Coming Down to Earth: Satellite-to-Street View Synthesis for Geo-Localization

Aysim Toker Qunjie Zhou Maxim Maximov Laura Leal-Taixé
Technical University of Munich
{aysim.toker, qunjie.zhou, maxim.maximov, leal.taixe}@tum.de

Figure 1: For a collection of satellite and street images, our method synthesizes the street view for each satellite input (right). It also simultaneously determines the geographic location of a query street image by matching it with the closest satellite image in the database (left→right). This is done in one single architecture which allows for end-to-end training.

Abstract

The goal of cross-view image based geo-localization is to determine the location of a given street view image by matching it against a collection of geo-tagged satellite images. This task is notoriously challenging due to the drastic viewpoint and appearance differences between the two domains. We show that we can address this discrepancy explicitly by learning to synthesize realistic street views from satellite inputs. Following this observation, we propose a novel multi-task architecture in which image synthesis and retrieval are considered jointly. The rationale behind this is that we can bias our network to learn latent feature representations that are useful for retrieval if we utilize them to generate images across the two input domains. To the best of our knowledge, ours is the first approach that creates realistic street views from satellite images and localizes the corresponding query street-view simultaneously in an end-to-end manner. In our experiments, we obtain state-of-the-art performance on the CVUSA and CVACT benchmarks. Finally, we show compelling qualitative results for satellite-to-street view synthesis.

1. Introduction

Estimating the geographic location of an image is a fundamental problem in computer vision with applications in autonomous driving, robotics, and augmented reality. Originally, the problem was cast as an image retrieval task [27, 9, 37, 6, 2, 31, 38, 26], where the goal is to determine the geographic location of a query street view image by comparing it against a database of GPS-tagged street images. The main limitation of this approach is that, even though there are large databases available for this type of imagery, the coverage varies a lot between different regions of the world, and it is generally sparse in rural areas.

Satellite imagery, on the other hand, is broadly available for most parts of the world with services like Google maps. This encouraged researchers to focus on cross-view image-based geo-localization [36, 16, 33, 17, 28, 4, 30, 29] as a more general and inclusive alternative. The overall idea is to predict the latitude and longitude of a street-level image by matching it against a GPS-tagged satellite database. Even though this approach helps to cover vast parts of the world, the significant domain gap between a pair of street view and top-view satellite images, shown in Figure 1, makes cross-view image based geo-localization extremely challenging. For instance, the appearance of the two images can vary significantly as they are typically taken at different times and with different cameras, leading to illumination changes. The biggest challenge, however, comes from the dramatically different viewpoints of street and satellite images – even for human eyes, it is far from obvious that two images show the same location. Satellite images cover a broader area in comparison to the ego-centric viewpoint of the street images. On the other hand, there are a lot of additional fea-

tures in street view images, like facades, that are not visible in the top-view satellite images which would otherwise be extremely useful for precise location retrieval.

In order to alleviate the difficulty of learning cross-view features, [28, 29] use a simple polar coordinate transformation as a preprocessing step for image retrieval. Intuitively, this mimics the real viewpoint transformation from the overhead view to the ground-view. Nevertheless, there is still a significant appearance gap between polar-transformed and real street images. The two views do not overlap perfectly, which limits the retrieval performance. In the last few years, Generative Adversarial Networks (GANs) [8] have proven to be a powerful tool for generating realistic looking images. Recent works [22, 23, 43] applied them for cross-view image synthesis between aerial and ground-level images but they do not evaluate their effectiveness for the geo-localization task. [24] is the first to use pre-trained synthesized images [22] to train a retrieval network for geo-localization. However, this is done in two stages and therefore does not allow for end-to-end training. They obtained less accurate retrieval results than methods based on polar transformations [28, 29]. This suggests that, while GANs create images that *look more realistic*, polar-transformation is more suitable to map the *content* of the images across the two domains.

In this work, our goal is to address the drastic viewpoint difference of the two domains by synthesizing realistic-looking and content-preserving street images from their satellite counterparts for geo-localization. To that end, we integrate a cross-view synthesis module and a geo-localization branch in a single architecture. The main insight here is that these two network components mutually reinforce each other: Learning to generate street images from satellite inputs naturally helps the image retrieval branch, since our network learns to extract local features that are useful across the two input domains. Vice versa, the retrieval branch incentivizes our network to create realistic street views that replicate the content of a given satellite image. Additionally, our network uses polar transformed satellite images as a starting point (*i.e.* as an input to the GAN). This makes the image generation easier, since the spatial layout of the polar transformed image and the street view is approximately the same.

Contribution We propose a novel geo-localization method that is trained jointly for the multi-task setup of both synthesizing ground images from satellite images and retrieving cross-view image matches. We devise a single network for both of these tasks which can be trained in an end-to-end manner. Our method shows strong empirical results, both in terms of the retrieval accuracy and synthesis quality. For geo-localization, we obtain state-of-the-art performance on standard large-scale cross-view retrieval benchmarks. Moreover, our pipeline generates highly realistic street views that strongly resemble real, panoramic street images. Remarkably, our method outperforms existing cross-view synthesis approaches that use semantic labels as supervision during training.

2. Related work

The main challenge of cross-view image based geo-localization is the drastic appearance and viewpoint differences between satellite and street view image pairs. For our discussion of existing work, we distinguish between methods that directly extract viewpoint invariant features on the input images and methods that apply an explicit viewpoint transformation to the inputs.

Domain-invariant features A central question in cross-view image based geo-localization is how to extract features that are invariant to the appearance gap between satellite and street view images. Early works [15, 21, 5, 3] built hand-crafted pipelines where extracted features have an explicit semantic interpretation, like detecting buildings.

Following the success of deep learning for several computer vision tasks, new approaches successfully applied deep convolutional neural networks (CNN) to learn feature representations for cross-view image based geo-localization. The first approaches in this line of work were based on the AlexNet [13] model pretrained on ImageNet [7] and the Places [42] dataset. Originally, the pre-trained weights were used directly to match image pairs without any additional training [35]. Further improvements were proposed in [36], which refined the features of the satellite images to make them more coherent with the pre-trained descriptors on the street level. [16] utilized a siamese network to learn features for both street view and 45° aerial images with a contrastive loss. In subsequent work, [33] explored several CNN architectures and concluded that a triplet CNN trained with a soft-margin triplet loss is most suitable to extract deep features from cross-view image pairs.

Most of these approaches used a standard fully connected layer to combine local features into a global feature representation. In contrast to that, [11] embedded a learnable NETVLAD [1] layer to aggregate local CNN features. [17] showed that orientation information, in the form of hand-crafted UV maps, helps to convey the approximate viewpoint difference to the network during training. Recently, [4] applied both spatial and channel-wise attention to the feature maps and trained them with a hard exemplar reweighting triplet loss.

Despite the abundant progress in improving the architecture and losses for learning cross-view feature representations, trying to overcome the domain gap purely via feature learning remains a challenging open problem.

Viewpoint transformation Instead of focusing merely on the feature representation, recent approaches transformed the input images to explicitly address the viewpoint discrepancy between satellite and street view images.

The first among these approaches [39] synthesized street-level information from top-view satellite inputs. In particular, they learned to map semantic labels from the satellite view to the ground view and use them to create street view information. Similarly, [22] applied a conditional GAN that creates ground images from the satellite view and vice versa. [18] proposed to generate street views from satellite images by utilizing depth maps and semantic labels. Even though all of these approaches generate novel viewpoints between satellite and street observations, they do not explicitly apply it to the geo-localization problem.

A key formalism towards closing the domain gap between top-view and ground-level images was proposed by [28]. The main insight here is that the viewpoint transformation can be approximated with a simple change of coordinates – in particular, a polar coordinate transformation. This approximately preserves the content of the satellite images, but the resulting images are far from realistic street views. Using the transformed images for retrieval, [28] further proposed to learn spatial attention maps to aggregate the local CNN features into global image descriptors. On top of the polar transformation, [29] trained a siamese network with a dynamic matching module to learn discriminative feature representations along the horizontal direction. To that end, features from the ground view are shifted such that they correlate with the polar transformed images.

More realistic street-view images can be generated with GANs [22] which was later used as an additional input to train a retrieval network [24]. While the obtained images look realistic, this approach lacks a strong incentive for the image generation to preserve the content of the input images, which negatively impacts the retrieval performance. Consequently, most existing cross-view synthesis approaches require semantic maps for a sufficient preservation of content [22, 39, 18].

In this paper, we take a different approach. We show that we do not need semantic maps to obtain realistic-looking and content-preserving street-from-satellite images. We observe that existing works treat the tasks of image synthesis and retrieval separately, even though the synergies are clear. We show that our proposed multi-task training of image synthesis and retrieval in an end-to-end manner leads to state-of-the-art results, both in terms of geo-localization and cross-view image generation.

3. Method

In this section, we describe our proposed multi-task approach to geo-localization, see Figure 2 for an overview. The main idea is to jointly address the cross-view image retrieval and satellite-to-street view synthesis in a single framework. Specifically, we project a given pair of satellite and street images into their latent feature space and use those features simultaneously for both tasks. On one hand, the retrieval branch makes sure that the *content* of the generated images is true to the real scene depicted. At the same time, the image synthesis biases our model to learn features that are consistent across the two input domains which, in turn, benefits the localization.

Initially, we apply a polar transformation to the satellite inputs [29, 28], which maps their content to an approximate street view, see Section 3.1. We then synthesize a realistic street view from the polar-transformed images, see Section 3.2. At the same time, the network learns to set satellite-street pairs in correspondence in the image retrieval branch, which we outline in Section 3.3. Finally, we provide details on the learning procedure in Section 3.4. Also, see our supplementary material for more technical implementation details.

3.1. Polar transformation

As shown in earlier work [28, 29], we can partially bridge the domain gap of our input pairs with a simple polar coordinate transformation of the top-view satellite inputs:

$$\begin{aligned} x_i^{\mathrm{s}} &= \frac{W_{\mathrm{s}}}{2} + \frac{W_s}{2}\frac{y_i^{\mathrm{ps}}}{H_{\mathrm{ps}}}\sin\left(\frac{2\pi}{W_{\mathrm{ps}}}x_i^{\mathrm{ps}}\right) \\ y_i^{\mathrm{s}} &= \frac{H_{\mathrm{s}}}{2} - \frac{H_s}{2}\frac{y_i^{\mathrm{ps}}}{H_{\mathrm{ps}}}\cos\left(\frac{2\pi}{W_{\mathrm{ps}}}x_i^{\mathrm{ps}}\right) \end{aligned} \tag{1}$$

Here, $(x_i^{\mathrm{s}}, y_i^{\mathrm{s}})$ and $(x_i^{\mathrm{ps}}, y_i^{\mathrm{ps}})$ are pixel coordinates of the satellite and polar transformed images, respectively. The dimensions are specified by $W_{\mathrm{s}} \times H_{\mathrm{s}}$ and $W_{\mathrm{ps}} \times H_{\mathrm{ps}}$. In this formulation, circular lines in the top-view satellite images become horizontal lines in the ground view. Vice-versa, radial lines correspond to vertical lines in the new set of coordinates. In particular, the north-line, which is a vertical line originating from the center of the satellite image, corresponds to the vertical line at $\frac{W_{\mathrm{ps}}}{2}$ in the transformed image.

Overall, this transformation produces image pairs that respect the content of the scene, *i.e.*, they have roughly the same arrangement of objects in the scene. However, that alone is not sufficient to completely close the domain gap between the two views: The overlap is typically not perfect and a lot of features, like, e.g., the sky as seen from the ground-view, can simply not be recovered in that manner. Consequently, in the next step, we convert the polar transformed images to street images using a generative model.

3.2. Generator and discriminator networks

Generative Adversarial Networks (GANs) [8] are nowadays broadly used for image synthesis tasks in computer

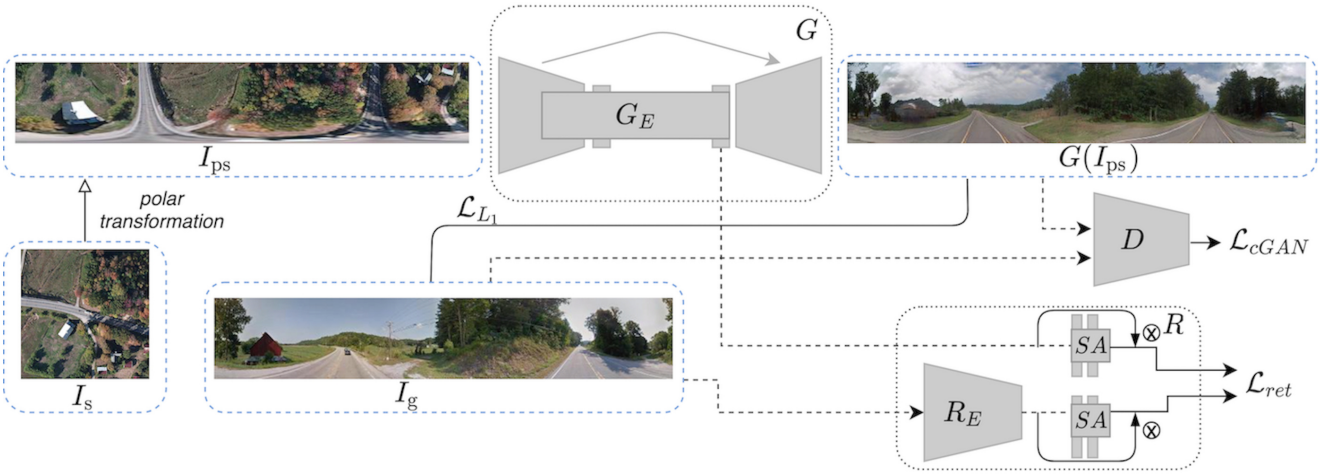

Figure 2: An overview of our network. We convert the pixel coordinates of the top-view satellite image I_{s} to I_{ps}. Then our generative network G synthesizes the street image $G(I_{\text{ps}})$. In the same forward pass, the network feeds the projected satellite features $G_E(I_{\text{ps}})$ and the corresponding ground image I_{g} to the retrieval branch. Network R_E extracts the local features from the real street view analogous to G_E. SA is a spatial-aware attention module that aggregates the extracted local features into global image descriptors. $\mathcal{L}_{cGAN}$, $\mathcal{L}_{L_1}$, $\mathcal{L}_{ret}$ are the loss functions that we used for learning, see Section 3.4.

vision. The main appeal of this class of architectures is that they are able to generate highly realistic images. This is typically done via adversarial training of two opposing networks, the generator G and the discriminator D. We follow the lines of recent conditional GAN method [12] since our goal is to synthesize realistic street views that, at the same time, replicate the content from a reference satellite image.

The first component of our model is the *generator* G which takes a polar-transformed satellite image I_{ps} as an input and translates it into a photo-realistic street panorama $G(I_{\text{ps}})$. The polar-coordinate representation, in this context, is a highly useful preprocessing step since the general outline of the transformed image already resembles the actual street view. This takes some of the burden of bridging the satellite-street domain gap from the generator. The generated images $G(I_{\text{ps}})$, as well as the ground-truth street views I_{g}, are then fed to the *discriminator* D which tries to determine whether the respective images are real or fake. The feedback from this discriminator in turn incentivizes the generator to create images that are indistinguishable from real street views.

In the remainder of this section, we briefly outline the architecture of the two network components G and D. For further details, we refer the interested reader to our supplementary material.

Generator Our generator network G is designed as a U-Net [25] architecture, which consists of residual blocks [10]. The first few downsampling layers, together with the network bottleneck, are called the image encoder G_E. Specifically, G_E consists of 3 residual downsampling blocks that reduce the spatial size by a factor of 4 each. On this reduced resolution, the bottleneck layers further refine the latent features with 6 residual blocks. In the remainder of the generator $G \setminus G_E$ we use 3 residual upsampling blocks to obtain a synthesized street-level image $G(I_{\text{ps}})$ with the same resolution as the polar-transformed input image I_{ps}. Between all downsampling and upsampling blocks we use skip connections as a standard trick to improve the network's convergence. Furthermore, we use instance normalization [32] after each residual block and spectral normalization [20] after each convolution layer.

Discriminator We construct the discriminator D as a PatchGAN [12, 14] classifier. For a given $H_{\text{ps}} \times W_{\text{ps}}$ street-view image, the discriminator D downscales the spatial size to smaller patches and classifies each patch as either real or fake. The patch-wise strategy is particularly beneficial for synthesizing street view images, which typically consist of recurring patterns of streets, trees, and buildings. Since the global coherency is secondary in this context, the classifier can place a higher emphasis on fine-scale details.

3.3. Retrieval network

Having defined our image synthesis module, we now describe our retrieval branch R. The goal is to localize a given query street image I_{g} by matching it against a database of satellite images. R consists of two parts: An encoder block R_E for I_{g} and a spatial attention module SA that converts obtained local features of street and satellite im-

ages into global descriptors. For R_E, we use a modified ResNet34 [10] backbone which extracts local features for the street-view input. We do not, however, compute an analogous latent encoding of the satellite inputs I_{ps} here. Instead, we reuse the features from the generator encoder $G_E(I_{\text{ps}})$.

This is the core idea of our *multi-task* setup: By using the learned features $G_E(I_{\text{ps}})$ for both the synthesis and retrieval tasks, we allow these two aspects of the learning procedure to interact and reinforce each other. The retrieval part by itself is limited to detect and identify similar objects. The learned features from the image synthesis task, on the other hand, provide an explicit notion of domain transfer, since we learn to translate images across the two domains. In turn, the retrieval network compels the generator branch to learn features that are eventually useful for image matching – this yields realistic generated images that also faithfully depict the content of the scene.

Spatial-aware feature aggregation The generator and retrieval feature encoders G_E and R_E learn local feature representations on both the polar-transformed satellite and the street images. In order to convert these local features $F_{\text{ps}} := G_E(I_{\text{ps}})$ into a global descriptor $\tilde{F}_{\text{ps}}$, we use a spatial-aware feature aggregation [28] layer. For a given set of input features, this module predicts k spatial attention masks $A_1, \dots, A_k \in \mathbb{R}^{H \times W}$. These masks A_i are obtained by max-pooling $F_{\text{ps}} \in \mathbb{R}^{H \times W \times C}$ along the channel dimension C and refining the obtained features with two consecutive full-connected layers. The global feature components $\tilde{F}_{\text{ps},i} \in \mathbb{R}^C$ are then defined as a weighted combination of the input features and the attention masks A_i:

$$\tilde{F}_{ps,i} := \left\langle F_{ps}, A_i \right\rangle_F. \tag{2}$$

Here, $\langle \cdot, \cdot \rangle_F$ denotes the Frobenius inner product. Finally, we obtain a global descriptor $\tilde{F}_{\text{ps}}$ by stacking $\tilde{F}_{ps,1}, \dots, \tilde{F}_{ps,k}$ into one kC-dimensional feature vector.

3.4. Learning

The goal of our method is to jointly retrieve the correct satellite match for a given query street view, as well as synthesizing the corresponding street view from the satellite image. To that end, we devise the following loss function:

$$\mathcal{L} = \lambda_{cGAN}\mathcal{L}_{cGAN} + \lambda_{L_1}\mathcal{L}_{L_1} + \lambda_{ret}\mathcal{L}_{ret}. \tag{3}$$

During training, we then update the weights of the three components of our model G, D and R in an adversarial manner:

$$\min_{G,R} \max_{D} \mathcal{L}(G, R, D). \tag{4}$$

In the remainder of this section, we describe in detail how the three components of our composite loss in Equation (3) are defined.

Conditional GAN loss For the image generation task, we define a conditional GAN loss [12]:

$$\begin{aligned}\mathcal{L}_{cGAN}(G, D) = \mathbb{E}_{I_{\text{ps}}, I_{\text{g}}}\big[\log D(I_{\text{ps}}, I_{\text{g}})\big] + \\ \mathbb{E}_{I_{\text{ps}}}\big[\log(1 - D(I_{\text{ps}}, G(I_{\text{ps}})))\big].\end{aligned} \tag{5}$$

While the discriminator D tries to classify images into real (for I_{g}) and fake (for $G(I_{\text{ps}})$), the generator G tries to minimize the loss by creating realistic images. The corresponding satellite image I_{ps} is applied as a condition to both the discriminator and the generator.

Reconstruction loss The second component in Equation (3) is a L_1 reconstruction loss which minimizes the distance between the predicted $G(I_{ps})$ and the ground-truth street-level images I_g:

$$\mathcal{L}_{L_1}(G) = \mathbb{E}_{I_{\text{g}}, I_{\text{ps}}}\big[\|I_{\text{g}} - G(I_{\text{ps}})\|_1\big]. \tag{6}$$

While, in principle, $\mathcal{L}_{cGAN}$ suffices to obtain meaningful translations, $\mathcal{L}_{L_1}$ still helps the network to capture low-level image features and thereby steers the image synthesis to convergence.

Retrieval loss Finally, we use a supervised retrieval loss for the geo-localization task, which is specified as a weighted soft-margin ranking loss [11]:

$$\begin{aligned}\mathcal{L}_{ret}(G_E, R_E, SA) = \\ \mathbb{E}_{I_{\text{ps}}, I_{\text{g}}}\, \mathbb{E}_{\tilde{I}_{\text{g}} \neq I_{\text{g}}}\big[\log(1 + e^{\alpha \text{d}(I_{\text{g}}, I_{\text{ps}}) - \alpha \text{d}(\tilde{I}_{\text{g}}, I_{\text{ps}})})\big].\end{aligned} \tag{7}$$

Here, the distance metric between a pair of ground and satellite images I_g and I_{ps} is defined as the squared L_2 distance between the learned features of both images:

$$\text{d}(I_{\text{g}}, I_{\text{ps}}) := \|SA(R_E(I_{\text{g}})) - SA(G_E(I_{\text{ps}}))\|_2^2. \tag{8}$$

Intuitively, $\mathcal{L}_{ret}$ aims at decreasing the distance of positive matches in the latent space and pushes negative pairs apart.

4. Experiments

In our experiments, we evaluate the performance of our method both in terms of geo-localization and cross-view image synthesis. Overall, our results indicate that these two tasks reinforce each other and the joint training improves performance significantly. We present our quantitative results on cross-view geo-localization in Section 4.2 and on street view synthesis in Section 4.3, with comparisons to state-of-the-art baselines. Furthermore, in Section 4.3, we present qualitative results and comparisons. Finally, we provide an ablation study for further insights into how the different components of our method contribute to our results, see Section 4.4.

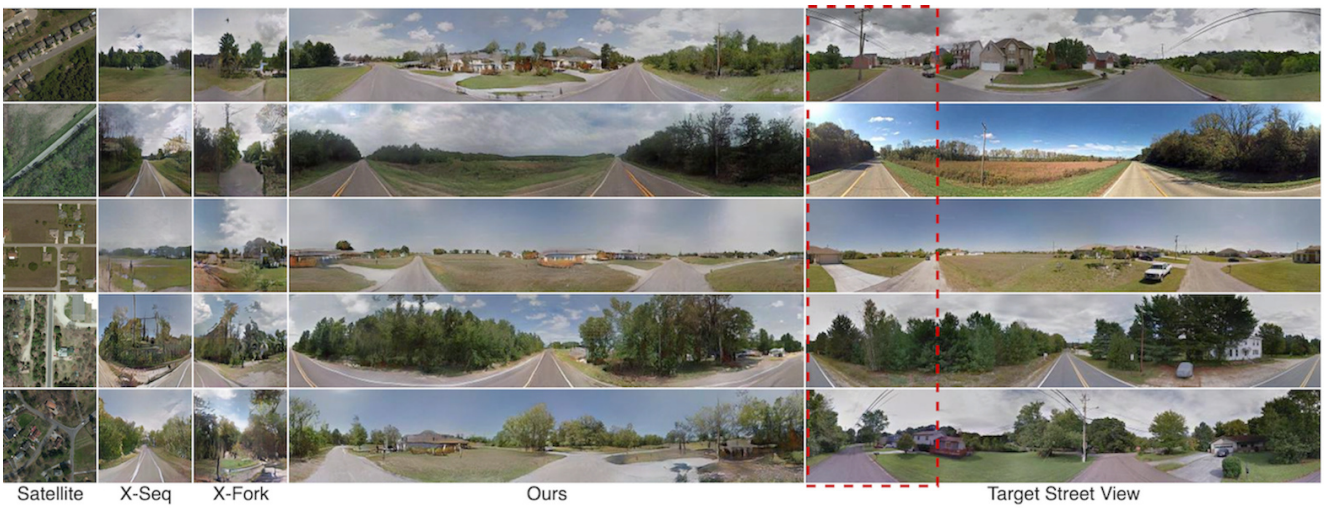

Figure 3: Qualitative comparisons for cross-view image synthesis on the CVUSA benchmark. We compare the images generated by our method with the best baselines X-Fork and X-Seq [22]. Note, that they focus on synthesizing the first quarter of the street view (which is equivalent to the red, dashed boxes on the target street-view), our method is able to create coherent full street view panoramas.

Method	CVUSA_val				CVACT_val				CVACT_test			
	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%
CVM-Net [11]	22.47	49.98	63.18	93.62	20.15	45	56.87	87.57	4.06	16.89	24.66	56.38
Liu, et al. [17]	40.79	66.82	76.36	96.12	46.96	68.28	75.48	92.01	19.9	34.82	41.23	63.79
Regmi, et al. [24]	48.75	-	81.27	95.98	-	-	-	-	-	-	-	-
CVFT [30]	61.43	84.69	90.49	99.02	61.05	81.33	86.52	95.93	34.39	58.83	66.78	95.99
SAFA [28]	89.84	96.93	98.14	99.64	81.03	92.8	94.84	98.17	55.5	79.94	85.08	94.49
DSM [29]	91.96	97.50	**98.54**	**99.67**	82.49	92.44	93.99	97.32	35.55	60.17	67.95	86.71
Ours	**92.56**	**97.55**	98.33	99.57	**83.28**	**93.57**	**95.42**	**98.22**	**61.29**	**85.13**	**89.14**	**98.32**

Table 1: A summary of our quantitative geo-localization experiments. We compare the recall-k (R@k) retrieval accuracy of our method with the current state-of-the-art on the CVUSA [39] and CVACT [17] benchmarks.

4.1. Datasets

We consider the standard large scale cross-view benchmarks CVUSA [39] and CVACT [17], which consist of 44,416 and 137,218 pairs of top view satellite and panoramic street view images, respectively. Images depict streets of both rural and urban scenes. The orientation of the images is normalized, such that the north direction corresponds to the top part of satellite images and the center of the street-level images.

For CVUSA, the first 35,532 ground-to-aerial image pairs are used for training and the remaining 8,884 pairs for validation. Additionally, the images in CVUSA are endowed with street view semantic segmentation labels, which we do not use since our method does not depend on any additional information.

For the sake of consistency, the authors of CVACT [17] chose the same training and validation set sizes as in CVUSA. The remaining 92,802 pairs comprise the test set. Additionally, the CVACT dataset provides UTM coordinates for each satellite-street pair. This large test set allows for a thorough investigation of the generalization ability of our proposed algorithm.

Moreover, since a lot of image pairs in CVACT were taken in close proximity to each other, for the test set of this benchmark, a retrieved satellite image is considered correct, as long as the distance to the actual ground truth match is less than 5 meters – *i.e.* there might be multiple correct satellite matches for a given street level image.

4.2. Cross-view geo-localization

Recall metric For our geo-localization results on the CVUSA and CVACT benchmarks, we followed the standard evaluation protocol from prior work [17, 28, 29, 30].

The main performance indicator here is the recall-k (R@k) retrieval accuracy. Since our algorithm produces L_2 distances for each potential street view to satellite match, we can output a set of plausible matches. The R@k value is therefore defined as the ranking of the satellite ground-truth images which are correctly classified in one of the top-k matches for a given street view image. In particular, the R@1 metric measures the fraction of correct one-to-one matches.

Discussion Table 1 shows our recall-k retrieval accuracies R@1, R@5, R@10, and R@1%, in comparison to existing methods. There are two things we would like to point out: First of all, our method outperforms prior works on R@1 and R@5. The R@1 metric is a crucial criterion since it quantifies the percentage of exact matches. More importantly, our network shows strong results on the CVACT test set, which contains a large, city-scale collection of images. Specifically, our approach outperforms the best baselines by a significant margin for all considered metrics.

4.3. Ground view image synthesis

Metrics For street view synthesis, we use the PSNR, SSIM, and Sharpness difference (SD) metrics. These quantify the pixel level difference between synthesized and ground-truth street views in terms of primitive geometric properties, see [22, 19, 34] for definitions. Additionally, we examine two task-specific metrics which were proposed by [22]. The idea is to assess the similarity of our generated images and the real street views by comparing the predictions of a separate image classifier. Specifically, the class predictions of the images are assigned to one of 365 different categories from the Places dataset [41] with a standard AlexNet [13] image classifier. We can then measure the top-1 and top-5 classification scores (CS-1 and CS-5) of the fake images. Additionally, we compute the KL-divergence between the class label distributions of the real and synthesized street images. Finally, we compare the perceptual similarity score (LPIPS) by using the AlexNet [13] backbone, see [40] for a definition.

Discussion Table 2 contains a summary of our image synthesis results in terms of the quality metrics mentioned above. We compare our generated images to the current state-of-the-art in cross-view image synthesis. The first baseline method [39] predicts auxiliary semantic segmentation labels on the satellite images, maps them to the street view, and uses them to generate a corresponding ground-level image. [22] introduced two different architectures: X-Fork and X-Seq. Moreover, [22] shows comparisons to the generic image-to-image translation architecture Pix2pix [12] on cross-view synthesis. Along the same lines, we provide comparisons to [12] analogously to the setting discussed in [22].

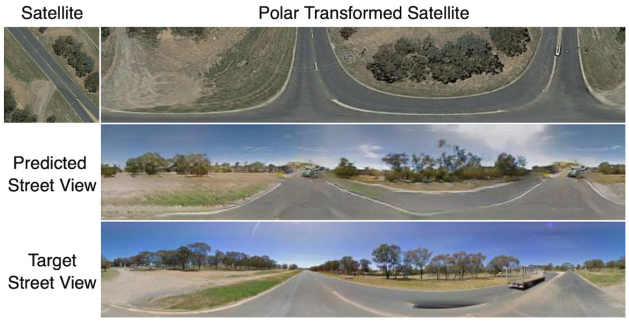

Figure 4: Qualitative result on CVACT. We show the input satellite-to-street pair, as well as the polar transformed satellite image and our synthesized street view. Note, that the considered baselines [22] cannot be applied here since there are no ground-truth annotated semantic maps on CVACT.

We want to point out that all the baselines we consider here require semantic segmentation masks during training. Remarkably, our method still outperforms these existing methods, even though it does not rely on supervision in terms of these semantic labels.

Qualitative Experiments Our quantitative results show that our method indeed synthesizes more realistic and accurate street views than prior approaches. First, we present a qualitative comparison on the validation set of CVUSA in Figure 3. Note, that here we consider the current state-of-the-art baselines X-Seq and X-Fork which specialize on generating the first quarter of the street view. Additionally, we show qualitative results on the test set of CVACT in Figure 4. On this benchmark, the other baselines unfortunately cannot be applied, since they require semantic segmentation labels during training which are not available for CVACT. Our method is able to generate highly plausible street-views, despite the fact that it uses less supervision than prior approaches [39, 22].

4.4. Ablation Study

Geo-localization First of all, we investigate how the different components of our method affect the retrieval performance on CVUSA [39]. To that end, we perform the following ablations and report the results in Table 3: The central question here is how much the image synthesis branch impacts the geo-localization task. First, we train the retrieval branch of our network without the generator *decoder head* and the *discriminator* network (**i.**). This means that the local features extracted by the generator encoder are only passed on to the R branch and no street-view images are generated. This modified network leads to a lower recall accu-

Method	SSIM(↑)	PSNR(↑)	SD(↑)	CS-1(↑)	CS-5(↑)	KL Scores(↓)	LPIPS(↓)
Zhai, et al.et al. [39]	0.414	11.502	10.631	13.97	42.09	27.43 ± 1.63	-
Pix2pix [12]	0.392	11.671	12.537	7.33	25.81	59.81 ± 2.12	0.595
X-Seq [22]	0.423	12.820	12.451	15.98	42.91	15.52 ± 1.73	0.590
X-Fork [22]	0.435	13.064	12.684	20.58	50.51	11.71 ± 1.55	0.609
Ours	**0.447**	**13.895**	**15.221**	**33.23**	**65.85**	**3.59 ± 0.92**	**0.474**

Table 2: Quantitative experiments on image synthesis on CVUSA benchmark. ↑ indicates higher is better, vice versa ↓ indicates lower is better.

racy since it cannot use the reinforced features from the image synthesis. The second experiment again omits the discriminator network and the GAN loss $\mathcal{L}_{cGAN}$ but still predicts a street view, solely based on the $\mathcal{L}_1$ reconstruction loss (**ii.**). This modification leads to a less significant drop in accuracy, since the reconstruction of the street view supports the retrieval part at least to some extent. Another aspect we want to examine (**iii.**) is how the retrieval performance alters if we pass on the generated images $G(I_{\text{ps}})$ to the retrieval branch R instead of the latent bottleneck features from G_E. Here, we again observe a decrease in performance. The generated images themselves are simply not sufficient to convey the same information richness as the bottleneck features – ultimately, the retrieval branch has less latent information available. Overall, the results in Table 3 suggest that the image synthesis indeed benefits the retrieval accuracy. This can by and large be attributed to the fact that learning to generate cross-view images yields local features that are more coherent across the different input domains.

Method	CVUSA			
	R@1	R@5	R@10	R@1%
i. w/o $G\&D$	88.06	96.47	97.88	99.62
ii. w/o $\mathcal{L}_{cGAN}$	91.92	97.22	98.29	99.65
iii. w/ $G(I_{ps})$	89.98	96.78	98.04	**99.66**
Ours	**92.56**	**97.55**	**98.33**	99.57

Table 3: Ablation study on our geo-localization experiments. We show the retrieval recall-k (R@k) accuracies for different versions of our full pipeline, see Section 4.4 for more details.

Street view synthesis For the image synthesis task, we consider the following ablations of our full pipeline, see Table 4 for a summary of the results: First of all, we measure the image quality of the pure, polar-transformed satellite images in comparison to the input street images. These results confirm that a simple geometric coordinate transformation is clearly inferior to a learned, generative model. Furthermore, we train our image generation branch without the retrieval head R. These results suggest that joint training indeed benefits the image synthesis task. The reason for that is, again, that the multi-task learning incentivizes our network to learn superior features which ultimately improves the performance of both tasks at the same time.

Method	CVUSA		
	SSIM↑	PSNR↑	SD↑
I_{ps} vs I_{g}	0.2892	10.7325	14.2291
w/o R	0.4392	13.6858	15.0843
Ours	**0.4472**	**13.8952**	**15.2215**

Table 4: Ablation study on our image synthesis experiments. This shows, that our generator produces the most accurate images, in combination with the retrieval branch.

5. Conclusion

We presented "Coming Down to Earth", a new framework for cross-view image-based geo-localization. Our model integrates image synthesis and retrieval in one architecture which is end-to-end trainable. The key insight is that satellite-to-street view synthesis promotes a latent feature space that is coherent across the two input domains, which benefits the localization. The image retrieval branch, on the other hand, naturally incentivizes the generator to create images that faithfully depict the content of the scene. Remarkably, our method outperforms existing cross-view synthesis approaches, even though it does not rely on any additional semantic information. Finally, we obtain state-of-the-art performance in terms of cross-view geo-localization on both considered benchmarks CVUSA and CVACT.

Acknowledgements We would like to thank Marvin Eisenberger for valuable discussions. This research was supported by the Humboldt Foundation through the Sofja Kovalevskaja Award and the Helmholtz Association under the joint research school "Munich School for Data Science - MUDS".

References

[1] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5297–5307, 2016. 2

[2] Relja Arandjelović and Andrew Zisserman. Dislocation: Scalable descriptor distinctiveness for location recognition. In *Asian Conference on Computer Vision*, pages 188–204. Springer, 2014. 1

[3] Mayank Bansal, Harpreet S Sawhney, Hui Cheng, and Kostas Daniilidis. Geo-localization of street views with aerial image databases. In *Proceedings of the 19th ACM international conference on Multimedia*, pages 1125–1128, 2011. 2

[4] Sudong Cai, Yulan Guo, Salman Khan, Jiwei Hu, and Gongjian Wen. Ground-to-aerial image geo-localization with a hard exemplar reweighting triplet loss. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8391–8400, 2019. 1, 2

[5] Francesco Castaldo, Amir Zamir, Roland Angst, Francesco Palmieri, and Silvio Savarese. Semantic cross-view matching. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 9–17, 2015. 2

[6] David M Chen, Georges Baatz, Kevin Köser, Sam S Tsai, Ramakrishna Vedantham, Timo Pylvänäinen, Kimmo Roimela, Xin Chen, Jeff Bach, Marc Pollefeys, et al. City-scale landmark identification on mobile devices. In *CVPR 2011*, pages 737–744. IEEE, 2011. 1

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2, 3

[9] James Hays and Alexei A Efros. Im2gps: estimating geographic information from a single image. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2008. 1

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4, 5

[11] Sixing Hu, Mengdan Feng, Rang MH Nguyen, and Gim Hee Lee. Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7258–7267, 2018. 2, 5, 6

[12] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 4, 5, 7, 8

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 2, 7

[14] Chuan Li and Michael Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *European conference on computer vision*, pages 702–716. Springer, 2016. 4

[15] Tsung-Yi Lin, Serge Belongie, and James Hays. Cross-view image geolocalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 891–898, 2013. 2

[16] Tsung-Yi Lin, Yin Cui, Serge Belongie, and James Hays. Learning deep representations for ground-to-aerial geolocalization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5007–5015, 2015. 1, 2

[17] Liu Liu and Hongdong Li. Lending orientation to neural networks for cross-view geo-localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5624–5633, 2019. 1, 2, 6

[18] Xiaohu Lu, Zuoyue Li, Zhaopeng Cui, Martin R Oswald, Marc Pollefeys, and Rongjun Qin. Geometry-aware satellite-to-ground image synthesis for urban areas. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 859–867, 2020. 3

[19] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*, 2015. 7

[20] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018. 4

[21] Arsalan Mousavian and Jana Kosecka. Semantic image based geolocation given a map. *arXiv preprint arXiv:1609.00278*, 2016. 2

[22] Krishna Regmi and Ali Borji. Cross-view image synthesis using conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3501–3510, 2018. 2, 3, 6, 7, 8

[23] Krishna Regmi and Ali Borji. Cross-view image synthesis using geometry-guided conditional gans. *Computer Vision and Image Understanding*, 187:102788, 2019. 2

[24] Krishna Regmi and Mubarak Shah. Bridging the domain gap for ground-to-aerial image matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 470–479, 2019. 2, 3, 6

[25] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 4

[26] Torsten Sattler, Michal Havlena, Konrad Schindler, and Marc Pollefeys. Large-scale location recognition and the geometric burstiness problem. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1582–1590, 2016. 1

[27] Grant Schindler, Matthew Brown, and Richard Szeliski. City-scale location recognition. In *2007 IEEE Conference on*

Computer Vision and Pattern Recognition, pages 1–7. IEEE, 2007. 1

[28] Yujiao Shi, Liu Liu, Xin Yu, and Hongdong Li. Spatial-aware feature aggregation for image based cross-view geo-localization. In *Advances in Neural Information Processing Systems*, pages 10090–10100, 2019. 1, 2, 3, 5, 6

[29] Yujiao Shi, Xin Yu, Dylan Campbell, and Hongdong Li. Where am i looking at? joint location and orientation estimation by cross-view matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4064–4072, 2020. 1, 2, 3, 6

[30] Yujiao Shi, Xin Yu, Liu Liu, Tong Zhang, and Hongdong Li. Optimal feature transport for cross-view image geo-localization. In *AAAI*, pages 11990–11997, 2020. 1, 6

[31] Akihiko Torii, Relja Arandjelovic, Josef Sivic, Masatoshi Okutomi, and Tomas Pajdla. 24/7 place recognition by view synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1808–1817, 2015. 1

[32] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 4

[33] Nam N Vo and James Hays. Localizing and orienting street views using overhead imagery. In *European conference on computer vision*, pages 494–509. Springer, 2016. 1, 2

[34] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 7

[35] Scott Workman and Nathan Jacobs. On the location dependence of convolutional neural network features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 70–78, 2015. 2

[36] Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-area image geolocalization with aerial reference imagery. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3961–3969, 2015. 1, 2

[37] Amir Roshan Zamir and Mubarak Shah. Accurate image localization based on google maps street view. In *European Conference on Computer Vision*, pages 255–268. Springer, 2010. 1

[38] Amir Roshan Zamir and Mubarak Shah. Image geo-localization based on multiplenearest neighbor feature matching usinggeneralized graphs. *IEEE transactions on pattern analysis and machine intelligence*, 36(8):1546–1558, 2014. 1

[39] Menghua Zhai, Zachary Bessinger, Scott Workman, and Nathan Jacobs. Predicting ground-level scene layout from aerial imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 867–875, 2017. 3, 6, 7, 8

[40] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 7

[41] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. 7

[42] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495, 2014. 2

[43] Xinge Zhu, Zhichao Yin, Jianping Shi, Hongsheng Li, and Dahua Lin. Generative adversarial frontal view to bird view synthesis. In *International conference on 3D Vision (3DV)*, pages 454–463. IEEE, 2018. 2