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

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

A. Architecture details

We outlined the general structure of our model in the main paper. Here, we provide more technical details about our network in terms of the generator G, discriminator D, and the retrieval R branches.

Generator. The generator G is constructed as a U-Net [9] architecture consisting of residual blocks [3]. This design is inspired by existing image-to-image translation models [1, 14, 7] with similar architectures. We provide the exact components of our generator in Table 1. For a given satellite image, G uses residual downsampling blocks (see Figure 1) that in total reduce the spatial size by a factor of 64. On the reduced resolution, our bottleneck refines the features with 6 residual blocks, see Figure 2. The last part of G then consists of residual upsampling blocks (see Figure 3) that mirror the downsampling blocks from before, such that the inputs and outputs of G have the same spatial size. Additionally, we add a self-attention (non-local) block [12] after the first upsampling block. This strategy helps to learn global dependencies in the image, as was shown in prior work [14, 1]. Finally, we use an instance normalization layer (IN) [10] after each upsampling and downsampling block and spectral normalization [8] after each convolution layer.

Discriminator. The discriminator D is designed as a modified PatchGAN classifier [4], see Table 2. In this context, the satellite-street input pair comprises the *real* input whereas the *fake* input consists of the input satellite and generated street $G(I_{ps})$ images. Similarly to the generator, we use non-local self-attention blocks on the 28x154 resolution. We also again use spectral normalization [8] after each convolution layer, which regularizes each individual set of features to a spectral radius of 1. Note, that spectral normalization is not used after the last convolution layer of both the generator and discriminator.

Retrieval. The retrieval branch R of our network determines the corresponding satellite image for a given street panorama. This is done by finding a global feature encoding for both satellite and street view input images. The local features for the satellite inputs are not computed here, because we can reuse the features from the encoder part of the generator $G_E(I_{ps})$. Specifically, we use the output from *the last residual block in the bottleneck*. In order to obtain an equivalent set of features for the street-view input images, we use a modified ResNet34 [3] feature extractor. This yields a set of features for both inputs with the same spatial and channel sizes. Afterward, we convert the local features of both inputs to global descriptors by using the spatial attention module SA which we explain in the main paper.

Generator
Satellite (3, 112, 616)
Conv (32, 112, 616) + IN
(enc1) Resblock Down (64, 56, 308) + IN
(enc2) Resblock Down (128, 28, 154) + IN
(enc3) Resblock Down (256, 14, 77) + IN
Resblock (256, 14, 77) x 6
+ *concat (enc3)* Resblock Up (128, 28, 154) + IN
+ *concat (enc2)* Non-local Block (256, 28, 154)
Resblock Up (64, 56, 308) + IN
+ *concat (enc1)* Resblock Up (64, 112, 616) + IN
Conv (3, 112, 616) + Tanh

Table 1: Exact technical specifications of our generator G.

Discriminator
Satellite+Street (6, 112, 616)
4x4 Conv + LeakyReLU(0.1) (64, 56, 308)
4x4 Conv + LeakyReLU(0.1) (128, 28, 154)
Non-local Block (128, 28, 154)
4x4 Conv + LeakyReLU(0.1) (256, 14, 77)
4x4 Conv + LeakyReLU(0.1) (512, 14, 76)
4x4 Conv (1, 14, 75)

Table 2: Exact structure of our discriminator module D.

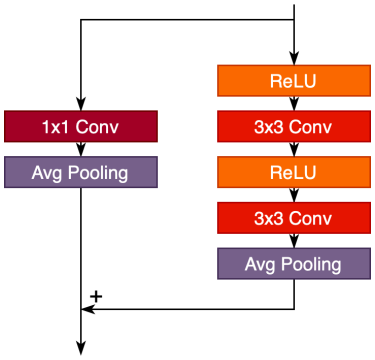

Figure 1: Residual downsampling blocks.

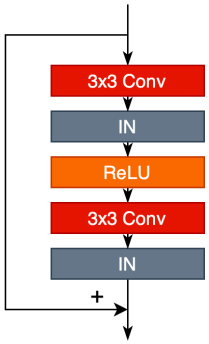

Figure 2: Residual blocks.

B. Implementation details

We implement our network in PyTorch using Adam optimizer [5]. The momentum parameters $\beta 1$ and $\beta 2$ are set to 0.5 and 0.999, respectively, and the learning rate for all three networks is set to $1e-4$. The resolution of both the ground images and polar transformed satellite images is 112x616. Furthermore, we normalized the pixel intensity values to the interval [-1, 1]. For the weighted soft-margin loss, we use the exhaustive mini-batch strategy to create the triplets within a batch. For the batch size B (we choose $B = 32$), this strategy sums up the triplet loss for all $2B(B-1)$ combinations of positive and negative pairs, see [11] for more details. Moreover, we use a hard negative mining strategy after the loss converged as a fine-tuning step during training. Specifically, we sort all triplets in the current batch by relevance, i.e., the loss value, and discard a certain percentage of the triplets with the least amount of surplus information. For the spatial aware feature aggregation, we use $k = 8$ attention masks. Finally, we choose the hyper-parameter from our weighted soft margin loss as $\alpha = 10$. During training, we follow the standard protocol for GAN optimization [2]. In particular, we alternate between updating the weights of each network individually with one update step per cycle. The gradients for the generator are coming from both the discriminator and retrieval networks. Additionally, the weights for the individual loss functions are set to: $\lambda_{ret} = 1000$, $\lambda_{L_1} = 100$ and $\lambda_{GAN} = 1$.

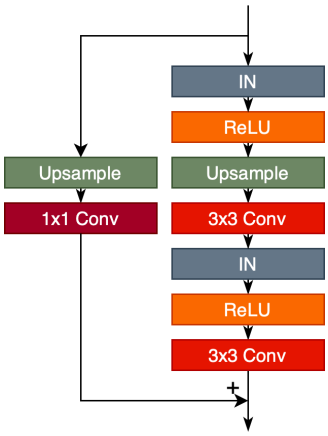

Figure 3: Residual upsampling blocks.

C. Additional qualitative results

Geo-localization. In the localization branch R, our algorithm produces L_2 distances between the features of satellite-street pairs. This means that for a given street image, our algorithm can output a set of the closest satellite matches and rank them according to plausibility. For a given street image, the recall-k retrieval accuracy (R@k) then measures whether the ground-truth satellite pair is among the first k predicted matches. Here, we visualize some of the closest satellite images for a given query street image for examples from CVUSA (see Figure 4) and the CVACT test set (see Figure 5).

Cross-view synthesis. For a more complete picture, we show additional qualitative results for satellite-to-street view synthesis for both considered benchmarks in Figure 6 and Figure 7.

Figure 4: Geo-localization results on CVUSA [13]. For a given query street view (left), we show the closest satellite matches produced by our method. Green boxes denote the ground truth match.

Ablation Study. We now assess the qualitative difference between our full architecture and the ablations **ii.** and **iii.** in Table 3 in the main paper, see Figure 8. Even though the ablation **ii.** only slightly underperforms our main pipeline in terms of the image retrieval task, the synthesized images are significantly less realistic. The $\mathcal{L}_1$ loss itself is sufficient to represent low frequencies, which yields latent features that represent the overall image structure, but lacks high frequency details producing blurry images. The qualitative comparison between ablation **iii.** and our main pipeline illustrates that $\mathcal{L}_{cGAN}$ synthesizes photo-realistic street-views. On the other hand, the qualitative difference between them highlights the importance of using the latent representation in our full architecture.

Figure 5: Geo-localization results on the CVACT [6] test set. For each street image, we show the closest five matches predicted by our method, as well as the ground truth match (green box). Since the coverage of the CVACT dataset is quite dense, there are at times multiple matches that are considered correct, as long as the distance to the ground truth match is less than 5 meters (e.g. in the last row, the second satellite image is also a correct match). Our method consistently retrieves not only the closest match, but also multiple images in the same region which yields a robust localization performance.

Figure 6: Qualitative cross-view synthesis examples on the CVUSA dataset.

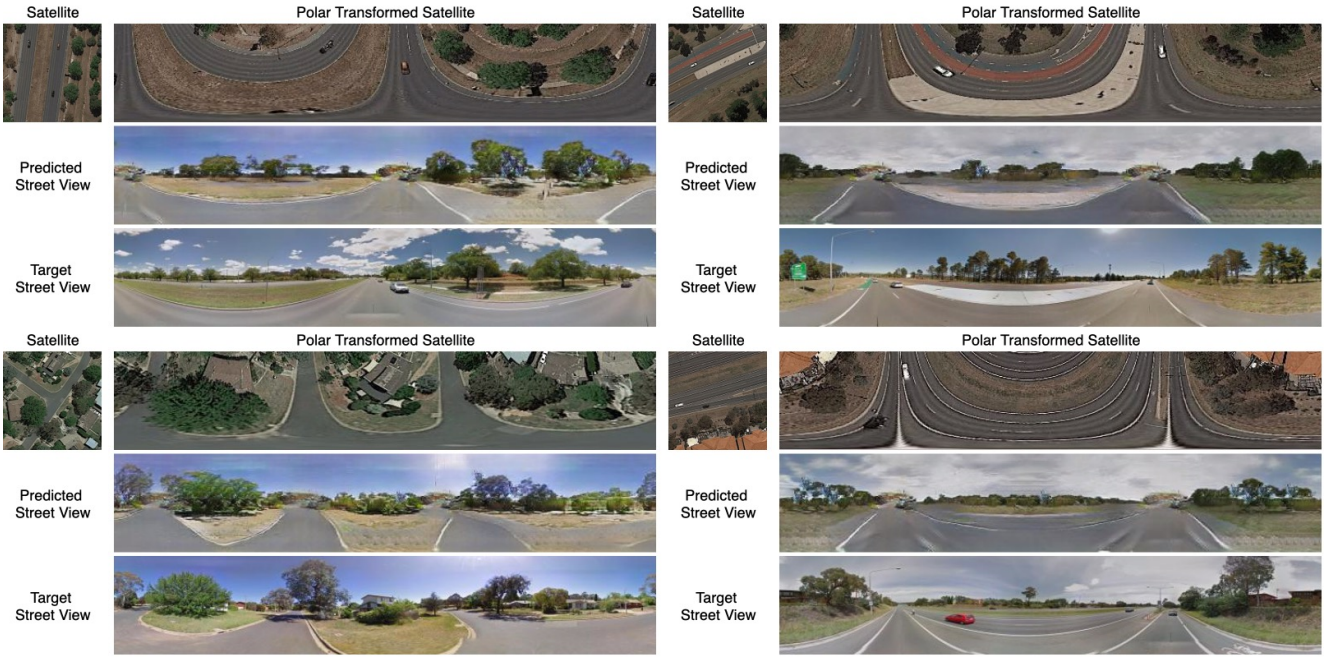

Figure 7: Qualitative examples on the CVACT dataset. The two examples on the left side are from the validation set and the ones on the right side from the large-scale test set.

Figure 8: Qualitative comparisons corresponding to the results from our ablation study in Table 3 in the main paper. We show the synthesized images for three different examples for ablations **ii.** and **iii.**, our method and the ground truth street views.

References

[1] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 1

[2] 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] 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. 1

[4] 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. 1

[5] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 2

[6] 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. 4

[7] Maxim Maximov, Ismail Elezi, and Laura Leal-Taixé. Ciagan: Conditional identity anonymization generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5447–5456, 2020. 1

[8] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018. 1

[9] 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. 1

[10] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 1

[11] 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. 2

[12] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 1

[13] 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

[14] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International Conference on Machine Learning*, pages 7354–7363. PMLR, 2019. 1