Can Location Embeddings Enhance Super-Resolution of Satellite Imagery? Supplementary Material

Daniel Panangian* Ksenia Bittner
The Remote Sensing Technology Institute
German Aerospace Center (DLR), Wessling, Germany
{daniel.panangian, ksenia.bittner}@dlr.de

S1. Dataset

In our experiments, we utilize only the RGB channels from Sentinel-2 imagery, simulating true color imagery (TCI) using Sentinel Hub's L1C Optimized Script[1]. To refine the training data, we apply a filtering step using WorldCover segmentation labels from the S2-NAIP dataset [3], ensuring that the model is trained on diverse urban environments. Specifically, only tiles with at least 30% urban coverage are included, resulting in approximately 4,442 tiles in the training set and 95 tiles in the validation set. The test set comprises around 1,883 tiles, selected to evaluate the model's ability to generalize to regions beyond the United States by incorporating diverse global locations such as forests, mountains, and urban landscapes. Additionally, during inference, we apply the model to various cities outside the United States, as detailed in Sec. S3. This comprehensive evaluation ensures the model's robustness across varying geographical and environmental contexts, which is critical for the building segmentation task.

S2. Implementation Details

S2.1. Baselines and Comparisons

We use SatLAS-SR [3] as our primary baseline, a model that extends ESRGAN [2] by integrating CLIP loss, an object-aware discriminator, and a feature extractor from a foundation model for remote-sensing super-resolution. SatLAS-SR achieved better results when using multiple Sentinel-2 inputs of the same area taken in different time, but for our experiments, we only use a single Sentinel-2 image that we focus on utilizing location-based features rather than than incorporating temporal data. we also incorporate CLIP loss and the object-aware discriminator, as these components demonstrated the most significant improvements in the paper. To further compare with current state-of-the-art methods, we fine-tune Stable Diffusion [1] for the super-resolution task using ControlNet. Stable Diffusion traditionally generates images using text prompts, while ControlNet [4] allows for an additional input, such as an image, to guide the generation process. This conditioning mechanism enables greater control over the output by leveraging input images like edges, depth maps, or other structural cues, which the model uses alongside the text prompt to shape the generated image. For a fair comparison, we fine-tune Stable Diffusion on our dataset and use a dummy text prompt for each sample to simulate similar input conditions across all models. Given that Stable Diffusion operates on 512×512 pixel images, we first upsample both Sentinel and NAIP imagery to 512×512 before downscaling them back to their original resolution after processing.

S2.2. Training Details

For all experiments, our model is implemented in PyTorch and runs on an NVIDIA A100-SXM4 GPU with a batch size of 16. For training hyperparameters, we follow Satlas-SR [3]. All models are trained from scratch using the Adam optimizer, with the learning rate initialized to 10^{-4}. For both the generator and discriminator, we employ the *large* variant mentioned in the SatLAS-SR paper, which includes 256 features, 128 grow channels, and 30 blocks. For our location-matching discriminator and location embedding features, we use 64 features. The loss function is a weighted combination of several components: the pixel loss $\lambda_{\text{pix}} = 1.0$, perceptual loss $\lambda_{\text{perceptual}} = 1.0$, GAN loss $\lambda_{\text{GAN}} = 0.1$, CLIP loss $\lambda_{\text{CLIP}} = 1.0$, OpenStreetMap loss $\lambda_{\text{OSM}} = 0.3$, and our location-matching loss $\lambda_{\text{loc\_match}} = 1.0$.

*Corresponding author

[1] L1C Optimized Script

S3. Qualitative results

We performed super-resolution inference over a large area in Malmö, Denmark to assess the performance of different methods at scale. The comparison includes four sets of figures: the low-resolution Sentinel-2 input (Fig. 1), outputs from Satlas-SR (Fig. 2), Stable Diffusion (SD) + ControlNet (Fig. 3), and our method (Fig. 4). Starting with SD + ControlNet, while it offers some improvement in generating finer details, it suffers from significant inconsistencies in both color and texture across patches. These inconsistencies become especially problematic when stitching the patches together, resulting in a disjointed appearance with noticeable blocky patterns. The overall image looks fragmented, as if each patch comes from a completely different area. This lack of cohesion is particularly disruptive in regions with homogeneous textures, such as agricultural fields or open urban areas, where smooth transitions are expected. In the other hand, Satlas-SR and our method show much more consistent color and texture across patches. However, Satlas still displays subtle block patterns. These patterns arise because the method does not share sufficient context between patches, leading to noticeable tiling effects in areas where the texture is expected to remain constant. Our method, on the other hand, effectively overcomes this limitation by ensuring smooth transitions between patches, preserving both color and texture consistency throughout the entire scene. This leads to a more visually cohesive result, without the blocky artifacts seen in the other methods.

A closer inspection reveals several challenges in applying super-resolution independently to each patch, as seen in the results from Satlas-SR and ours in 5. For example, in urban residential areas, we observe that roads are often interrupted between patches, creating discontinuities that disrupt the visual flow of the scene. Similarly, in industrial zones, large structures like buildings can become incomplete, with parts missing in certain patches. In suburban neighborhoods, several buildings appear fragmented or distorted, breaking the continuity of the image. Our method addresses these inconsistencies by using information from neighboring patches to maintain coherence across the entire area.

S4. Ablation study

We analyze the impact of adding self-attention and location embeddings (via cross-attention) on the model's performance. We compare four configurations: a baseline model (Satlas-SR), a model with self-attention, a model with both self-attention and location embeddings, and the final model, which includes a location matching discriminator. The performance metrics for these models, including PSNR, SSIM, LPIPS, and CLIP Score, are shown in Table 1.

The baseline model exhibits balanced performance across metrics. Adding self-attention in the model leads to improvements, with slightly higher PSNR, SSIM, and CLIP scores, indicating better image quality and semantic alignment. However, the inclusion of location embeddings results in a trade-off. While the PSNR and SSIM drop to 14.5957 and 0.2453 respectively, the CLIP score improves to 0.9373. Finally, the addition of the location matching discriminator in the final model slightly decreases all metrics although the CLIP score remains relatively high at 0.9261.

This results reveals that while the addition of location embeddings and a location matching discriminator enhances semantic alignment, it introduces trade-offs in traditional image quality metrics.

Model	**PSNR**	**SSIM**	**LPIPS**	**CLIP Score**
ESRGAN + CLIP+ OSM	15.6481	0.2621	0.4786	0.9227
+ Self-Att.	15.7768	0.2663	0.4624	0.9334
+ Location	14.5957	0.2453	0.4668	0.9373
+ Loc. Disc	14.3973	0.2221	0.4855	0.9261

Table 1. Ablation study results of different model configurations.

S5. Location control

In this section, we evaluate the performance of the location control feature in our model, which incorporates geographic coordinates into the super-resolution process. To test this feature, we applied SR to the same input image while varying the coordinates used in the location embeddings. We compared the results generated using the original coordinates of the input image with outputs produced using coordinates from different cities in USA, as shown in Figure 6. These locations were specifically chosen for their distinct geographic and environmental characteristics. While the model effectively maintains the overall content of the images (i.e., shapes, textures, and spatial arrangement of objects), the only noticeable change when altering the coordinates is in the hue of the images. The structural elements, such as building shapes and road layouts, remain consistent across different locations. However, this limited change in image characteristics reveals a limitation of the current location control mechanism: it does not fully capture the regional variations that would be expected when changing geographic context. These variations could include differences in architectural styles, natural features, and other location-specific characteristics, which are not reflected in the generated images.

References

[1] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. 2021. 1

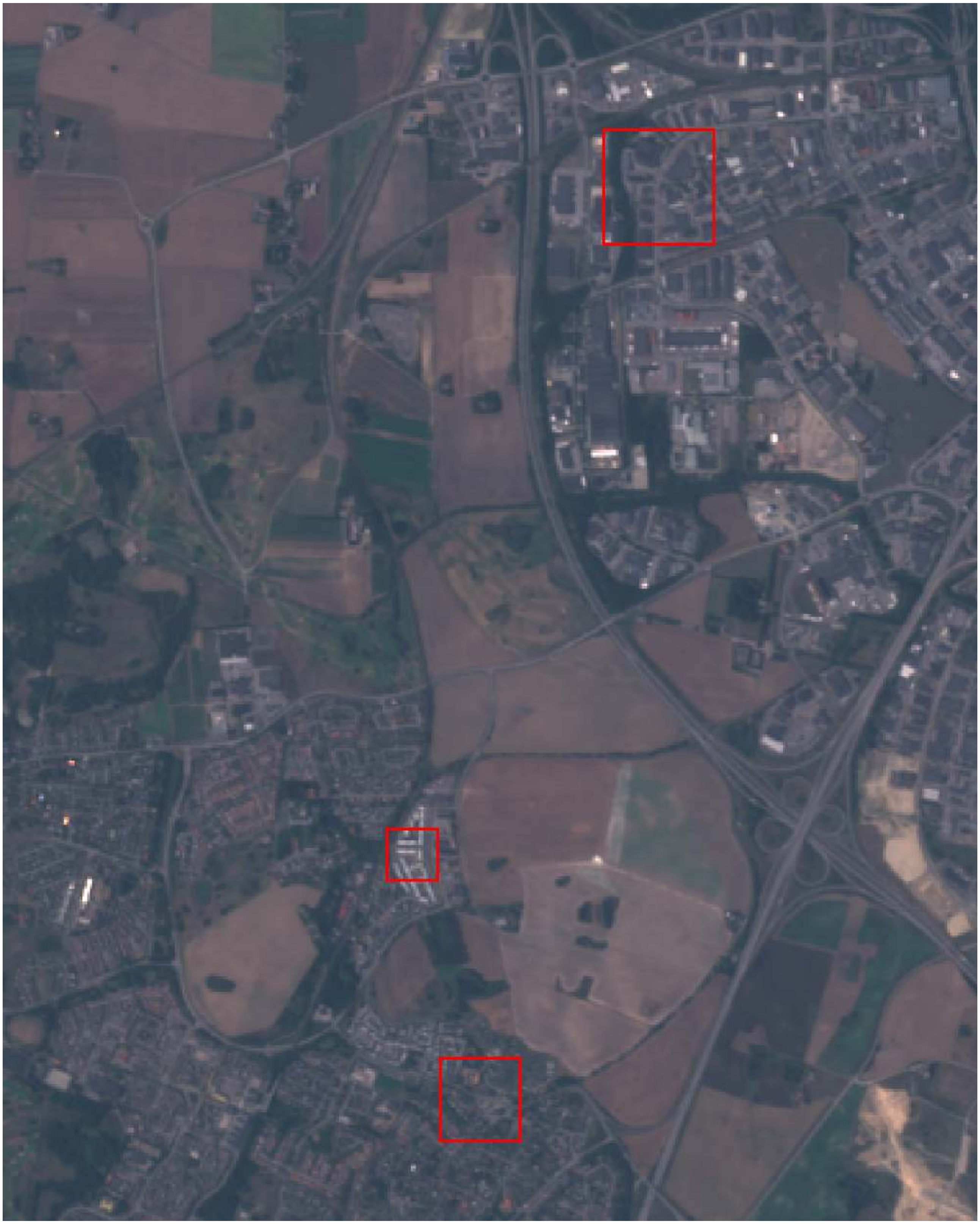


Figure 1. Sentinel-2 input image of the Malmö region

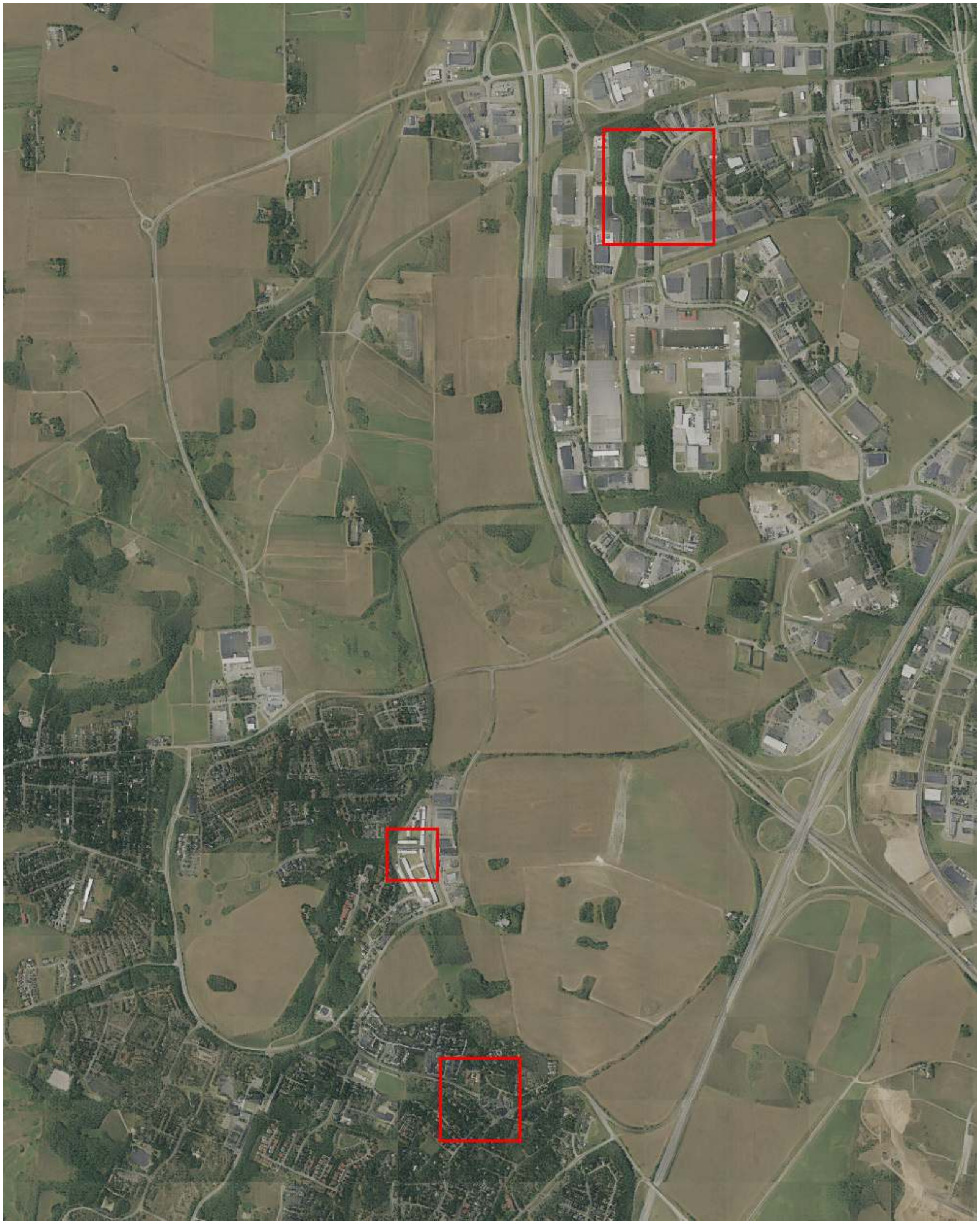


Figure 2. Super-resolution output from Satlas-SR for the Malmö region

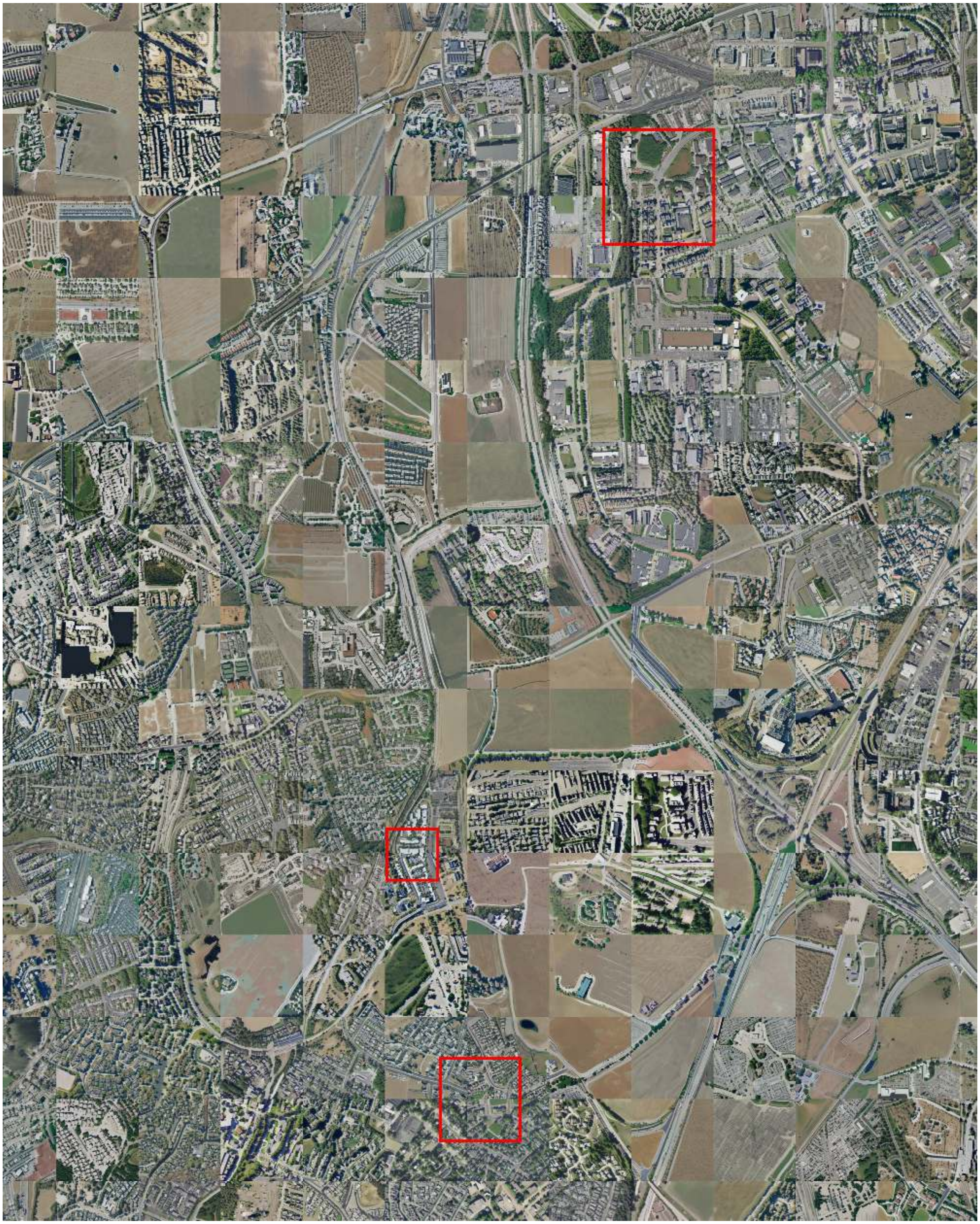


Figure 3. Super-resolution output from Stable Diffusion with ControlNet for the Malmö region

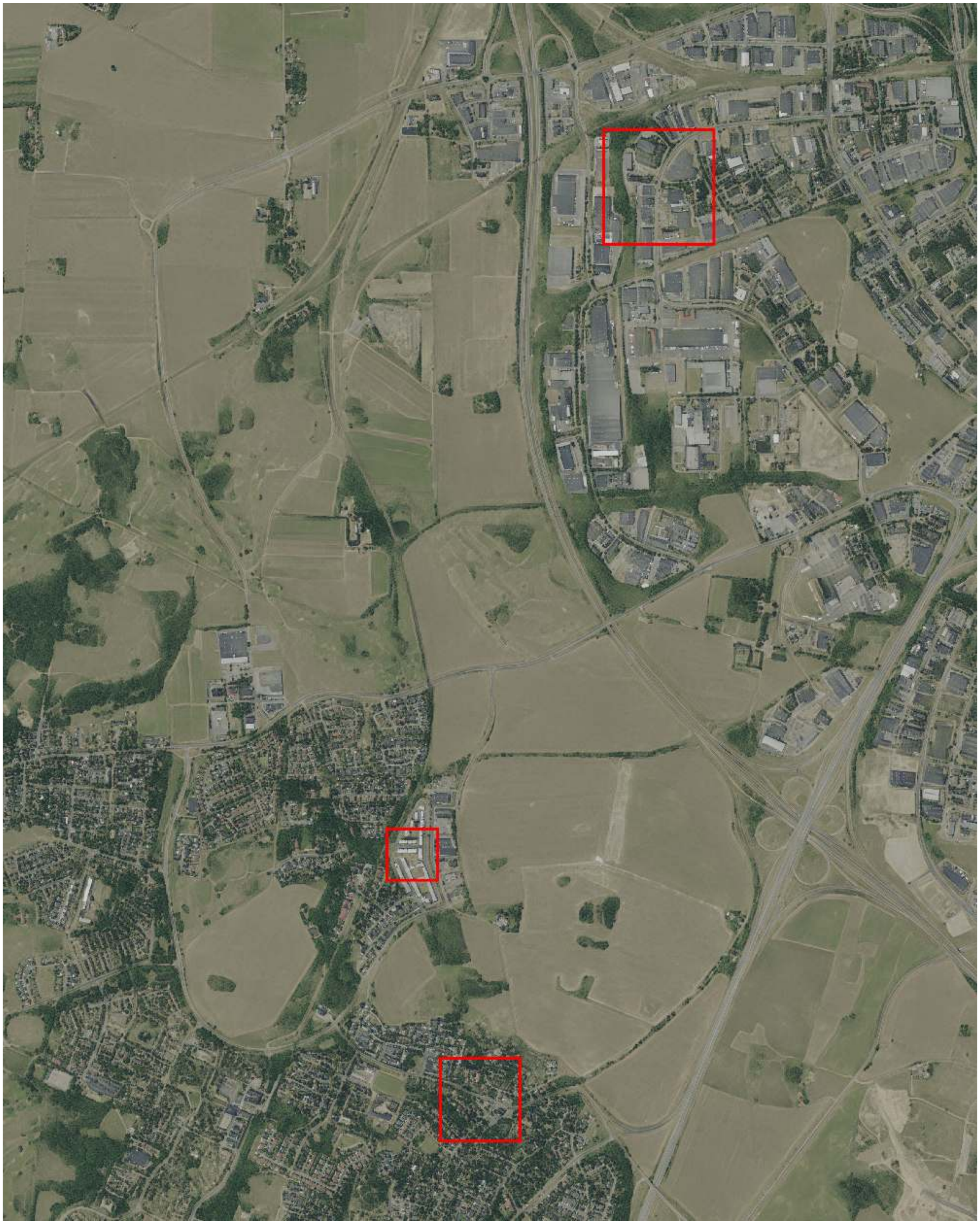
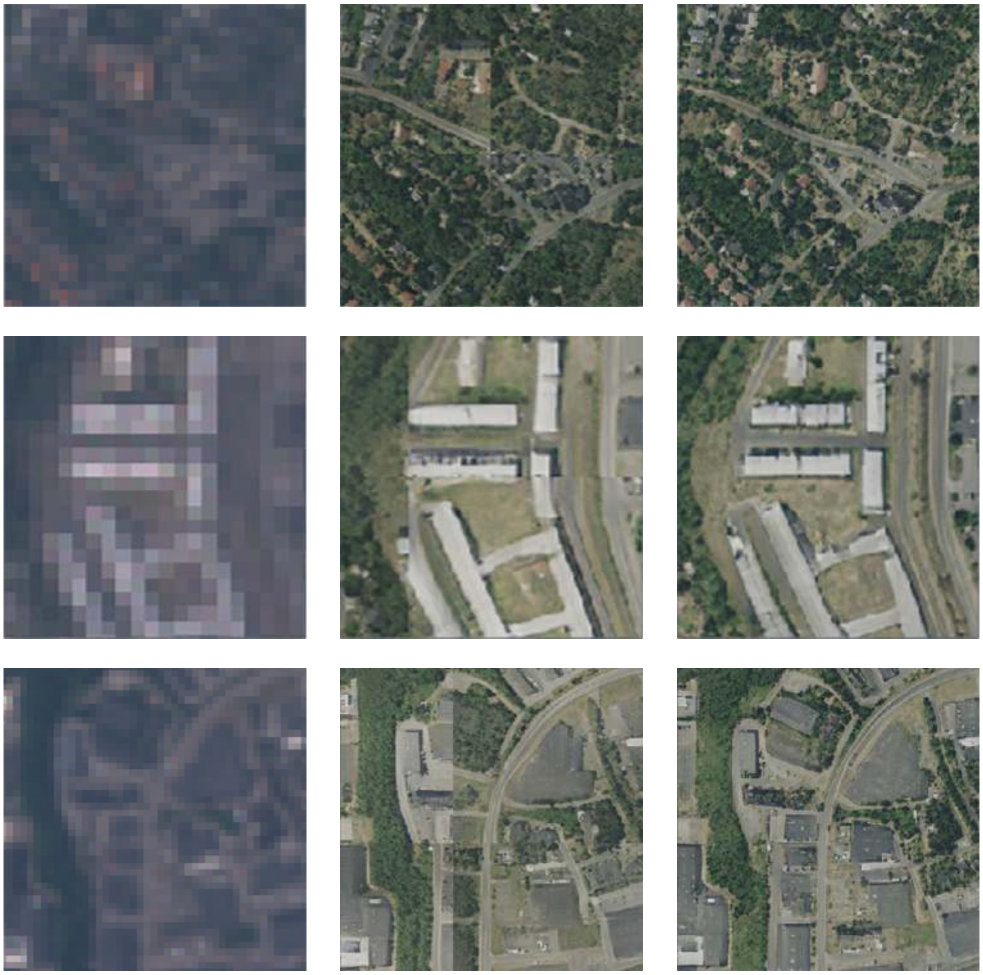


Figure 4. Super-resolution output from our method for the Malmö region



(a) Sentinel-2

(b) Satlas-SR

(c) Ours

Figure 5. Comparison of super-resolution outputs for regions of interest from the Malmö inference

[2] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Chen Change Loy, Yu Qiao, and Xiaoou Tang. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018. 1

[3] Piper Wolters, Favyen Bastani, and Aniruddha Kembhavi. Zooming out on zooming in: Advancing super-resolution for remote sensing, 2023. 1

[4] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. 1

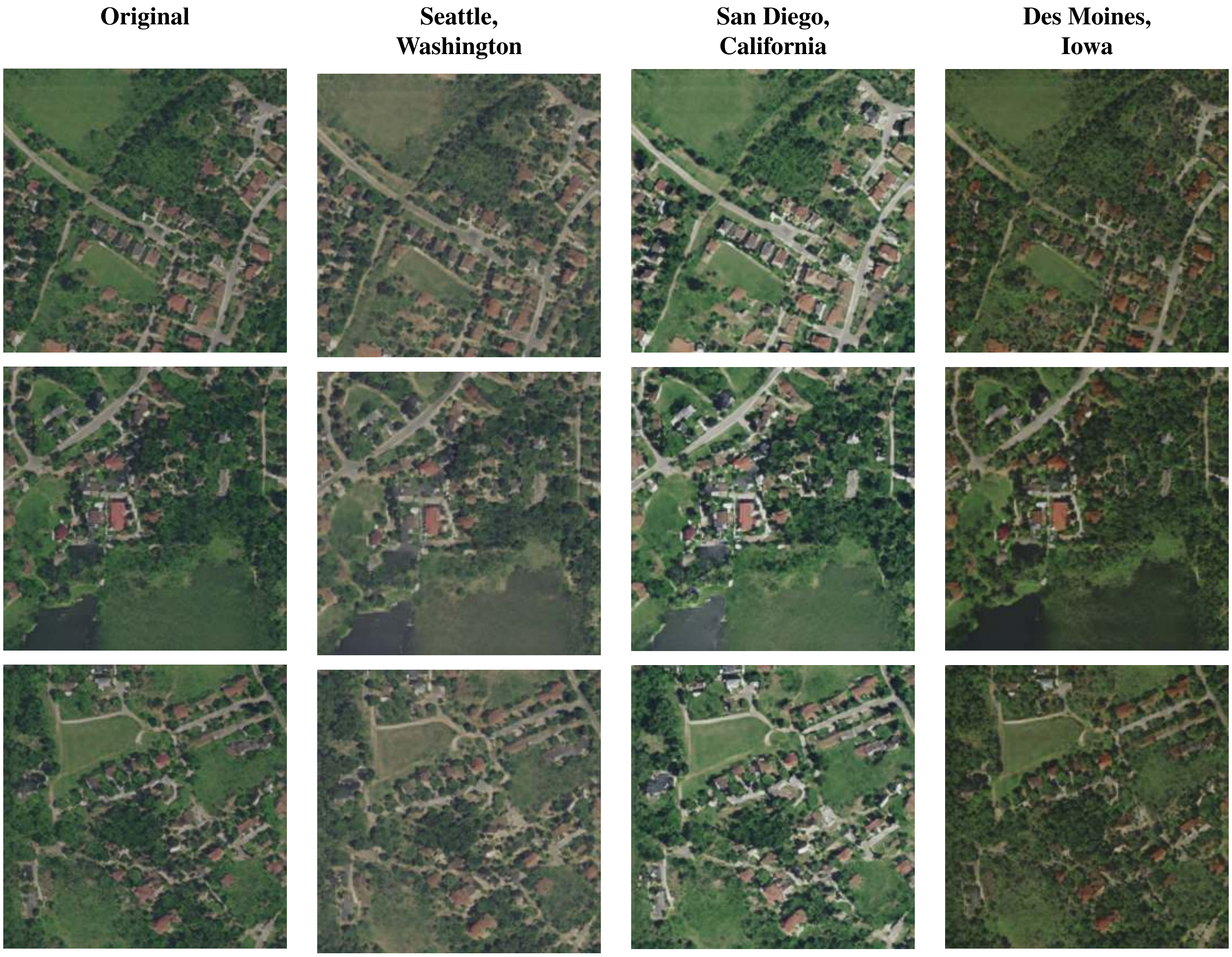


Figure 6. Super-resolution results using different geographic coordinates. The first column shows the output when using the original coordinates from an area in Germany. The subsequent columns show the results of applying super-resolution with different coordinates in USA.