Optimal Feature Transport for Cross-View Image Geo-Localization

Yujiao Shi,[1,2] **Xin Yu,**[1,2] **Liu Liu,**[1,2] **Tong Zhang,**[1,3] **Hongdong Li**[1,2]

[1]Australian National University, Canberra, Australia.
[2]Australian Centre for Robotic Vision, Australia.
[3]Motovis Australia Pty Ltd
firstname.lastname@anu.edu.au

Abstract

This paper addresses the problem of cross-view image geo-localization, where the geographic location of a ground-level street-view query image is estimated by matching it against a large scale aerial map (*e.g.*, a high-resolution satellite image). State-of-the-art deep-learning based methods tackle this problem as deep metric learning which aims to learn global feature representations of the scene seen by the two different views. Despite promising results are obtained by such deep metric learning methods, they, however, fail to exploit a crucial cue relevant for localization, namely, the spatial layout of local features. Moreover, little attention is paid to the obvious domain gap (between aerial view and ground view) in the context of cross-view localization. This paper proposes a novel *Cross-View Feature Transport (CVFT)* technique to explicitly establish cross-view domain transfer that facilitates feature alignment between ground and aerial images. Specifically, we implement the CVFT as network layers, which transports features from one domain to the other, leading to more meaningful feature similarity comparison. Our model is differentiable and can be learned end-to-end. Experiments on large-scale datasets have demonstrated that our method has remarkably boosted the state-of-the-art cross-view localization performance, *e.g.*, on the CVUSA dataset, with significant improvements for top-1 recall from 40.79% to 61.43%, and for top-10 from 76.36% to 90.49%. We expect the key insight of the paper (*i.e.*, explicitly handling domain difference via domain transport) will prove to be useful for other similar problems in computer vision as well.

1 Introduction

This paper studies the problem of ground-to-aerial cross-view image-based localization. The task is: given a ground-level image of street scene (*e.g.*, a Google street view image), we want to estimate where this image was taken, in terms of estimating its geo-location relative to a large (and geo-tagged) satellite map covering the region of interest (*e.g.*, an entire city). Such a cross-view localization problem is not only practically relevant, thanks to the wide-accessibility of high-resolution satellite imagery, but also scientifically important and challenging. The challenges mainly come from the significant differences in their dif-
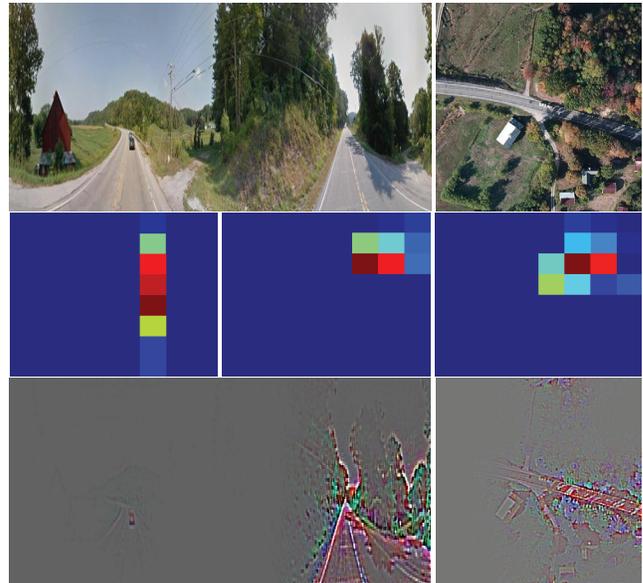
Figure 1: Illustration of the proposed Cross-View Feature Transport (CVFT). First row: a panoramic ground-level query image (left), and its corresponding aerial image (right); Second row: left: an example feature map learned from the ground image shown above; middle: the learned feature map from the aerial image shown above; Right: the transported ground-level feature map using our CVFT. It is clear that CVFT effectively aligns the two feature domains, bringing the two corresponding feature maps closer in the feature space; Bottom row: visualizing the two feature maps (Zeiler and Fergus 2014).

ferent viewpoints (one is from overhead, and the other at ground level) and vast variations in visual appearance.

In this paper, we propose an effective mechanism to explicitly take into account the two issues: (1) preserving local feature spatial layout information, and (2) implementing a mechanism to account for feature domain differences. We first employ a two-branch Siamese-like CNN architecture to extract features from ground and aerial images separately. As seen in Figure 1, there is a spatial domain correspondence between ground panorama and aerial imagery. In or-

der to exploit the spatial information of local features, we propose a Cross-View Feature Transport (CVFT) module to explicitly transform feature maps from one domain (ground) to the other one (aerial).

Specifically, we re-arrange feature maps extracted from one branch by a transport matrix. Then the transformed features from one domain will lie much closer to the corresponding features in the other domain in terms of positions, facilitating similarity comparisons. Moreover, the spatial layout information is also embedded in the features, which makes our features more discriminative.

Our transport matrix is achieved by optimizing a cross-view feature transport problem using differentiable Sinkhorn operation. Since all the layers in our network are differentiable, our CVFT module can be trained in an end-to-end fashion. Extensive experimental results demonstrate that our CVFT improves cross-view geo-localization performance by a large margin on standard benchmarks.

The contributions of this work are as follows:

- A new state-of-the-art cross-view image-based geo-localization method, which outperforms all previous best performing deep networks developed for this task.

- A Cross-View Feature Transport (CVFT) module for domain transfer. Our CVFT module explicitly establishes feature transformations between two different domains, facilitating cross-domain feature similarity matching.

2 Related works

Cross-view image-based Geo-localization: Benefited from recent advance of deep learning, most of the existing works on cross-view image matching adopt two-branch structured CNNs to learn representations for ground and aerial images and treat this task as a deep metric learning problem. The seminal work (Workman and Jacobs 2015) fine-tuned AlexNet (Krizhevsky, Sutskever, and Hinton 2012) on Imagenet (Russakovsky et al. 2015) and Places (Zhou et al. 2014), and then applied the network to extract features for cross-view matching/localization. Workman *et al.*(Workman, Souvenir, and Jacobs 2015) further indicated that fine-tuning the aerial branch by minimizing the distance between aerial and ground images leads to better localization performance. Later, Vo and Hays (Vo and Hays 2016) conducted thorough experiments on the investigation of the formulation and solutions towards this problem, *i.e.*, binary classification or image retrieval, Siamese or Triplet network architecture. Building upon the Siamese network, Hu *et al.*(Hu et al. 2018) embedded a NetVLAD layer on top of a VGG network to extract descriptors that are invariant against large viewpoint changes. Liu & Li (Liu and Li 2019) incorporated per-pixel orientation information into a CNN to learn orientation-selective features for cross-view localization. Though effective these methods are, they all dismiss the domain difference and discard the spatial layout information in ground and aerial images.

Optimal Transport: The optimal transport (OT) theory provides a principled way of comparing distributions and offers an optimal transport plan for matching distributions. However, as the classic OT is a Linear Programming problem,
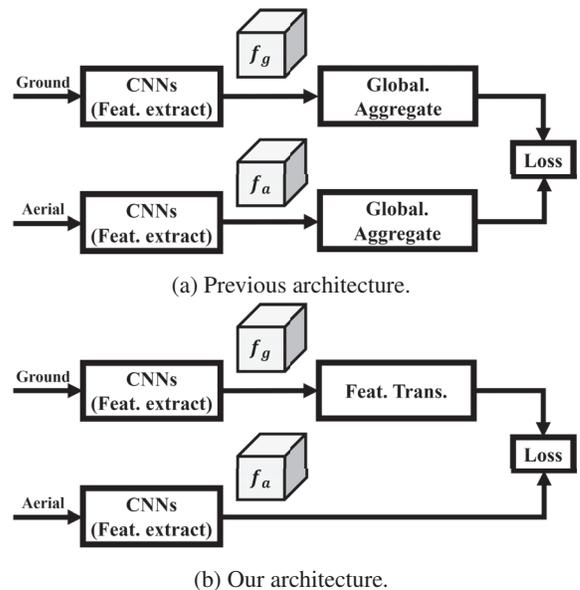

(a) Previous architecture.

(b) Our architecture.

Figure 2: Comparison between traditional methods and ours. (a) Existing deep metric learning network with a global feature aggregation layer;(b) Our new network with Feature Transport. Global Feature Aggregation tends to discard very useful spatial layout information. In contrast, our Feature Transport maps feature from the ground-view domain to the aerial domain, accounting for their domain difference. Details of the Feature Transport will be explained in the main text.

it suffers from expensive computation complexity in terms of (distribution or data) dimensions. Regarding this deficiency, Cuturi *et al.*(Cuturi 2013) proposed an entropy regularized optimal transport problem which can be solved by the Sinkhorn-Knopp algorithm (Sinkhorn and Knopp 1967; Knight 2008) efficiently. After that, it has been successfully applied to solve various machine learning problems, such as domain adaptation (Courty et al. 2017), multi-label learning (Frogner et al. 2015), feature dimensionality reduction (Gautheron, Redko, and Lartizien 2018), generative models (Park and Thorpe 2018) and so on. Inspired by the optimal transform, we construct a feature transport module in this paper to bridge the domain gap and encode the correspondence between ground and aerial features, and the Sinkhorn-Knopp algorithm is employed as a solver to the feature transport problem.

3 Motivation: Cross-View Feature Transport

Existing deep learning based cross-view localization methods often adopt a forced feature embedding procedure that brings matching ground-and-aerial image pairs closer in the feature embedding space while far apart for those non-matching ones. For instance, a Global Aggregation Operation (such as 'global pooling' or VLAD-type feature aggregation (Arandjelovic et al. 2016)) is often adopted.

However, these global aggregation steps often discard (or destroy) the spatial layout information of local deep features

(in other words, the relative positions between those local features in the image plane). This spatial layout information is however very useful and plays a pivotal role in aiding human's geo-localization (*e.g.*, re-orientation or way-finding). For example, human usually memorizes the relative spatial position of a certain object or building in a scene for the localization and navigation purpose. Thus it is important to exploit such cues for the image-based geo-localization.

Besides, there are obvious domain differences between a ground-view image and a satellite image (namely, the former is a panoramic image seen from the ground-level, while the latter depicts an overhead (bird-eye's) view of a geographic region). Hence, simply computing the distance between two deep-features from different domains may neglect the important domain correspondences, thus yielding inferior estimation.

In this paper, we take a direct (and explicit) approach to tackle this issue to account for the domain gap, such as drastic appearance changes as well as geometric distortions. Inspired by the idea of Optimal Transport (OT), we aim to resort a transport matrix to achieve this goal. Specifically, we propose a novel Feature Transport block (in the form of differentiable neural layers) which is able to transport (transfrom or map) features from one domain (*e.g.*, ground) to another (*e.g.*, aerial). Figure 2 shows a comparison between the traditional method using global feature aggregation, and our new method with Feature Transport.

To speed up the computation and make the feature transport objective can be optimized by a deep neural network, we employ an entropy-regularized optimal feature transport which can be solved by the differentiable Sinkhorn operation. The achieved feature transport plan is a doubly stochastic matrix encoding the relationships between cross-domain features. It can be used to transport ground features into the aerial domain such that matching ground and aerial features can be spatially aligned. In the next section, we will give more technical details about the process.

4 The Proposed CVFT Framework

Figure 3 illustrates the overview of our proposed CVFT framework. We first employ a two-branch Siamese-like architecture to learn feature maps for ground and aerial images separately. In order to align the ground-view features to aerial-view features, we then exploit the idea of optimal transport and construct a Feature Transport Module to formulate the feature alignment. To be specific, we design a cost generation sub-network to formulate the feature transport problem, and a Sinkhorn solver to optimize this problem and obtain the feature transport matrix. Then, the feature transport matrix is applied to the ground-view feature maps. In this way, we re-arrange the ground-view feature maps and thus align them to their corresponding aerial ones as well as bridge the domain gap. Our neural network is optimized by a metric learning objective in an end-to-end fashion. Detailed illustrations of each block in the framework will be presented in the following content.

4.1 Feature Extraction

Due to the powerful feature representation ability of deep convolutional neural networks, we employ a CNN, such as VGG (Simonyan and Zisserman 2014), as our backbone network to extract features from input images. Since our backbone CNN, VGG network, outputs the feature maps with a high dimension (*i.e.*, $8 \times 8 \times 512$), we apply another convolutional layer to reduce the feature dimension along the channel dimension rather than the spatial dimension. In this manner, we preserve the spatial layout information of the extracted features. Our final output feature dimension is $8 \times 8 \times 64 = 4096$, which is a commonly used feature representation dimension for image retrieval/matching task (Hu et al. 2018).

4.2 Optimal Feature Transport

Optimal Transport (OT) theory was originally developed for finding the optimal transport plan $\mathbf{P}^*$ that can best align two probability distributions μ_s and μ_g. This is done by minimizing the following transport distance: $\langle \mathbf{P}, \mathbf{C} \rangle_F$, where $\langle \cdot, \cdot \rangle_F$ is the Frobenius norm between two matrices, $\mathbf{C} \in \mathcal{R}^{n_s \times n_t}$ is a cost matrix measuring cost for transporting between source and target samples. $\mathcal{P}$ is a set of transport plans and $\mathbf{P} \in \mathcal{R}^{n_s \times n_t}$. n_s is the number of source samples $\mathbf{x}_s$ and n_t is the number of target samples $\mathbf{x}_t$.

In the cross-view image-based geo-localization, we aim to expose the relationship (*i.e.*,feature transport plan) between ground and aerial features so as to facilitate the cross-domain feature alignment. With this purpose, we generalize the optimal transport from distributions to features.

Cost Matrix Generation: In the cross-view feature transport, we first need to define the cost matrix between the ground features $\mathbf{f}^i(g) \in \mathcal{R}^{h \times w}$ and aerial features $\mathbf{f}^i(a) \in \mathcal{R}^{h \times w}$, where h, w indicate the height and width of feature maps[1], and i is the index of feature channels c.

Ideally, both the ground and aerial images should be used to compute the cost matrix $\mathbf{C}$. However, it is computationally expensive to calculate the cost matrix for each gallery image in large-scale retrieval problems. To speed up the efficiency of calculating $\mathbf{C}$, we propose to use one image from one domain to learn how to transform into the other domain. The other domain information is learned from the final objective function and the cost-matrix generation network is updated through backpropagation. As seen in Figure 3, we employ a regression block to generate a cost matrix $\mathbf{C}$.

Sinkhorn Solver: As the original OT is a Linear Programming problem that requires formidable computational costs regarding to the data dimension, we employ an entropy regularized version (Cuturi 2013) which formulated the feature transport problem as a strictly convex problem that can be solved by the Sinkhorn operation efficiently:

$$\mathbf{P}^* = \arg \min_{\mathbf{P} \in \mathcal{P}} \langle \mathbf{P}, \mathbf{C} \rangle_F - \lambda h(\mathbf{P}), \quad (1)$$

where $h(\mathbf{P})$ is an entropy regularization term of $\mathbf{P}$ and λ is a trade-off weight. The entropy regularization term forces the

---

[1]For simplicity, we resize the ground and aerial feature maps into the same size before they are fed into the feature transport module.
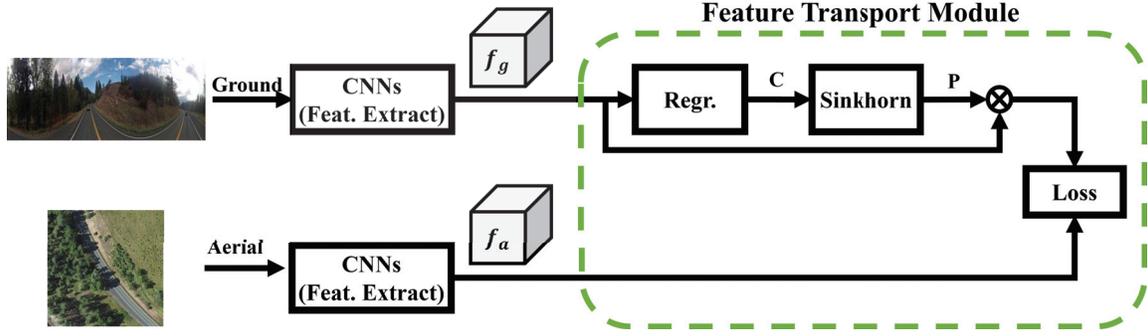
Figure 3: The framework of our method. A series of convolutional layers are applied to function as feature extractors, and then the extracted features are fed into a feature transport module to construct the correspondence between ground and aerial domains.

solution of Equation (1) to be smoother as λ increases, and sparser as λ decreases. $\mathbf{P}^*$ is a doubly stochastic matrix.

Given the cost matrix $\mathbf{C}$, the feature transport objective, as presented in Equation (1), is constructed. We then adopt the Sinkhorn-Knopp algorithm to optimize the objective function and generate feature transport plans for the cross-view features.

Specifically, the Sinkhorn algorithm first applies an exponential kernel on the cost matrix $\mathbf{C}$, yielding $\mathbf{C}' = \exp(-\lambda \mathbf{C})$. Second, the Sinkhorn algorithm normalizes rows and columns iteratively in order to convert $\mathbf{C}'$ to a doubly stochastic matrix. Denote row $\mathcal{N}^r(\cdot)$ and column $\mathcal{N}^c(\cdot)$ normalization as follows:

$$\mathcal{N}_{i,j}^r(\mathbf{C}') = \frac{c'_{i,j}}{\sum_{k=1}^N c'_{i,k}}, \quad \mathcal{N}_{i,j}^c(\mathbf{C}') = \frac{c'_{i,j}}{\sum_{k=1}^N c'_{k,j}}, \quad (2)$$

where $c'_{i,j}$ represents an element in $\mathbf{C}'$.

For the m-th iteration, the output of Sinkhorn operation $\mathcal{S}(\cdot)$ can be represented in a recursively manner:

$$\mathcal{S}^m(\mathbf{C}') = \begin{cases} \mathbf{C}' & m = 0, \\ \mathcal{N}^c(\mathcal{N}^r(S^{m-1}(\mathbf{C}'))) & otherwise. \end{cases} \quad (3)$$

When iterations converge, we thus obtain our feature transport matrix $\mathbf{P}^* = \mathcal{S}^m(\mathbf{C}')$. The Sinkhorn operation is differentiable and its gradient with respect to the input can be calculated by unrolling the sequence of the row and column normalization operations (Cuturi 2013; Santa Cruz et al. 2017). For example, the partial derivatives when $m = 1$ can be defined as follows:

$$\frac{\partial \mathcal{S}^1}{\partial c'_{s,t}} = \frac{\partial \mathcal{N}_{s,t}^c}{\partial \mathcal{N}_{s,t}^r} \cdot \sum_{j=1}^N \left[\frac{[j=t]_{\mathbb{I}}}{\sum_{k=1}^N c'_{s,k}} - \frac{c'_{s,j}}{(\sum_{k=1}^N c'_{s,k})^2} \right], \quad (4)$$

where $[\cdot]_{\mathbb{I}}$ represents an indication function, and s, t and j represent the indices of the row and columns in $\mathbf{C}'$.

After obtaining $\mathbf{P}^*$, we can transport the feature maps between the ground-view and aerial-view images as follows:

$$\mathbf{f}^i(a) = n_a \mathbf{P}^* \mathbf{f}^i(g), \quad \mathbf{f}^i(g) = n_g \mathbf{P}^{*T} \mathbf{f}^i(a), i = 1, \ldots, c, \quad (5)$$

where n_a and n_g represent the feature numbers of aerial and ground images in each channel, respectively.

4.3 Triplet Loss on transported features

After transporting the ground-view feature maps, the ground-view features will be aligned to the corresponding overhead-view features. Then, we apply a metric learning objective to learn feature embedding for both the aligned ground-view images as well as aerial-view images. The triplet loss is widely used as an objective function to train deep neural networks for image localization and matching tasks (Hu et al. 2018; Liu and Li 2019; Vo and Hays 2016). The goal of the triplet loss is to bring matching pairs closer while pushing non-matching pairs far apart. Therefore, we employ a weighted soft-margin triplet loss, $\mathcal{L}$, as our objective to train our neural network, expressed as:

$$\mathcal{L} = \log(1 + e^{\gamma(d_{pos} - d_{neg})}), \quad (6)$$

where d_{pos} and d_{neg} represent the ℓ_2 distance of all the positive and negative aerial features to the anchor ground feature, γ is a parameter which controls the convergence speed (Hu et al. 2018).

5 Experiments

Training and Testing Datasets: We conduct our experiments on two standard benchmark datasets, namely CVUSA (Zhai et al. 2017) and CVACT (Liu and Li 2019), for evaluation and comparisons. These two datasets are both cross-view datasets, and each contains 35,532 ground-and-aerial image pairs for training. CVUSA provides 8,884 image pairs for testing and CVACT provides the same number of pairs for validation (denoted as CVACT_val). CVACT (Liu and Li 2019) also provides 92,802 cross-view image pairs with accurate Geo-tags to evaluate Geo-localization performance (denoted as CVACT_test). Figure 4 illustrates some sample pairs from these two datasets.

Implementation Details: We employ VGG16 with pretrained weights on ImageNet (Deng et al. 2009) as our backbone network to extract image features. The convolutional feature maps of the conv5_3 layer in VGG16 are extracted as our image deep features. Similar to the work of (Hu et al. 2018; Liu and Li 2019), we set γ to 10 for the weighted soft-margin triplet loss. Our network is trained using Adam optimizer (Kingma and Ba 2014) with a learning rate of 10^{-5}

Figure 4: Samples of ground and aerial image pairs. Left are ground panorama images and right are the aerial images. The first two pairs are from CVUSA dataset and the second two pairs are from CVACT dataset.

and batch size of $B_s = 12$. We exploit an exhaustive mini-batch strategy (Vo and Hays 2016) to construct the maximum number of triplets within each batch. For instance, for each ground-view image, there is 1 positive satellite image and $B_s - 1$ negative satellite images, and we can construct $B_s(B_s - 1)$ triplets in total. Similarly, for each satellite image, there is 1 positive ground-view image and $B_s - 1$ negative ground-view images, and thus $B_s(B_s - 1)$ triplets can also be established. Hence, we have $2B_s(B_s - 1)$ triplets in total within each batch.

Evaluation Metrics: We apply the same evaluation metric as (Hu et al. 2018; Liu and Li 2019; Vo and Hays 2016), known as r@K, to exam the performance of our method as well as compare with state-of-the-art cross-view localization methods. r@K measures the performance that how many satellite images in the database need to be retrieved to find the true matching image. More detailed explanations of r@K metric can be found in (Hu et al. 2018; Vo and Hays 2016).

5.1 Effects of Cross-view Feature Transport

In this part, we conduct an ablation study to demonstrate the effectiveness of our proposed Cross-View Feature Transport (CVFT) module and how much our CVFT contributes to improving the geo-localization performance.

Baseline Networks: To illustrate the effectiveness of our proposed CVFT module, we simply remove it from our framework while keeping all the other parts unchanged as our first baseline, named Our network wo/ CVFT. Then we retrain this baseline. In this case, we still preserve the relative spatial information of local features but do not explicitly construct the cross-view correspondences in learning the network. The performance of our first baseline is shown in the fourth row in Table 1. The recalls for the top-1 candidate on CVUSA and CVACT_val datasets are 41.68% and 45.28%, respectively. In comparison to our proposed network, we conclude that it is difficult for our first baseline to
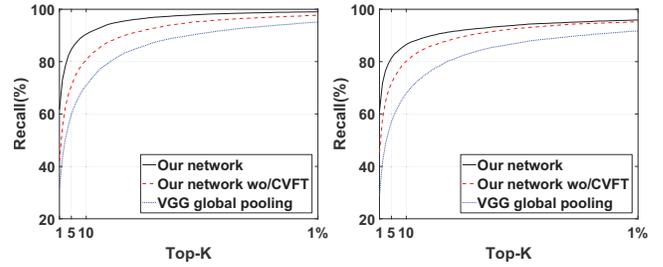

Figure 5: Comparison of Recalls on CVUSA (**Left**) and CVACT_val (**Right**) datasets. Our method outperforms baseline methods by a large margin, with a relative improvement of 19.75% and 15.77% at Top-1 for CVUSA and CVACT_val datasets, respectively.

establish such spatial correspondences or transformations, thus leading to inferior localization performance.

Furthermore, in order to demonstrate the importance of the feature spatial layout information, we employ another baseline, called VGG global pooling. For this baseline, we apply global max pooling to the feature maps extracted by the layer of conv5_3 in VGG16 as our image representation. Then we retrain this baseline. As seen in Table 1, the recall performance for our second baseline is the worst. This phenomenon indicates that the feature spatial layout information plays a critical role in the cross-view geo-localization task.

Our Network with CVFT: Since our proposed network not only explores the spatial layout information but also explicitly establishes the domain correspondences via our CVFT, our network achieves the best performance as indicated by Table 1. As visible in Figure 5, the performance of our network is consistently better than our baselines, and our network outperforms the baseline networks by a large margin on both CVUSA and CVACT_val datasets. For example, we obtain over **15%** improvements on recall@1 on both datasets with the help of our CVFT module.

5.2 Visualization of Feature Alignment

The core insight of our method is that our CVFT is able to align the feature maps from a cross-view image pair. To verify this intuition, we show two examples of cross-view feature map alignments in Figure 7. As visible in Figure 7, we demonstrate two features from the ground-view images (the first column) as well as two features from the aerial-view images (the second column). After applying our estimated feature transport plan $\mathbf{P}^*$ to the ground-view features, we obtain transported ground features (the third column). Compared with the features directly extracted from the aerial images, our transported ground features are well aligned to them. This experiment not only shows the effectiveness of our CVFT but also demonstrates that our network can explicitly establish the relationship between the ground and aerial domains, thus facilitating the cross-view matching. Note that our transport matrix $\mathbf{P}$ is different from spatial transformer networks (STN) (Jaderberg et al. 2015), since cross-view feature transformation cannot be modeled

Table 1: Recall performance on CVUSA (Zhai et al. 2017) and CVACT_val dataset (Liu and Li 2019).

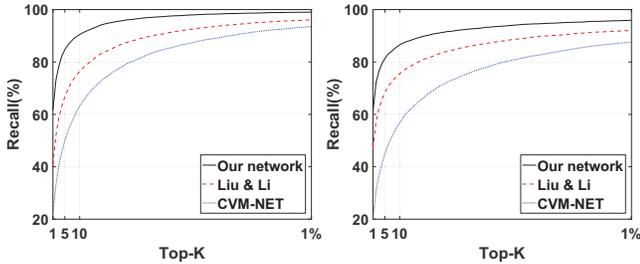|	CVUSA				CVACT_val			
	r@1	r@5	r@10	r@1%	r@1	r@5	r@10	r@1%
Liu & Li (Liu and Li 2019)	40.79	66.82	76.36	96.12	46.96	68.28	75.48	92.04
CVM-NET (Hu et al. 2018)	22.47	49.98	63.18	93.62	20.15	45.00	56.87	87.57
VGG global pooling	31.53	59.85	70.91	95.09	28.98	57.04	67.96	91.72
Our network wo/ CVFT	41.68	70.71	80.71	97.70	45.28	71.74	80.14	95.28
Our network	**61.43**	**84.69**	**90.49**	**99.02**	**61.05**	**81.33**	**86.52**	**95.93**
Our network (orientation noise $=\pm20°$)	52.07	76.87	84.58	97.77	54.81	77.70	83.37	94.92

Figure 6: This figure shows that with our cross-view feature transport module, we outperform all state-of-the-art methods by a large-margin on CVUSA (**Left**) and CVACT_val (**Right**) datasets.

Table 2: Comparison of r@1% by state-of-the-art methods on CVUSA dataset.

	Workman	Zhai	Vo	CVM-NET	Liu	Ours
r@1%	34.30	43.20	63.70	93.62	96.12	**99.02**

by geometric transformation of few parameters. Detailed illustration and experiments are presented in supplementary material.

5.3 Comparisons with the State-of-the-Art

Retrieval on Cross-View Geo-localization: We compare our method with the state-of-the-art cross-view localization methods, including Workman *et al.*(Workman, Souvenir, and Jacobs 2015), Vo *et al.*(Vo and Hays 2016), Zhai *et al.*(Zhai et al. 2017), CVM-NET (Hu et al. 2018) and Liu & Li (Liu and Li 2019). Following the previous works, we also report the recall at top 1% performance in Table 2.

As indicated by Table 2, our method outperforms the state-of-the-art methods significantly. However, as the size of the database increases, the number of gallery images increases. Thus, using the retrieved results at top 1% may be impractical. For instance, there would be more than 80 images in the retrieved results in CVUSA. Therefore, using recalls at top-1, 5, 10 would be more reasonable since the results can be easily analyzed by humans. The first two rows in Table 1 illustrate the performance of state-of-the-art methods on recalls at top-1, 5, 10, and Figure 6 presents the complete r@K performance. Compared to the state-of-the-art, our method outperforms the second best method (Liu and Li
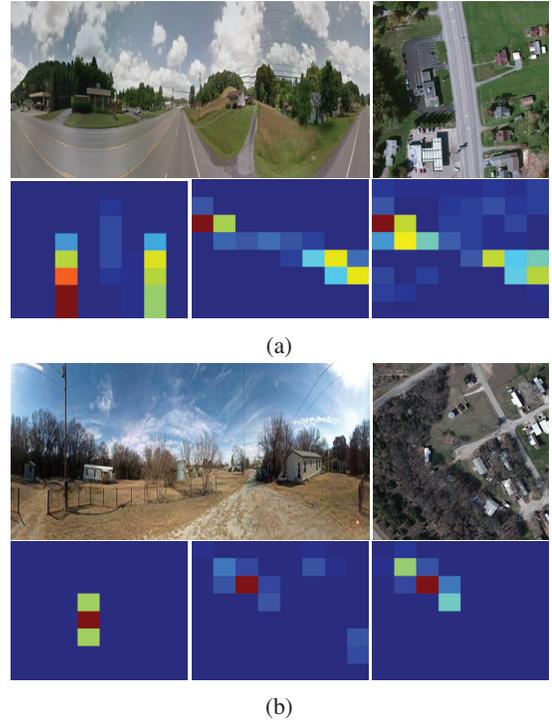

(a)

(b)

Figure 7: Our method successfully aligns cross-view feature maps by using CVFT feature transport. For the second row of each subfigure: (Left:) Input ground-view feature map; (Middle:) The corresponding satellite image feature map; (Right:) CVFT-transported ground feature map. Note the two feature maps are better aligned.

2019) by a large margin. In particular, our network improves 20.64% and 14.09% on the recalls at top-1 for CVUSA and CVACT_val datasets, respectively. Table 1 also indicates our method is more practical for real-world cases.

Accurate Geo-localization: We also conduct an evaluation on the large-scale CVACT_test dataset (Liu and Li 2019) to show the effectiveness of our method for accurate city-scale Geo-localization applications. In addition, we also compare with the state-of-the-art methods, CVM-NET (Hu et al. 2018) and Liu & Li (Liu and Li 2019), on this task.

In this experiment, we follow the evaluation protocol used in (Arandjelovic et al. 2016). To be specific, a query image
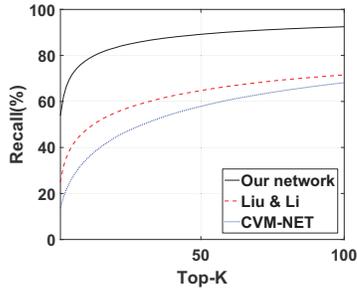
Figure 8: This graph shows that we obtain much higher localization performance (measured by recalls) than previous state-of-the-art methods.

is deemed correctly geo-localized if at least one of the top K retrieved database satellite images is within the distance $d = 25m$ from the ground-truth position of the query image. The percentage of correctly localized queries (recall) is then plotted for different values of K, marked as recall@K. The r@K performance is shown in Figure 8. As expected, our method still significantly outperforms the second-best method (Liu and Li 2019), with an improvement of 28.87% at top-1. We also show some localization examples in Figure 9 and 10, thus demonstrating the superiority of our method for localizing cross-view images in different scenarios.

5.4 Robustness to Orientation Perturbations

The ground-level panoramas and overhead-view aerial images in the CVUSA and CVACT datasets are north aligned. However, it is worthy investigating the impact of orientation perturbation on the cross-view localization performance. As the north direction of a satellite map is often available, we discuss the case that the north direction of a ground-level panorama is not given. In practice, we can estimate an approximate north direction by exploiting some scene clues, such as shadows of trees. Thus we examine the performance of our algorithm on $\pm 20°$ orientation noise. The recall results are presented in the last row in Table 1. It can be seen that our method still outperforms the state-of-the-art (CVM-NET and Liu & Li) although they are tested on the orientation aligned ground-level panoramas.

6 Conclusion

We propose a new cross-view image-based geo-localization deep network in this paper. Our major contribution as well as the central innovation is the introducing of a novel feature transport module (CVFT) which effectively bridges the cross-view domain gap by transforming features from one domain to the other. In contrast to conventional approaches using deep metric learning to match corresponding feature pairs, our method provides a superior and sensible solution. We believe the key idea of this paper (explicitly handling domain differences) is also valuable for solving many other problems in computer vision and machine learning. For the task of large-scale cross-view localization, our method significantly improves performance in terms of top-K recall rate, demonstrating the power of our feature transport frame-
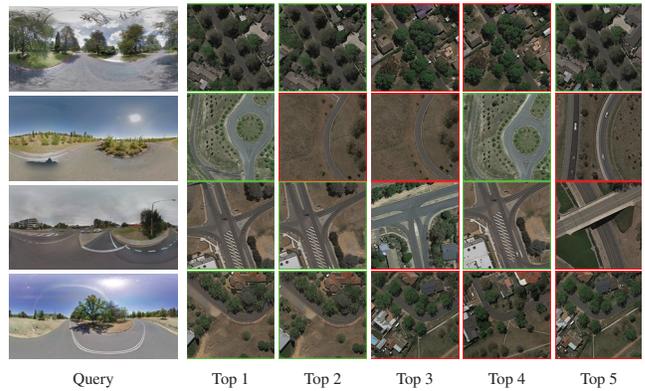

Figure 9: Sample localization results by our method on ACT_test dataset. From left to right: ground-view query image and the Top 1-5 retrieved satellite images. Green and red borders indicate correct and incorrect retrieved results, respectively. Note that our method can localize versatile query images (*e.g.*, suburbs and traffic intersection). (Best viewed in color on screen.)
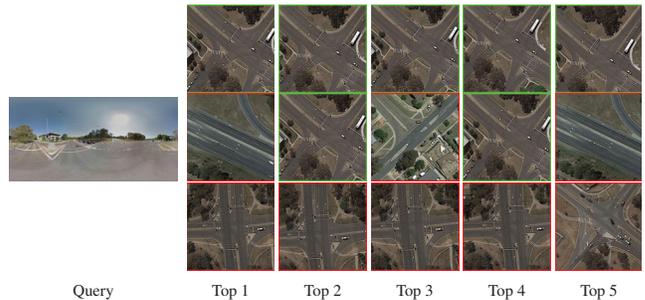

Figure 10: Comparisons of different methods on ACT_test dataset. first row: top-5 recall results by our method; Second row: Liu & Li result (Liu and Li 2019); third row: CVM-NET (Hu et al. 2018). Green and red borders indicate correct and incorrect retrieved results, respectively. Note that our method remains robust for this challenging scenario with many repetitive features.

work. Our method currently assumes the input query image is a full 360° panorama, but we argue this restriction can be relaxed under the same theoretical framework of CVFT, and this is left as a possible future extension.

7 Acknowledgments

References

Arandjelovic, R.; Gronat, P.; Torii, A.; Pajdla, T.; and Sivic, J. 2016. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5297–5307.

Courty, N.; Flamary, R.; Tuia, D.; and Rakotomamonjy, A. 2017. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence* 39(9):1853–1865.

Cuturi, M. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, 2292–2300.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Frogner, C.; Zhang, C.; Mobahi, H.; Araya, M.; and Poggio, T. A. 2015. Learning with a wasserstein loss. In *Advances in Neural Information Processing Systems*, 2053–2061.

Gautheron, L.; Redko, I.; and Lartizien, C. 2018. Feature selection for unsupervised domain adaptation using optimal transport. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 759–776. Springer.

Hu, S.; Feng, M.; Nguyen, R. M. H.; and Hee Lee, G. 2018. Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jaderberg, M.; Simonyan, K.; Zisserman, A.; et al. 2015. Spatial transformer networks. In *Advances in neural information processing systems*, 2017–2025.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Knight, P. A. 2008. The sinkhorn–knopp algorithm: convergence and applications. *SIAM Journal on Matrix Analysis and Applications* 30(1):261–275.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In Pereira, F.; Burges, C. J. C.; Bottou, L.; and Weinberger, K. Q., eds., *Advances in Neural Information Processing Systems 25*. Curran Associates, Inc. 1097–1105.

Liu, L., and Li, H. 2019. Lending orientation to neural networks for cross-view geo-localization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Park, S., and Thorpe, M. 2018. Representing and learning high dimensional data with the optimal transport map from a probabilistic viewpoint. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7864–7872.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* 115(3):211–252.

Santa Cruz, R.; Fernando, B.; Cherian, A.; and Gould, S. 2017. Deeppermnet: Visual permutation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3949–3957.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sinkhorn, R., and Knopp, P. 1967. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics* 21(2):343–348.

Vo, N. N., and Hays, J. 2016. Localizing and orienting street views using overhead imagery. In *European Conference on Computer Vision*, 494–509. Springer.

Workman, S., and Jacobs, N. 2015. On the location dependence of convolutional neural network features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 70–78.

Workman, S.; Souvenir, R.; and Jacobs, N. 2015. Wide-area image geolocalization with aerial reference imagery. In *Proceedings of the IEEE International Conference on Computer Vision*, 3961–3969.

Zeiler, M. D., and Fergus, R. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*, 818–833. Springer.

Zhai, M.; Bessinger, Z.; Workman, S.; and Jacobs, N. 2017. Predicting ground-level scene layout from aerial imagery. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 3.

Zhou, B.; Lapedriza, A.; Xiao, J.; Torralba, A.; and Oliva, A. 2014. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, 487–495.