ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XI-2-2026
XXV ISPRS Congress 2026 “From Imagery to Understanding”, Commission II, 4–11 July 2026, Toronto, Canada

From Aerial to Satellite: Can Super-Resolution Enable Label-Free Model Transfer?

Nina Merkle, Corentin Henry, Sandeep Kumar Jangir, Jens Hellekes, Felix Rauch, Pablo d’Angelo, Franz Kurz

Remote Sensing Technology Institute, German Aerospace Center (DLR), Oberpfaffenhofen, Germany
{nina.merkle, corentin.henry, sandeep.jangir, pablo.angelo, franz.kurz}@dlr.de

Keywords: Super-Resolution, Deep Learning, Aerial Imagery, Road Segmentation, Satellite Imagery, Remote Sensing

Abstract

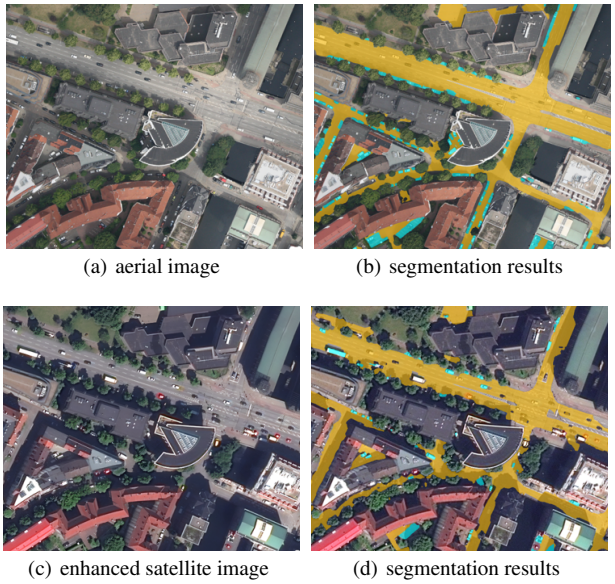
Satellite imagery enables large-scale remote sensing applications by providing frequent and large-scale coverage. However, its limited spatial resolution often restricts the use of satellite images in tasks that require detailed, fine-scale information. In contrast, aerial images offer a much higher spatial resolution, allowing the extraction of fine-grained features, but typically cover smaller, more localized areas. In this work, we investigate whether super-resolution (SR) methods can bridge the gap between aerial and high-resolution satellite imagery, enabling a label-free model transfer, meaning without fine-tuning our model with additional manual annotations. The idea is to enhance the spatial resolution of high-resolution satellite images, allowing models trained on aerial data to be directly applied to satellite images. Towards this goal, a state-of-the-art SR algorithm is used to upscale three high-resolution satellite images, matching the resolution of the aerial training data. Then, a segmentation network trained on an aerial image dataset is applied to segment roads and parking areas in the super-resolved satellite images. The approach is evaluated on an annotated dataset and compared to the results in the original satellite images. Additionally, we investigate its performance on a low-resolution aerial image. Our results demonstrate that SR facilitates the utilization of models trained on aerial image datasets for large-scale satellite applications without requiring new labels.

1. Introduction

Remote sensing imagery plays a crucial role in mapping and monitoring the Earth’s surface in a wide range of applications, including urban planning, transportation analysis, and environmental monitoring. Satellite imagery provides global coverage with frequent revisit times, allowing large-scale and long-term monitoring. However, the spatial resolution of most satellite sensors remains lower than that of aerial imagery, which limits the ability to extract fine-grained features such as small roads, parking areas, and detailed building attributes from these data. In contrast, aerial images provide significantly higher spatial resolution, enabling the extraction of fine-grained features, but generally cover smaller and more localized areas. Although both data sources provide complementary information, effectively combining them remains challenging.

A key challenge in integrating both aerial and satellite imagery in a single analysis pipeline lies in the domain gap between them. Differences in spatial resolution, viewing geometry, illumination conditions, and sensor characteristics cause models trained on aerial data to perform poorly when applied directly to satellite imagery (Peng et al., 2022). While domain adaptation methods can mitigate this issue, they often require additional annotated data or complex retraining pipelines (Lyu et al., 2025). Moreover, creating new pixel-accurate annotations for satellite images is time-consuming and costly.

Recent advances in deep learning-based super-resolution (SR) have shown impressive capability in recovering fine spatial details from coarse-resolution imagery. SR models can effectively enhance texture, edges, and structural features, narrowing the perceptual gap between low- and high-resolution domains. The evolution of single-image super-resolution (SISR) began with convolutional neural networks (CNNs) like SRCNN (Dong et al., 2014) focused on pixel-wise accuracy. This paradigm shifted with GANs, such as SRGAN (Ledig et al., 2017), which prioritized perceptual realism using adversarial losses. A key challenge remained: generalizing to real-world images with complex, unknown degradations. Blind SR (BISR) methods, notably Real-ESRGAN (Wang et al., 2021), addressed this by training on randomized degradation models, though often at the cost of overly smooth results. Concurrently, new architectures emerged. Transformer-based models, popularized by SwinIR (Liang et al., 2021), have been refined for efficiency and better feature aggregation in recent works, such as by using multi-range attention (Xie et al., 2025) or progressive feature filtering (Long et al., 2025). In parallel, diffusion models, such as SR3 (Saharia et al., 2022), now offer state-of-the-art generative fidelity. However,



(a) aerial image (b) segmentation results

(c) enhanced satellite image (d) segmentation results

Figure 1. Results of ■ road and ■ parking area segmentation from an aerial image and an enhanced WorldView-3 satellite.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-493-2026 | © Author(s) 2026. CC BY 4.0 License.

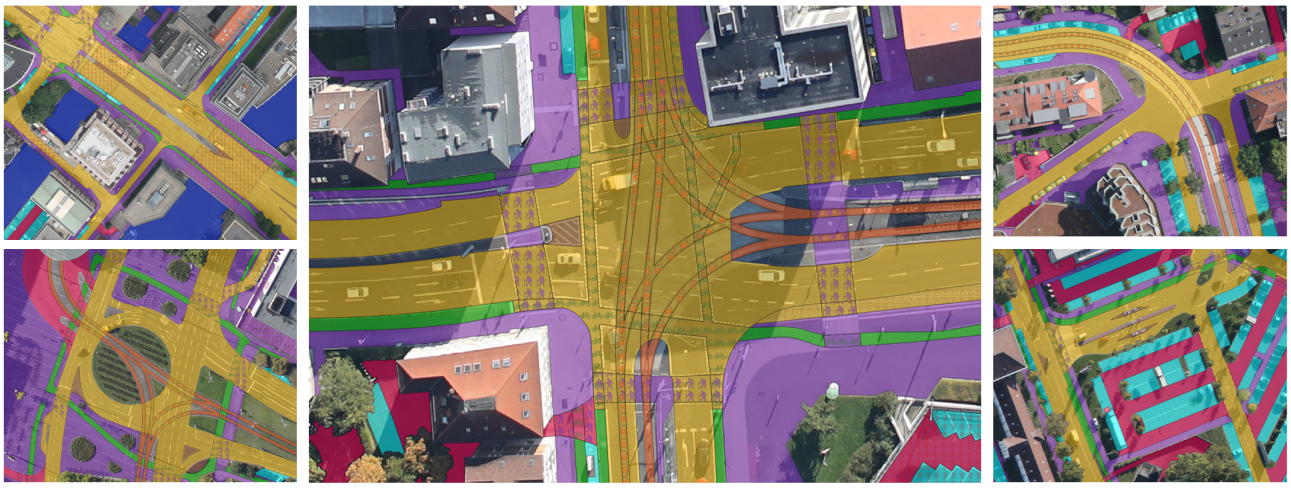


Figure 2. Example areas from the TIAS dataset with overlaid labels. Category colors are: parking area, road, access way, footway, bikeway, railroad bed, keep-out area, road shoulder, and water.

these diffusion models struggle to generalize to the diverse, real-world artifacts common in remote sensing. A trade-off persists between robust generalization to artifacts and the generation of high-fidelity detail, which the U2D2 framework (Jangir et al., 2025) aims to resolve. U2D2 is a multi-stage SR framework specialized for aerial and satellite imagery. It integrates the advantages of blind SR models with the image refinement capabilities of diffusion models. This hybrid, multi-stage approach achieves superior image quality when compared to conventional single-stage, end-to-end alternatives.

In this work, we investigate whether SR methods can bridge the gap between aerial and high-resolution satellite imagery, enabling a label-free model transfer. The idea is to enhance the ground sampling distance of high-resolution satellite images, allowing models trained on aerial data to be directly applied to satellite images. Towards this goal, the state-of-the-art SR algorithm U2D2 is used to upscale three high-resolution satellite images and low resolution aerial images, matching the resolution of the aerial training data (10 cm/px). Then, a segmentation network trained on the TIAS aerial image dataset (Merkle et al., 2024) is applied to segment roads and parking areas in the super-resolved images. Our approach is evaluated on an annotated dataset and compared against the results obtained on the original satellite images.

Our results demonstrate that SR methods enable a label-free model transfer between aerial and high-resolution satellite images. Specifically, SR effectively bridges the spatial resolution gap, allowing models trained on aerial data to achieve improved segmentation performance on high-resolution satellite imagery without requiring additional annotations (see illustration in Figure 1). Our experiments conducted on three test sites in Hamburg, Germany, highlight the potential of SR for efficient and scalable large-scale satellite mapping applications.

2. Dataset

In order to evaluate the potential of our super-resolution algorithm for a label-free model transfer, we use a dataset consisting of aerial images with corresponding ground truth data, OpenStreetMap (OSM) masks used as priors, satellite images and digital orthophotos acquired over the city of Hamburg in Germany. The individual data sources are described in detail in the following.

Aerial Images and Ground Truth Data: The aerial images and the corresponding ground-truth annotations used in this study are derived from the TIAS dataset (Merkle et al., 2024). TIAS is a novel dataset consisting of 51 aerial images with high-fidelity, fine-grained labels of traffic areas. The dataset provides a detailed representation of urban environments from a transportation perspective, supporting the reconstruction of traffic networks for motorized road vehicles, cyclists, pedestrians, and rail systems.

The images contained in the TIAS dataset were acquired over the German cities of Berlin, Brunswick, Cologne, Garmisch-Partenkirchen, Hamburg, Landsberg, Kaufbeuren, Munich, Münster, Oldenburg, and Wolfenbüttel. Of the 51 images, 45 are captured using the 3K (Kurz et al., 2012) and 4k (Kurz et al., 2014) camera systems of the German Aerospace Center (DLR) and have a Ground Sampling Distance (GSD) between 6 and 14 cm/px. The remaining 6 are ortho-projected aerial images with a GSD of 10 cm/px. The image sizes range from 17 to 22 Mpx.

The traffic areas in the TIAS dataset are categorized into nine classes: parking area, road, access way, footway, bikeway, railroad bed, keep-out area, road shoulder, and water. Additional attributes capture topological or contextual information, indicating whether areas are shared by multiple traffic participants, elevated (e.g., bridges), under construction, or difficult to interpret during annotation. The attribute unsure reflects the confidence in each annotation, used in case the annotator could not make decision about a clearly-visible yet hard-to-define object. Figure 2 illustrates five representative regions together with their corresponding labels from the TIAS dataset.

For this study, the three images from the city of Hamburg contained in the TIAS dataset were used as test data, while the remaining 48 images served for training and validation of the segmentation model. The test images were acquired by the 3K camera system (Kurz et al., 2012) on June 25, 2020. All experiments in this paper focused on three TIAS classes: road, access way, and parking area, where the road and access way classes were merged into a single category. In addition, we extended the

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-493-2026

parking area class to include all areas with the attribute “shared with parking area” such as roads shared with parking area.

OpenStreetMap (OSM) Masks: To leverage additional information, we use OSM masks as an additional input for our segmentation model. Although these masks are sometimes spatially or semantically inaccurate, they provide a useful baseline to distinguish roads and parking lots. The OSM masks are generated by querying the OSM database for ten traffic- and non-traffic-related infrastructure categories for each image extent, then rasterizing them into a 10-channel one-hot-encoded mask: road, parking area, access way, bikeway, footway, railroad bed, water, bridge, gas station, and building. In order to match the OSM rasters with the source, non-orthorectified images, we first estimate the extent of the orthorectified image to determine the region of interest to query. The resulting vector layers are then projected back onto the original images.

Satellite Images: The high-resolution satellite images used in this study were acquired by WorldView-3 and GeoEye-1, both operated by Maxar Technologies (Maxar, 2022). The two WorldView-3 images used have a spatial resolution of 30 cm and were acquired on June 17, 2020 with an 13° off nadir angle (ONA), and on June 24, 2020 with an 27° ONA, respectively. The GeoEye-1 image used in this study has a spatial resolution of 50 cm and was acquired on March 10, 2022 with a 28° ONA. The acquisition dates were selected to match the dates of the corresponding aerial and DOP images as closely as possible.

Digital Orthophotos (DOPs): To test our approach on lower resolution aerial images, we used Digital Orthophotos (DOPs) from the Federal Agency for Cartography and Geodesy (BKG) (BKG, 2025), which are distortion-corrected and georeferenced aerial images that provide an accurate, true-to-scale representation of the Earth’s surface. The surveying authorities of the German federal states release DOPs with varying GSDs. The BKG merges these regional datasets into unified, seamless, nationwide raster products available at ground resolutions of 20 cm (DOP20) and 40 cm (DOP40). For this study, we used DOP20 images acquired on March 7, 2021, over the city of Hamburg, Germany.

3. Methodology

The proposed workflow comprises three main stages. First, aerial and satellite images are co-registered and projected onto a common reference surface (see Section 3.1). Second, a state-of-the-art SR algorithm is applied to enhance the spatial resolution of the satellite imagery, thereby reducing the ground sampling distance discrepancy between the two domains (see Section 3.2). In the final stage, a segmentation network trained on aerial imagery is applied to the super-resolved satellite data (see Section 3.3).

3.1 Data Preparation

In order to enable the evaluation of segmentation results from images of the same scene captured by different sensors and at different times, the images must first be mapped to a common reference surface. The DOP20 images from Hamburg are defined as the reference, to which the other image datasets must be co-registered. The aerial imagery from the 3K camera system (Kurz et al., 2012) comes with a pre-calibrated interior orientation and exterior orientation derived from a GNSS/IMU system. Despite the high quality of the GNSS/IMU data, it is necessary to further refine absolute image orientations to obtain pixel or sub-pixel accurate absolute orientations. For this purpose, ground control points were manually measured in the DOP and aerial image. The 2D map coordinates of the ground control points were taken from the DOPs, whereas the height was derived from the digital surface model (DSM) from Hamburg provided by BKG. Additional tie points were calculated between aerial images and DOPs and transferred to a bundle block adjustment together with the manually-measured points in order to refine the internal and external parameters of the aerial image orientation. These parameters allow both the aerial image rasters as well as the TIAS vector annotations to be projected onto the DSM. Both datasets then match the DOP data perfectly, apart from elevated surfaces such as buildings, trees or bridges. These areas require special considerations during the evaluation, which we cover in Section 4.2.

The co-registration of the satellite images was performed by matching tie points between the different satellite images and the DOP images, similarly to the aerial images. The procedure consists of multiple steps. First, downscaled images are matched using the SIFT operator. For this, we use a simple but effective approach based on RANSAC to verify the pairwise matching, which can successfully handle image pairs with outlier rates of more than 90 %. After outlier removal, the pairwise matches are chained into tracks connecting multiple images, and we perform an initial bundle block adjustment. Using the corrected orientation, tie points are then refined using local least squares matching on the full resolution images, resulting in tie points with an accuracy of up to 0.1 px (d’Angelo, 2013). Using these high quality tie points, we perform a bundle block adjustment again to refine the rational polynomial coefficients (RPC) sensor model using a 0 order bias correction (Grodecki and Dial, 2003). Using these corrections, we orthorectify the satellite images to ensure a good co-registration of satellite and airborne imagery.

3.2 Satellite Image Super-Resolution

To enhance the DOP20 and high-resolution satellite images, we use U2D2 (Jangir et al., 2025), a $2\times$ blind SR framework designed to resolve the key trade-off between the robust generalization of degradation-aware models and the high perceptual quality of diffusion models. As demonstrated in (Jangir et al., 2025), U2D2 achieved the best visual and quantitative results among several blind image SR methods compared across various GSDs. It is specifically designed to handle the diverse characteristics of remote sensing data, demonstrating robust performance across a wide GSD range from 10 cm to 1.20 m, and to leverage the respective strengths of degradation-based and diffusion-based models while mitigating their weaknesses. Consequently, it produces sharper boundaries, reduced noise, and better resolvability; these improvements enable downstream tasks like segmentation and object detection to extract better features and increase their overall performance. One such result can be seen in object detection task (Jangir et al., 2025), where the U2D2 enhances the degraded 50 cm GSD images and improves the vehicle detection AP_{50} score from a 25.31% bicubic baseline to 70.70%, outperforming the second-best state-of-the-art method.

U2D2 decomposes this task into a two-stage process. First, a lightweight regression-based CNN, the DL-Upsampler (DLU), performs a robust $2\times$ upsampling. It is trained with a degradation module that simulates realistic noise, blur, and compression, enabling the DLU to output a clean, generalized base image. Second, a conditional diffusion model, the Diffusion Refiner

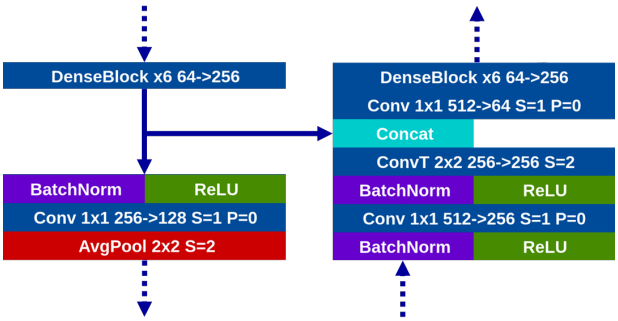


Figure 3. Architecture principle behind Dense-U-Net. On the left, the input to the encoder module is parsed by a dense block and a convolution layer (Conv), then down-sampled by a factor of 2 by an average pooling layer. Here, (S) indicates stride, and (P) indicates padding. On the right, the input to the corresponding decoder module is depth adapted by a convolution layer, is upsampled by a transposed convolution (ConvT), is concatenated with the output from the corresponding encoder module, and is again depth adapted its channel depth by a convolution layer to be inputted to a dense block with the same configuration as its encoder counterpart.

(DiffRef) based on SR3 (Saharia et al., 2022), performs the final refinement. By conditioning on both the clean DLU output and a simple bicubic interpolation of the input, the DiffRef focuses solely on injecting sharp, high-frequency details. This modular approach allows the framework to first clean the image and then separately add realistic texture, leveraging the respective strengths of both model classes.

3.3 Road and Parking Area Segmentation

For the automatic segmentation of roads and parking areas, we use a neural network called Dense-U-Net (Henry et al., 2021a), which is based on the well-known U-Net architecture (Ronneberger et al., 2015). Although the standard U-Net model has been the preferred type of architecture for semantic segmentation tasks in remote sensing applications, thanks to its effectiveness in restoring fine-grained spatial details, many architectures it inspired have used a shallow decoder which did not mirror the depth of the opposite encoder modules, contrary to the original design. This severely limits their fine-grained information extraction capabilities. Unlike other U-Net architectures, Dense-U-Net does not use a simple decoder consisting of consecutive deconvolutions; instead, it mirrors the encoder and thus reverses the inference direction (see Figure 3). The densely connected layers in this network enable faster training, the architecture provides greater capacity for learning complex semantics, and most deep learning libraries offer pretrained parameter sets on ImageNet. DenseNet-121 is a relatively lightweight backbone with few parameters, which reduces the risk of overfitting when training on small datasets such as the TIAS dataset—in contrast to larger models, which tend to overfit the training data and therefore perform worse on new, unseen images. Indeed, Dense-U-Net has often demonstrated its excellent generalization capabilities to diverse areas worldwide (Merkle et al., 2023; Schneibel et al., 2023).

As we want to experiment with extracting additional information from OSM masks, and following several studies which investigated the fusion of multi-modal and multi-temporal data via neural networks (Chlaily et al., 2021; Hong et al., 2021), we explore different data fusion techniques. Merkle et al. (2019) showed the benefits of combining RGB, near infrared (NIR) and thermal infrared (TIR) aerial images for vehicle segmentation and the influence of an early or late fusion within the network. Instead of fusing data from different sensors, Audebert et al. (2017) investigated the inclusion of semantic data, and more specifically tested different network architectures to explore the usefulness of OSM in combination with RGB data for semantic labeling. Eventually we use a modified version of the fusion scheme from FuseNet (Hazirbas et al., 2017) to keep the aerial RGB encoder features separate from the OSM encoder features, where the fusion instead takes place in the skip connections to the decoder. We call it SkipFuse-Dense-U-Net (Henry et al., 2021b). Our intuition is that the fusion of two heterogeneous data types, if done like in FuseNet, will make the optimization of the main encoder harder and lead to lower performance, and therefore both feature map flows should be kept independent from one another.

4. Results & Discussion

In the following, we introduce the training parameters used for the SR approach, as well as the road and parking area segmentation network, in Section 4.1, and describe the metrics used in Section 4.2. The results obtained with the SR approach are presented in Sections 4.3, and the results of the segmentation models are quantitatively and qualitatively evaluated and presented in Section 4.4.

4.1 Trainings Parameters

SR: The U2D2 model (Jangir et al., 2025) used in this study uses its original pre-trained weights and was neither retrained nor fine-tuned on the data used here. The original model was trained on a combination of high-resolution aerial images from the open-source EAGLE dataset (Azimi et al., 2021) (GSDs from 5 to 45 cm) and WorldView-4 RGB images (30 cm GSD). During training, the degradation module dynamically created low-quality pairs with GSDs spanning 10 cm to 1.5 m. The DLU was trained for 500,000 iterations, with a learning rate initialized at 2×10^{-5} and reduced to 2×10^{-6} for the final 50,000 iterations. The Diffusion Refiner (DiffRef) model was trained for 4.5 million iterations, using a 500-step linear noise schedule with levels increasing from 1×10^{-4} to 1×10^{-2}.

Segmentation: We train both Dense-U-Net-121 and SkipFuse-Dense-U-Net-121 on 48 images of the TIAS dataset and test them on the remaining 3. We find optimal hyperparameters via an earlier validation phase: the trainings run for 100 epochs on an 80 GB NVIDIA A100 GPU, with a batch size of 39 and 27 respectively, and a patch size of 512×512 px. We use a cross-entropy loss, an AdamW optimizer, and an initial learning rate of $1e^{-4}$, decaying by a factor of 0.90 if the mean intersection over union (mIoU) evaluated on the non-ortho-projected test images does not increase by at least 0.1 % after 2 epochs (“reduce on plateau” scheduler).

To decrease the false positive rates of the models, we further extend the training set in each epoch by randomly sampling images from two extension datasets containing images with only background pixels, therefore requiring no additional annotation effort. These images were gathered from the countryside of the cities of Hamburg and Dortmund, Germany, and contain challenging examples including crop fields, snowy fields, lakes, and forests. The pool of such data being larger than the TIAS dataset itself, the equivalent of only 20 % of the TIAS training set is randomly sampled from it and added to each epoch, ensuring

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-493-2026 | © Author(s) 2026. CC BY 4.0 License.

Figure 4. Comparison between aerial and satellite images for two example areas in Hamburg, Germany. For the DOP20 and satellite images, the scenes are shown with and without super-resolution (SR).

that background patches are not over-represented in the training, while a different collection of such patches is presented in each epoch.

The non-ortho-projected training, training extension and test images are all resampled to a 10 cm/px GSD via bicubic upsampling prior to training the models, to limit the domain gap between their native resolution and the exclusively 10 cm/px target GSD in our experiments. We also apply uniform random 90° rotation and horizontal flip augmentations.

During inference, the models process the images with a patch size of 3584×3584 px and an overlap of 358 px. In the overlapping areas, the pixels are merged based on extrema logits override: the logit value furthest from 0, either negative or positive, is retained so as to only consider the strongest signal from the model overall (i. e. the most confident value, in layman’s terms). This helps to reduce artifacts on seams where patches meet, as models will typically output less extreme logit values on patch borders due to lacking image context, which will then be overridden by pixels predicted with more extreme logit values using complete image context from the model’s receptive field.

4.2 Evaluation and Metrics

The orthorectification of annotations created on top of off-nadir images causes several issues during the evaluation phase, as they do not always have a perfect correspondence to the underlying DOP20 and satellite images. We will detail them below, before going over the metrics we have chosen for evaluating the performance of our models.

No-data areas: The orthorectified aerial images have an irregular shape due to their projection on a DSM, from which a binary no-data mask can be generated. It is used to mask out no-data areas in the DOP20 and the satellite images, which otherwise would contain far more data within the same bounds as the orthorectified aerial images.

Co-registration of elevated surfaces: Objects located above ground will not be correctly projected onto the DSM, leading to mismatches of variable amplitude compared to the underlying images. This particularly affects buildings, with the offset increasing the higher the buildings is. Thus, no performance evaluation can be performed on those areas, and we therefore filter them out in the annotations using OSM building footprint masks. Though not completely matching with one another, this still removes most of the badly co-registered annotation pixels.

Visibility versus completeness: Our annotations were created on off-nadir images, where not all surface pixels were visible due to the occlusion by trees, buildings and similar infrastructure. Some pixels visible in nadir-looking images therefore have no correspondence in the off-nadir images, as well as in the original annotations. While all annotated objects located on the ground remain correct, holes may appear in areas where tree crowns, buildings, or walls were present in the original images. Moreover, due to the temporal difference between the original images and their DOP and satellite counterparts, some objects visible in the off-nadir imagery may no longer be, though this is an exceedingly rare occurrence in our imagery.

Segmentation metrics and interpretability: Due to the issues mentioned before, we have to take precautions when interpreting the performance metrics for each experiment. As we report classical pixel-wise metrics, the most critical aspect is that of false positives: a good prediction made on top of a missing label will be unfairly penalizing the model. This affects the intersection over union (IoU), Dice and precision metrics, which are therefore computed as a reference point and not as an absolute measure of a model’s capabilities. The recall metric, on the other hand, provides a reliable estimate of our model’s actual generalization capability to new image modalities, as it reflects how many of the annotated objects are correctly predicted, regardless of how incomplete the predictions may be.

4.3 Results of the Super-Resolution Enhancement

Figure 4 illustrates the results of the U2D2 model on both DOP20 aerial images and satellite imagery from WV3 and GeoEye. The DOP20 images are enhanced 2×, improving the GSD from 20 to 10 cm. For these images, the U2D2 model

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-493-2026 | © Author(s) 2026. CC BY 4.0 License.

Source	GSD	Upsampling strategy	OSM	Road				Parking Areas			
				IoU*	Dice*	Prec.*	Rec.	IoU*	Dice*	Prec.*	Rec.
3K native	10 cm	–		69.52	82.02	79.79	84.38	44.89	61.97	66.71	57.85
			✓	**72.98**	**84.38**	81.47	**87.51**	**47.70**	**64.59**	67.37	**62.03**
3K ortho	10 cm	–		68.90	81.59	80.88	82.31	42.59	59.74	**69.39**	52.44
			✓	72.69	84.19	**83.90**	84.47	44.57	61.66	64.03	59.46
DOP20	20 cm	10 cm Bicubic (B)		45.96	62.97	**70.66**	56.80	18.16	30.74	**43.82**	23.67
			✓	45.95	62.96	69.76	57.38	**22.73**	**37.04**	43.64	**32.18**
		10 cm SR		44.77	61.85	61.54	62.17	15.87	27.40	38.81	21.17
			✓	**50.59**	**67.19**	68.15	**66.26**	19.46	32.58	37.55	28.78
WV3 (13° ONA)	30 cm	10 cm B		36.66	53.65	**73.94**	42.10	5.17	9.83	21.28	6.39
			✓	42.65	59.79	59.11	60.50	9.95	18.09	**31.36**	12.72
		20 cm B → 10 cm SR		44.21	61.31	60.83	61.81	6.66	12.49	27.46	8.08
			✓	**49.61**	**66.32**	64.17	68.63	10.17	18.47	28.81	13.59
		15 cm SR → 10 cm B		46.95	63.90	59.48	69.03	8.50	15.66	26.18	11.17
			✓	47.92	64.79	57.49	74.22	**12.56**	**22.31**	30.97	**17.44**
		15 cm SR → 7.5 cm SR → 10 cm B		43.41	60.54	54.87	67.53	8.83	16.22	26.50	11.69
			✓	46.76	63.72	54.74	**76.23**	11.51	20.65	30.23	15.68
WV3 (27° ONA)	30 cm	10 cm B		27.12	42.67	**82.24**	28.81	3.08	5.97	21.26	3.47
			✓	34.31	51.09	64.78	42.18	4.77	9.10	**28.15**	5.43
		20 cm B → 10 cm SR		37.90	54.97	67.30	46.46	2.95	5.74	24.34	3.25
			✓	**44.33**	**61.43**	68.18	55.89	5.47	10.36	25.60	6.50
		15 cm SR → 10 cm B		41.20	58.36	68.04	51.09	4.54	8.69	24.68	5.28
			✓	43.85	60.97	65.71	56.87	**5.86**	**11.07**	25.35	7.08
		15 cm SR → 7.5 cm SR → 10 cm B		39.04	56.16	64.21	49.90	5.37	10.20	23.74	6.50
			✓	44.28	61.38	62.73	**60.09**	5.84	11.03	24.11	**7.16**
GeoEye (28° ONA)	50 cm	10 cm B		3.15	6.10	52.64	3.24	0.18	0.36	17.17	0.18
			✓	13.66	24.04	57.65	15.19	1.38	2.73	**26.81**	1.44
		20 cm B → 10 cm SR		21.31	35.14	53.03	26.27	0.83	1.65	12.75	0.88
			✓	**31.98**	**48.47**	**64.31**	38.89	2.67	5.20	12.81	3.27
		40 cm B → 20 cm SR → 10 cm SR		18.69	31.49	63.52	20.94	0.93	1.83	12.09	0.99
			✓	31.73	48.18	62.84	39.06	**3.13**	**6.07**	15.32	**3.79**
		25 cm SR → 10 cm B		13.19	23.31	62.87	14.30	1.30	2.56	17.77	1.38
			✓	20.97	34.67	59.05	24.54	2.40	4.68	17.12	2.71
		25 cm SR → 12.5 cm SR → 10 cm B		17.58	29.90	57.89	20.15	1.39	2.75	14.41	1.52
			✓	28.50	44.35	51.15	**39.15**	2.77	5.39	14.80	3.29

Table 1. Quantitative evaluation of the effect of super-resolution on the performance of road and parking-area segmentation. The table lists the data sources, including the off-nadir angle (ONA) during image acquisition, the original GSD of each dataset, and the upsampling strategy. Two segmentation models are applied to the different images, with one model using OSM as an additional input source. Metrics marked with * have a limited reliability due to the incomplete annotations.

excels at removing compression artifacts and increasing the resolvability of objects like lane markings, buildings, and cars. For the satellite images, a two-stage process was used. First, the native 30 cm WV3 and 50 cm GeoEye images were upsampled to 20 cm GSD using bicubic interpolation; this was done to preserve as much original information as possible. These 20 cm interpolated images were then super-resolved to 10 cm GSD using the U2D2 model. As U2D2 is designed to work across a range of GSDs for both aerial and satellite data, its application reveals a strong difference between the interpolated images and the final SR images. The model not only enhances spatial resolution but also improves the resolvability of lane markings, vehicles, and buildings while removing noise artifacts. While the enhancement is more pronounced for the higher-quality WV3 image (native 30 cm), the SR image from the GeoEye data still provides a significant improvement over the original 50 cm GSD image. This demonstrates that SR is a valuable pre-processing step for increasing base image quality despite various degradations.

4.4 Results of Road and Parking Areas Segmentation

The results of the experiments conducted are summarized in Table 1. To assess the model's generalization capability to orthorectified images and to verify the quality of the projected labels, we evaluated the model on both the original 3K image and its ortho-projected counterpart. Overall, both setups achieved very similar and consistently high performance results. For roads, only minor differences were observed with a maximum difference of $\pm 0.6\,\%$ IoU, $\pm 0.5\,\%$ Dice, $\pm 2.4\,\%$ precision, and $\pm 3.0\,\%$ recall. The difference for parking areas lies within a couple of percents (up to $\pm 3.2\,\%$ IoU, $\pm 3.0\,\%$ Dice, $\pm 3.3\,\%$ precision, and $\pm 5.4\,\%$ recall).

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-493-2026 | © Author(s) 2026. CC BY 4.0 License.

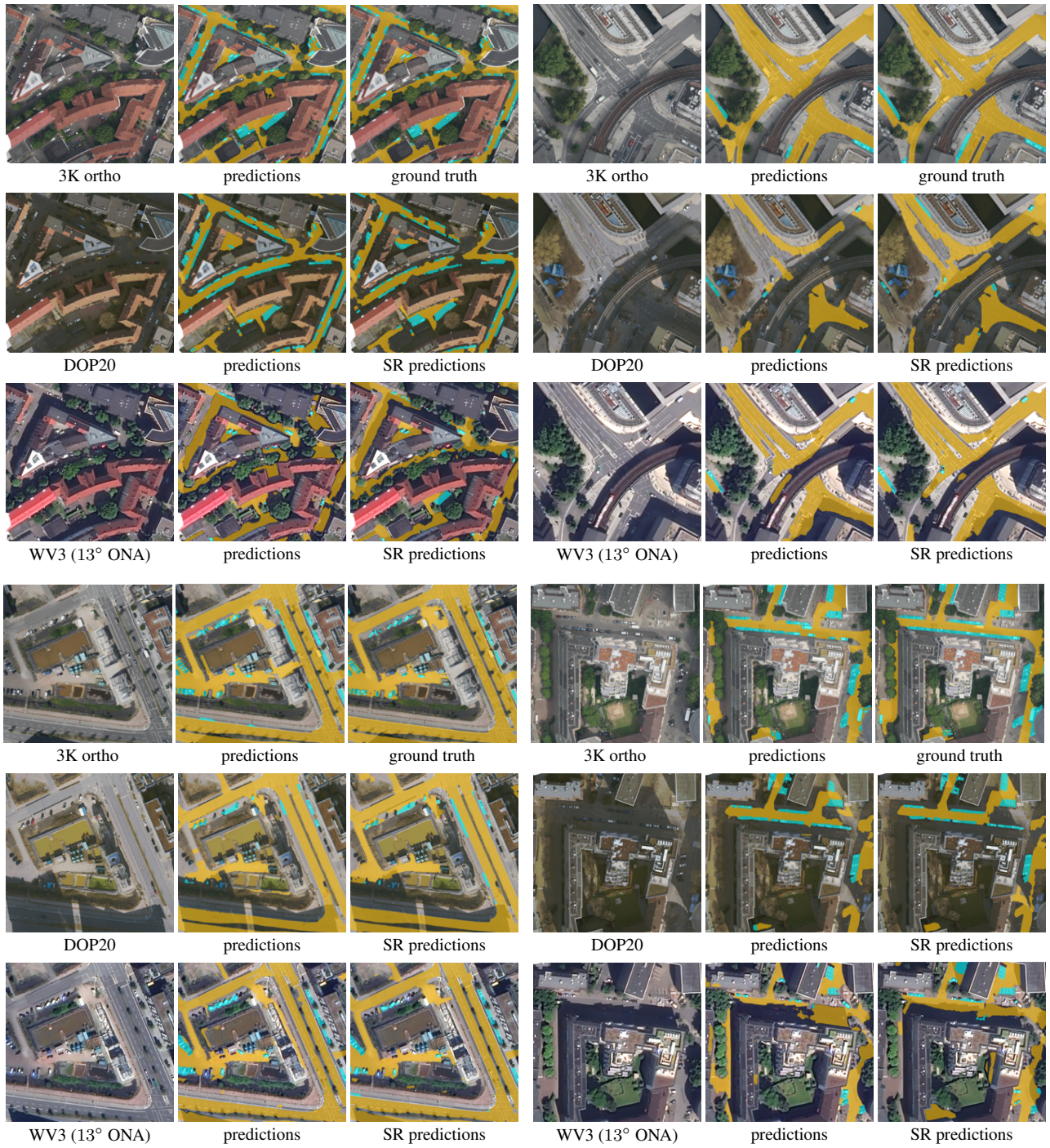


Figure 5. Qualitative evaluation of the effect of super-resolution on the performance of ■ road and ■ parking area segmentation. This figures shows four sample areas in Hamburg, Germany. For each area, the 3K image is shown together with the corresponding predictions and ground truth. In addition, the DOP20 and WV3 images are presented along with their predictions, both on the original images and on the super-resolution-enhanced versions.

The inclusion of OSM masks as priors in the input to the model, i.e. using the SkipFuse-Dense-U-Net-121 model in place of Dense-U-Net-121, systematically results in a significant improvement of all performance metrics (except precision in some cases), a finding consistent with our earlier experience with both models (Henry et al., 2021b).

Another observation is that, regardless of the upsampling strategy, incorporating an OSM mask as prior information consistently improves performance across all metrics except precision. For the road class, improvements range from 1 to 6 % IoU and from 1 to 8 % in recall for the DOP20 and WV3 13° ONA images. The gains are even more pronounced for the two most challenging satellite images and for the parking area class. The increase in overall scores despite a slight reduction in precision indicates that the number of additional true positive detections substantially outweighs the increase in false positives.

The results obtained on the 30 cm satellite imagery are comparable to, or even surpass, those achieved on the DOP20 data for the road class, both with and without SR: $-1.0\,\%$ IoU, $-0.9\,\%$ Dice, $+3.3\,\%$ precision, and $+10.0\,\%$ recall for the road class.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-493-2026 | © Author(s) 2026. CC BY 4.0 License.

In contrast, the performance of parking area segmentation in the satellite imagery still lags behind, with $-10.2\,\%$ IoU and $-14.7\,\%$ recall. These results show the large potential of satellite imagery for use in fine-grained road segmentation. However, its use for segmenting smaller objects, such as parking areas, requires further investigation.

For all setups tested, the use of SR consistently led to better road segmentation results compared to direct bicubic interpolation to a 10 cm/px GSD (4.3 to $7.0\,\%$ IoU improvement for the DOP20 and WV3 13° ONA images). Greater variance is observed in the results for the parking area class, where mixed outcomes were obtained (from -3.3 to $+2.6\,\%$ IoU change when using SR on the same images). This may come from the fact that the over-smoothing tendency of the SR algorithm tends to wipe some of the small contextual clues which indicate the location of parking area (e. g. very thin markings), though visual inspection does not reveal a significant impact of surface-to-surface contrast on either side of the border between roads and parking areas. Future works could look into making the models resilient to SR artifacts and bicubic blurring through data augmentation, in order to handle both upsampling types equally and identify which upsampling strategy is actually superior in that regard.

When comparing the influence of the off-nadir angle, Table 1 shows that the results for the 27° ONA WV3 image are worse than those for the 13° ONA image, with drops of $5.3\,\%$ in IoU and $12.7\,\%$ in recall for the road class using the best-performing upsampling strategy (20 cm B + 10 cm SR). The performance drops even further with the GeoEye image with a 28° ONA and a 50 cm/px GSD: $-17.6\,\%$ IoU and $-29.7\,\%$ recall. This is to be expected, as both represent the highest levels of difficulty, particularly in the challenging city scene selected for this study. In contrast, the 13° ONA 30 cm image is as optimal as satellite imagery can be for the task at hand.

When looking at the different upsampling strategies, we observe for the road class that starting with bicubic interpolation to 20 cm or 40 cm, followed by SR, provides the best balance between precision and recall, achieving the highest IoU and Dice scores (despite some uncertainty in the precision measure). In contrast, starting with SR and then adjusting the image GSD to 10 cm via bicubic interpolation appears to be more effective for maximizing recall alone. This pattern does not hold for the parking area class, which remains challenging to detect in the DOP20 images, and even more so in the satellite imagery.

The qualitative results in Figure 5 confirm the findings of the numerical analysis presented in Table 1 and discussed above. Furthermore, the complexity of the scenes from the city center of Hamburg is highlighted by the example images, as well as by the differences resulting from the temporal discrepancy between the original images and their DOP20 and satellite counterparts. More specifically, the DOP20 images, acquired in March, show significantly fewer tree crowns. This facilitates the model’s detection of roads and parking areas compared to the WV3 image from June. In contrast, new construction sites visible in the DOP20 images from 2021 (compared to the other images acquired in 2020) affect the model’s performance. Since roads within fenced-off construction sites are not annotated in the TIAS dataset, the model produces less accurate results in these areas.

In summary, the experiments demonstrate that the use of SR methods can help bridge the gap between aerial and high-resolution satellite imagery, enabling a label-free model transfer, i.e. without requiring additional manual annotations. Incorporating SR and OSM priors provides notable improvements on road segmentation, while performance for parking area segmentation remains more variable and challenging, highlighting the need for further research on small objects detection in both aerial and satellite imagery. The influence of off-nadir angle, image resolution, and temporal discrepancies between acquisitions further emphasizes the importance of carefully selecting input imagery and pre-processing strategies.

5. Conclusion & Future Work

In this study, we investigated the use of SR to reduce the spatial resolution disparity between aerial and high-resolution satellite imagery, thereby enabling a label-free transfer of a segmentation model trained on aerial images to satellite data. By applying a state-of-the-art SR method to upscale satellite images to 10 cm GSD and subsequently deploying a segmentation network trained on an aerial dataset, we assessed the feasibility of SR-supported model transfer for road and parking area segmentation. The experimental results across three test sites in Hamburg, Germany, indicate that SR can effectively support the transfer of models trained on aerial imagery to satellite imagery. For the road class, SR consistently improves segmentation accuracy compared to bicubic upsampling, particularly by increasing the results’ recall. The performance on parking areas remains more variable, reflecting the sensitivity of this class to fine-grained visual cues that may be attenuated by SR. However, the overall findings confirm that high-quality satellite imagery at 30 cm GSD already approaches or meets the performance obtained on DOP20 data, highlighting its suitability for fine-grained urban monitoring. This capability supports scalable and cost-efficient mapping workflows and opens opportunities for future research (e.g., in suburban and rural areas), on improving the robustness of the model to SR artifacts, optimizing upsampling strategies, and extending SR-enabled transfer to broader semantic classes and geographic regions.

References

Audebert, N., Le Saux, B., Lefevre, S., 2017. Joint Learning From Earth Observation and OpenStreetMap Data to Get Faster Better Semantic Maps. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Azimi, S. M., Bahmanyar, R., Henry, C., Kurz, F., 2021. Eagle: Large-scale vehicle detection dataset in real-world scenarios using aerial imagery. *2020 25th international conference on pattern recognition (ICPR)*, IEEE, 6920–6927.

BKG, 2025. Digitale Orthophotos und Satellitenbilddaten. Technical report.

Chlaily, S., Mura, M. D., Chanussot, J., Jutten, C., Gamba, P., Marinoni, A., 2021. Capacity and Limits of Multimodal Remote Sensing: Theoretical Aspects and Automatic Information Theory-Based Image Selection. *IEEE Transactions on Geoscience and Remote Sensing*, 59(7), 5598-5618.

d’Angelo, P., 2013. Automatic Orientation of large multitemporal Satellite Image Blocks. *Proceedings of International Symposium on Satellite Mapping Technology and Application 2013*, 1–6.

Dong, C., Loy, C. C., He, K., Tang, X., 2014. Learning a deep convolutional network for image super-resolution. *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13*, Springer, 184–199.

Grodecki, J., Dial, G., 2003. Block Adjustment of High-Resolution Satellite Images Described by Rational Functions. *Photogrammetric Engineering Remote Sensing*, 69(1), 59-70.

Hazirbas, C., Ma, L., Domokos, C., Cremers, D., 2017. FuseNet: Incorporating Depth into Semantic Segmentation via Fusion-Based CNN Architecture. *Computer Vision – ACCV 2016*, Springer International Publishing, Cham, 213–228.

Henry, C., Fraundorfer, F., Vig, E., 2021a. Aerial Road Segmentation in the Presence of Topological Label Noise. *2020 25th International Conference on Pattern Recognition (ICPR)*, 2336–2343.

Henry, C., Hellekes, J., Merkle, N., Azimi, S. M., Kurz, F., 2021b. Citywide Estimation of Parking Space Using Aerial Imagery and OSM Data Fusion with Deep Learning and Fine-Grained Annotation. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIII-B2-2021, 479–485.

Hong, D., Gao, L., Yokoya, N., Yao, J., Chanussot, J., Du, Q., Zhang, B., 2021. More Diverse Means Better: Multimodal Deep Learning Meets Remote-Sensing Imagery Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 59(5), 4340-4354.

Jangir, S. K., Mühlhaus, M., Merkle, N., Henry, C., Bahmanyar, R., 2025. A Generalized Two-Stage Framework for Blind Super-Resolution of Real-World Aerial and Satellite Imagery. *Journal of Selected Topics in Applied Earth Observations and Remote Sensing*. Submitted.

Kurz, F., Rosenbaum, D., Meynberg, O., Mattyus, G., Reinartz, P., 2014. Performance of a real-time sensor and processing system on a helicopter. *ISPRS Archives*, ISPRS Technical Commission I Symposium (Volume XL-1), ISPRS Archive, 189–193.

Kurz, F., Türmer, S., Meynberg, O., Rosenbaum, D., Runge, H., Reinartz, P., Leitloff, J., 2012. Low-cost optical Camera System for real-time Mapping Applications. *Photogrammetrie Fernerkundung Geoinformation*, Jahrgang 2012(2), 159–176.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z. et al., 2017. Photo-realistic single image super-resolution using a generative adversarial network. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4681–4690.

Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R., 2021. Swinir: Image restoration using swin transformer. *Proceedings of the IEEE/CVF international conference on computer vision*, 1833–1844.

Long, W., Zhou, X., Zhang, L., Gu, S., 2025. Progressive Focused Transformer for Single Image Super-Resolution. *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2279–2288.

Lyu, S., Zhao, Q., Zhou, Z., Li, M., Zhou, Y., Yao, D., Cheng, G., Zhou, H., Shi, Z., 2025. Deep Learning Based Domain Adaptation Methods in Remote Sensing: A Comprehensive Survey.

Maxar, 2022. WorldView-2 and WorldView-3 Satellite Imagery Products. Technical report.

Merkle, N., Azimi, S. M., Pless, S., Kurz, F., 2019. Semantic Vehicle Segmentation in Very High Resolution Multispectral Aerial Images Using Deep Neural Networks. *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, 5045–5048.

Merkle, N., Bahmanyar, R., Henry, C., Azimi, S. M., Yuan, X., Schopferer, S., Gstaiger, V., Auer, S., Schneibel, A., Wieland, M., Kraft, T., 2023. Drones4Good: Supporting disaster relief through remote sensing and AI. *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 3772–3776.

Merkle, N., Rauch, F., Henry, C., Hellekes, J., Kurz, F., 2024. TIAS: An aerial traffic infrastructure dataset to study transportation in urban environments. *GeoDPA - International Conference on Geoinformation Data, Processing and Applications*.

Peng, J., Huang, Y., Sun, W., Chen, N., Ning, Y., Du, Q., 2022. Domain Adaptation in Remote Sensing Image Classification: A Survey. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15, 9842-9859.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional networks for biomedical image segmentation. N. Navab, J. Hornegger, W. M. Wells, A. F. Frangi (eds), *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Springer International Publishing, Cham, 234–241.

Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D. J., Norouzi, M., 2022. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4), 4713–4726.

Schneibel, A., Gähler, M., Halbgewachs, M., Berger, R., Brauchle, J., Geßner, M., Gstaiger, V., Hein, D., Henry, C., Merkle, N., Klein, D., 2023. Using Earth Observation to Support First Aid Response in Crisis Situations– Lessons Learned from the Earthquake in Türkiye/Syria (2023). *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVIII-M-1-2023, 579–586. https://isprs-archives.copernicus.org/articles/XLVIII-M-1-2023/579/2023/.

Wang, X., Xie, L., Dong, C., Shan, Y., 2021. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1905–1914.

Xie, C., Zhang, X., Li, L., Fu, Y., Gong, B., Li, T., Zhang, K., 2025. MAT: Multi-range attention transformer for efficient image super-resolution. *IEEE Transactions on Circuits and Systems for Video Technology*.