Luis M Bergasa, and Eduardo Romera. 2015. Towards life-long visual localization using an efficient matching of binary sequences from images. In *IEEE International Conference on Robotics and Automation*. 6328–6335.

[4] Clemens Arth, Christian Pirchheim, Jonathan Ventura, Dieter Schmalstieg, and Vincent Lepetit. 2015. Instant outdoor localization and slam initialization from 2.5 d maps. *IEEE Transactions on Visualization and Computer Graphics* 21, 11 (2015), 1309–1318.

[5] Georges Baatz, Kevin Köser, David Chen, Radek Grzeszczuk, and Marc Pollefeys. 2012. Leveraging 3d city models for rotation invariant place-of-interest recognition. *International Journal of Computer Vision* 96, 3 (2012), 315–334.

[6] Hernán Badino, Daniel Huber, and Takeo Kanade. 2012. Real-time topometric localization. In *International Conference on Robotics and Automation*. 1635–1642.

[7] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. 2017. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2017).

[8] Mayank Bansal and Kostas Daniilidis. 2014. Geometric urban geo-localization. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3978–3985.

[9] Francesco Castaldo, Amir Zamir, Roland Angst, Francesco Palmieri, and Silvio Savarese. 2015. Semantic cross-view matching. In *IEEE International Conference on Computer Vision Workshops*. 9–17.

[10] Hui Chen, Guiguang Ding, Zijia Lin, Sicheng Zhao, and Jungong Han. 2019. Cross-Modal Image-Text Retrieval with Semantic Consistency. In *ACM International Conference on Multimedia*. 1749–1757.

[11] Xinlei Chen and Abhinav Gupta. 2015. Webly supervised learning of convolutional networks. In *IEEE International Conference on Computer Vision*. 1431–1439.

[12] Zetao Chen, Adam Jacobson, Niko Sünderhauf, Ben Upcroft, Lingqiao Liu, Chunhua Shen, Ian Reid, and Michael Milford. 2017. Deep learning features at scale for visual place recognition. In *IEEE International Conference on Robotics and Automation*. 3223–3230.

[13] Zetao Chen, Obadiah Lam, Adam Jacobson, and Michael Milford. 2014. Convolutional neural network-based place recognition. *arXiv preprint arXiv:1411.1509* (2014).

[14] Han-Pang Chiu, Varun Murali, Ryan Villamil, G Drew Kessler, Supun Samarasekera, and Rakesh Kumar. 2018. Augmented Reality Driving Using Semantic Geo-Registration. In *IEEE Conference on Virtual Reality and 3D User Interfaces*. IEEE, 423–430.

[15] Jinsoo Choi, Tae-Hyun Oh, and In So Kweon. 2018. Contextually customized video summaries via natural language. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1718–1726.

[16] C Mario Christoudias, Raquel Urtasun, Mathieu Salzmann, and Trevor Darrell. 2010. Learning to recognize objects from unseen modalities. In *European Conference on Computer Vision*. Springer, 677–691.

[17] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[18] Mark J Cummins and Paul M Newman. 2010. Fab-map: Appearance-based place recognition and mapping using a learned visual vocabulary model. In *International Conference on Machine Learning*. 3–10.

[19] Samyak Datta, Karan Sikka, Anirban Roy, Karuna Ahuja, Devi Parikh, and Ajay Divakaran. 2019. Align2Ground: Weakly Supervised Phrase Grounding Guided by Image-Caption Alignment. In *IEEE International Conference on Computer Vision*.

[20] Jianfeng Dong, Xirong Li, and Cees GM Snoek. 2016. Word2VisualVec: Image and Video to Sentence Matching by Visual Feature Prediction. *arXiv preprint arXiv:1604.06838* (2016).

[21] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. VSE++: Improved Visual-Semantic Embeddings. In *British Machine Vision Conference*.

[22] Muhammad Moazam Fraz, Paolo Remagnino, Andreas Hoppe, Bunyarit Uyyanonvara, Alicja R Rudnicka, Christopher G Owen, and Sarah A Barman. 2012. An ensemble classification-based approach applied to retinal blood vessel segmentation. *IEEE Transactions on Biomedical Engineering* 59, 9 (2012), 2538–2548.

[23] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. 2013. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*. 2121–2129.

[24] Emilio Garcia-Fidalgo and Alberto Ortiz. 2015. Vision-based topological mapping and localization methods: A survey. *Robotics and Autonomous Systems* 64 (2015), 1–20.

[25] Yunchao Gong, Qifa Ke, Michael Isard, and Svetlana Lazebnik. 2014. A multi-view embedding space for modeling internet images, tags, and their semantics. *International journal of computer vision* 106, 2 (2014), 210–233.

[26] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. 2016. Cross modal distillation for supervision transfer. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2827–2836.

[27] Timo Hackel, Nikolay Savinov, Lubor Ladicky, Jan D Wegner, Konrad Schindler, and Marc Pollefeys. 2017. Semantic3d. net: A new large-scale point cloud classification benchmark. *arXiv preprint arXiv:1704.03847* (2017).

[28] James Hays and Alexei A Efros. 2008. IM2GPS: estimating geographic information from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1–8.

[29] Martin Hirzer, Clemens Arth, Peter M Roth, and Vincent Lepetit. 2017. Efficient 3D tracking in urban environments with semantic segmentation. In *British Machine Vision Conference 2017*.

[30] Sixing Hu, Mengdan Feng, Rang MH Nguyen, and G Hee Lee. 2018. CVM-Net: Cross-View Matching Network for Image-Based Ground-to-Aerial Geo-Localization. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[31] Feiran Huang, Xiaoming Zhang, Zhoujun Li, Tao Mei, Yueying He, and Zhonghua Zhao. 2017. Learning Social Image Embedding with Deep Multimodal Attention Networks. In *Thematic Workshops of ACM Multimedia*. ACM, 460–468.

[32] Yao-Hung Hubert Tsai, Liang-Kang Huang, and Ruslan Salakhutdinov. 2017. Learning Robust Visual-Semantic Embeddings. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3571–3580.

[33] Brody Huval, Tao Wang, Sameep Tandon, Jeff Kiske, Will Song, Joel Pazhayampallil, Mykhaylo Andriluka, Pranav Rajpurkar, Toki Migimatsu, Royce Cheng-Yue, et al. 2015. An empirical evaluation of deep learning on highway driving. *arXiv preprint arXiv:1504.01716* (2015).

[34] Andrej Karpathy, Armand Joulin, and Fei-Fei Li. 2014. Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in Neural Information Processing Systems*. 1889–1897.

[35] Andrej Karpathy and Fei-Fei Li. 2015. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3128–3137.

[36] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. 2017. Simple does it: Weakly supervised instance and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 2.

[37] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539* (2014).

[38] Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. 2015. Associating neural word embeddings with deep image representations using fisher vectors. In *IEEE Conference on Computer Vision and Pattern Recognition*. 4437–4446.

[39] Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-Based & Neural Unsupervised Machine Translation. *arXiv preprint arXiv:1804.07755* (2018).

[40] Loic Landrieu and Martin Simonovsky. 2018. Large-scale point cloud semantic segmentation with superpoint graphs. In *IEEE Conference on Computer Vision and Pattern Recognition*. 4558–4567.

[41] Henning Lategahn, Johannes Beck, Bernd Kitt, and Christoph Stiller. 2013. How to learn an illumination robust image feature for place recognition. In *Intelligent Vehicles Symposium (IV)*. IEEE, 285–291.

[42] Jesse Levinson, Michael Montemerlo, and Sebastian Thrun. 2007. Map-Based Precision Vehicle Localization in Urban Environments.. In *Robotics: Science and Systems III*, Vol. 4. 1–8.

[43] Yunpeng Li, Noah Snavely, Dan Huttenlocher, and Pascal Fua. 2012. Worldwide pose estimation using 3d point clouds. In *European Conference on Computer Vision*. Springer, 15–29.

[44] Zhengqi Li and Noah Snavely. 2018. Megadepth: Learning single-view depth prediction from internet photos. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2041–2050.

[45] Tsung-Yi Lin, Serge Belongie, and James Hays. 2013. Cross-view image geolocalization. In *IEEE Conference on Computer Vision and Pattern Recognition*. 891–898.

[46] Tsung-Yi Lin, Yin Cui, Serge Belongie, and James Hays. 2015. Learning deep representations for ground-to-aerial geolocalization. In *IEEE Conference on Computer Vision and Pattern Recognition*. 5007–5015.

[47] Stephanie Lowry, Niko Sünderhauf, Paul Newman, John J Leonard, David Cox, Peter Corke, and Michael J Milford. 2016. Visual place recognition: A survey. *IEEE Transactions on Robotics* 32, 1 (2016), 1–19.

[48] Faisal Mahmood, Wenhao Xu, Nicholas J Durr, Jeremiah W Johnson, and Alan Yuille. 2019. Structured Prediction using cGANs with Fusion Discriminator. In *International Conference on Learning Representations Workshops*.

[49] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. 2015. Deep captioning with multimodal recurrent neural networks (m-rnn). *International Conference on Learning Representations* (2015).

[50] Bogdan C Matei, Nick Vander Valk, Zhiwei Zhu, Hui Cheng, and Harpreet S Sawhney. 2013. Image to LIDAR matching for geotagging in urban environments. In *IEEE Winter Conference on Applications of Computer Vision*. 413–420.

[51] Michael Milford, Chunhua Shen, Stephanie Lowry, Niko Suenderhauf, Sareh Shirazi, Guosheng Lin, Fayao Liu, Edward Pepperell, Cesar Lerma, Ben Upcroft, et al.

2015. Sequence searching with deep-learnt depth for condition-and viewpoint-invariant route-based place recognition. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 18–25.

[52] Niluthpol Chowdhury Mithun, Juncheng Li, Florian Metze, and Amit K Roy-Chowdhury. 2018. Learning Joint Embedding with Multimodal Cues for Cross-Modal Video-Text Retrieval. In *International Conference on Multimedia Retrieval (ICMR)*. ACM, 19–27.

[53] Niluthpol Chowdhury Mithun, Juncheng Li, Florian Metze, and Amit K Roy-Chowdhury. 2019. Joint embeddings with multimodal cues for video-text retrieval. *International Journal of Multimedia Information Retrieval* (2019), 1–16.

[54] Niluthpol Chowdhury Mithun, Rameswar Panda, Evangelos Papalexakis, and Amit Roy-Chowdhury. 2018. Webly Supervised Joint Embedding for Cross-Modal Image-Text Retrieval. In *ACM International Conference on Multimedia*.

[55] Niluthpol Chowdhury Mithun, Cody Simons, Robert Casey, Stefan Hilligardt, and Amit Roy-Chowdhury. 2018. Learning Long-Term Invariant Features for Vision-Based Localization. In *IEEE Winter Conference on Applications of Computer Vision*. 2038–2047.

[56] Tayyab Naseer, Gabriel L Oliveira, Thomas Brox, and Wolfram Burgard. 2017. Semantics-aware Visual Localization under Challenging Perceptual Conditions. In *IEEE International Conference on Robotics and Automation*.

[57] Nathan Piasco, Désiré Sidibé, Cédric Demonceaux, and Valérie Gouet-Brunet. 2018. A survey on Visual-Based Localization: On the benefit of heterogeneous data. *Pattern Recognition* 74 (2018), 90–109.

[58] Bryan Plummer, Matthew Brown, and Svetlana Lazebnik. 2017. Enhancing Video Summarization via Vision-Language Embedding. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[59] Robi Polikar. 2007. Bootstrap inspired techniques in computational intelligence: ensemble of classifiers, incremental learning, data fusion and missing features. *IEEE Signal Processing Magazine* 24, 4 (2007), 59–72.

[60] Andrzej Pronobis, Barbara Caputo, Patric Jensfelt, and Henrik I Christensen. 2006. A discriminative approach to robust visual place recognition. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*. 3829–3836.

[61] Noha Radwan, Abhinav Valada, and Wolfram Burgard. 2018. VLocNet++: Deep multitask learning for semantic visual localization and odometry. *arXiv preprint arXiv:1804.08366* (2018).

[62] Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. 2014. CNN features off-the-shelf: an astounding baseline for recognition. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 806–813.

[63] Krishna Regmi and Mubarak Shah. 2019. Bridging the domain gap for ground-to-aerial image matching. In *IEEE International Conference on Computer Vision*. 470–479.

[64] Axel Rottmann, Óscar Martínez Mozos, Cyrill Stachniss, and Wolfram Burgard. 2005. Semantic place classification of indoor environments with mobile robots using boosting. In *AAAI*, Vol. 5. 1306–1311.

[65] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. 2011. Fast image-based localization using direct 2d-to-3d matching. In *IEEE International Conference on Computer Vision*. 667–674.

[66] Torsten Sattler, Akihiko Torii, Josef Sivic, Marc Pollefeys, Hajime Taira, Masatoshi Okutomi, and Tomas Pajdla. 2017. Are Large-Scale 3D Models Really Necessary for Accurate Visual Localization?. In *IEEE Conference on Computer Vision and Pattern Recognition*. 6175–6184.

[67] Johannes L Schönberger, Marc Pollefeys, Andreas Geiger, and Torsten Sattler. 2018. Semantic Visual Localization. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[68] Zachary Seymour, Karan Sikka, Han-Pang Chiu, Supun Samarasekera, and Rakesh Kumar. 2019. Semantically-Aware Attentive Neural Embeddings for Long-Term 2D Visual Localization. In *British Machine Vision Conference*.

[69] Yujiao Shi, Liu Liu, Xin Yu, and Hongdong Li. 2019. Spatial-Aware Feature Aggregation for Cross-View Image based Geo-Localization. In *Advances in Neural Information Processing Systems*. 10090–10100.

[70] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. 2017. Revisiting unreasonable effectiveness of data in deep learning era. In *IEEE international conference on computer vision*. 843–852.

[71] Niko Sünderhauf, Sareh Shirazi, Feras Dayoub, Ben Upcroft, and Michael Milford. 2015. On the performance of ConvNet features for place recognition. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*. 4297–4304.

[72] Niko Sunderhauf, Sareh Shirazi, Adam Jacobson, Feras Dayoub, Edward Pepperell, Ben Upcroft, and Michael Milford. 2015. Place recognition with ConvNet landmarks: Viewpoint-robust, condition-robust, training-free. *Robotics: Science and Systems XII* (2015).

[73] Yicong Tian, Chen Chen, and Mubarak Shah. 2017. Cross-view image matching for geo-localization in urban environments. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1998–2006.

[74] Carl Toft, Erik Stenborg, Lars Hammarstrand, Lucas Brynte, Marc Pollefeys, Torsten Sattler, and Fredrik Kahl. 2018. Semantic Match Consistency for Long-Term Visual Localization. In *European Conference on Computer Vision*. 383–399.

[75] Liwei Wang, Yin Li, and Svetlana Lazebnik. 2016. Learning deep structure-preserving image-text embeddings. In *IEEE Conference on Computer Vision and Pattern Recognition*. 5005–5013.

[76] Shenlong Wang, Sanja Fidler, and Raquel Urtasun. 2015. Lost shopping! monocular localization in large indoor spaces. In *IEEE International Conference on Computer Vision*. 2695–2703.

[77] Tan Wang, Xing Xu, Yang Yang, Alan Hanjalic, Heng Tao Shen, and Jingkuan Song. 2019. Matching Images and Text with Multi-modal Tensor Fusion and Re-ranking. In *ACM International Conference on Multimedia*. 12–20.

[78] Weiyue Wang and Ulrich Neumann. 2018. Depth-aware cnn for rgb-d segmentation. In *European Conference on Computer Vision*. 135–150.

[79] Yuan Wang, Tianyue Shi, Peng Yun, Lei Tai, and Ming Liu. 2018. PointSeg: Real-Time Semantic Segmentation Based on 3D LiDAR Point Cloud. *arXiv preprint arXiv:1807.06288* (2018).

[80] Ryan W Wolcott and Ryan M Eustice. 2014. Visual localization within lidar maps for automated urban driving. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*. 176–183.

[81] Scott Workman, Richard Souvenir, and Nathan Jacobs. 2015. Wide-area image geolocalization with aerial reference imagery. In *IEEE International Conference on Computer Vision*. 3961–3969.

[82] Hao Wu, Jiayuan Mao, Yufeng Zhang, Yuning Jiang, Lei Li, Weiwei Sun, and Wei-Ying Ma. 2019. Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. In *IEEE Conference on Computer Vision and Pattern Recognition*. 6609–6618.

[83] Jianixn Wu and James M Rehg. 2008. Where am I: Place instance and category recognition using spatial PACT. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1–8.

[84] Yiling Wu, Shuhui Wang, Guoli Song, and Qingming Huang. 2019. Learning Fragment Self-Attention Embeddings for Image-Text Matching. In *ACM International Conference on Multimedia*. 2088–2096.

[85] Dan Xu, Wanli Ouyang, Elisa Ricci, Xiaogang Wang, and Nicu Sebe. 2017. Learning cross-modal deep representations for robust pedestrian detection. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[86] I Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. 2019. Billion-scale semi-supervised learning for image classification. *arXiv preprint arXiv:1905.00546* (2019).

[87] Fisher Yu, Jianxiong Xiao, and Thomas Funkhouser. 2015. Semantic alignment of LiDAR data at city scale. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1722–1731.

[88] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor Fusion Network for Multimodal Sentiment Analysis. In *Conference on Empirical Methods in Natural Language Processing*. 1103–1114.

[89] Sergey Zagoruyko and Nikos Komodakis. 2016. Wide residual networks. *arXiv preprint arXiv:1605.07146* (2016).

[90] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. 2017. Pyramid scene parsing network. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2881–2890.

[91] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, and Yi-Dong Shen. 2017. Dual-Path Convolutional Image-Text Embedding. *arXiv preprint arXiv:1711.05535* (2017).

[92] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2017).

[93] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2019. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision* 127, 3 (2019), 302–321.

[94] Yingying Zhu, Jiong Wang, Lingxi Xie, and Liang Zheng. 2018. Attention-based pyramid aggregation network for visual place recognition. In *ACM international conference on Multimedia*. 99–107.

[95] SM Zolanvari, Susana Ruano, Aakanksha Rana, Alan Cummins, Rogerio Eduardo da Silva, Morteza Rahbar, and Aljosa Smolic. 2019. DublinCity: Annotated LiDAR Point Cloud and its Applications. In *British Machine Vision Conference*.