deSEO: Physics-Aware Dataset Creation for High-Resolution Satellite Image Shadow Removal

Lorenzo Beltrame[1,4], Jules Salzinger[1], Filip Svoboda[2], Phillipp Fanta-Jende[1], Jasmin Lampert[1], Radu Timofte[3], Marco Körner[4,5,6]

[1] Austrian Institute of Technology, Giefinggasse 4, 1210 Vienna, Austria – (name.surname)@ait.ac.at
[2] University of Cambridge, William Gates Building, 15 JJ Thomson Ave., CB3 0FD Cambridge, UK – fs437@cam.ac.uk
[3] University of Würzburg, John Skilton Str. 4a, Hubland Nord, 97074 Würzburg, Germany – radu.timofte@uni-wuerzburg.de
[4] Technical University of Munich (TUM), TUM School of Engineering and Design, Chair of Remote Sensing Technology, Arcisstr. 21, 80333 Munich, Germany – marco.koerner@tum.de
[5] Technical University of Munich (TUM), Munich Data Science Institute (MDSI), 85748 Garching, Germany
[6] ELLIS Unit Jena, Friedrich Schiller University of Jena, 07743 Jena, Germany

Keywords: Shadow removal, Satellite imagery, Physics-informed, Dataset, GANs, Remote sensing

Abstract

Shadows cast by terrain and tall structures remain a major obstacle for high-resolution satellite image analysis, degrading classification, detection, and 3D reconstruction performance. Public resources offering geometry-consistent paired shadow/shadow-free satellite imagery are essentially missing, and most Earth-observation datasets are designed for shadow detection or 3D modelling rather than removal. Existing deep shadow-removal datasets either target ground-level or aerial scenes or rely on unpaired and weakly supervised formulations rather than explicit satellite pairs. We address this gap with *deSEO*, a *geometry-aware* and *physics-informed* methodology that, to the best of our knowledge, is the first to derive paired supervision for satellite shadow removal from the S-EO shadow detection dataset (Masquil et al., 2025) through a fully replicable pipeline. For each tile, deSEO selects a minimally shadowed acquisition as a weak reference and pairs it with shadowed counterparts using temporal and geometric filtering, Jacobian-based orientation normalisation, and LoFTR–RANSAC registration. A per-pixel validity mask restricts learning to reliably aligned regions, enabling supervision despite residual off-nadir parallax. In addition to this paired dataset, we develop a DSM-aware deshadowing model that combines residual translation, perceptual objectives, and mask-constrained adversarial learning. In contrast, a direct adaptation of a UAV-based SRNet/pix2pix architecture fails to converge under satellite viewpoint variability. Our model consistently reduces the visual impact of cast shadows across diverse illumination and viewing conditions, achieving improved structural and perceptual fidelity on held-out scenes. deSEO therefore provides the first reproducible, geometry-aware paired dataset and baseline for shadow removal in satellite Earth observation.

1. Introduction

The abundance of *Earth Observation (EO)* data in combination with advanced machine learning techniques is currently revolutionising the field of remote sensing. Among the many physical challenges to aerial and satellite-based remote sensing, terrain-induced shadows constitute an impediment to various forms of analysis (Li et al., 2016). This is particularly true in high-slope environments such as mountains (Giles, 2001), but also in cities where man-made structures can cast strong shadows on other regions of interest (Dare, 2005). Physics-based methods, grounded in light transport theory, have long provided solutions for shadow detection and removal. Still, those methods are known to produce artifacts that can interfere with downstream processing (Le and Samaras, 2022). Following recent trends in various fields, deep learning-based methods have been successfully applied to this task (Dong et al., 2023), yielding improvements in both deshadowing and downstream task performance (Zhu et al., 2024; Zhang et al., 2025).

Existing shadow-aware EO datasets, such as the *Aerial Imagery Shadow Detection (AISD)* dataset (Luo et al., 2020) and the *CUHK-Shadow* dataset (Hu et al., 2021), are primarily designed for shadow detection, rather than their removal. They rely on manual annotations, which are costly and potentially subjective, and they do not provide true shadow-free references. While the large-scale *Shadow-aware Earth Observation (S-EO)* dataset (Masquil et al., 2025) advances detection with geometry-aware, automatically generated masks, it remains detection-oriented and does not provide paired acquisitions suitable for weakly supervised deshadowing. More broadly, current resources seldom control for seasonal shifts, viewing geometry differences, or illumination changes, all of which introduce confounding factors that hinder the training and fair evaluation of shadow-removal models in high-resolution satellite imagery.

A further challenge is that, despite recent progress in deep learning for shadow removal, there is currently no public available dataset of geometry-consistent, paired shadowed and shadow-free EO satellite imagery suitable for weakly supervised deshadowing. Existing shadow-removal datasets either focus on ground-level scenes (Qu et al., 2017; Wang et al., 2018; Vasluianu et al., 2023), provide paired imagery acquired by *unmanned aerial vehicles (UAVs)* under controlled conditions (Luo et al., 2023), or adopt unpaired and weakly supervised formulations without true shadow-free references (Wang et al., 2024). None of these resources captures the multi-date, multi-angle satellite imaging geometry, with *rational polynomial coefficient (RPC)* camera models and *digital surface model (DSM)* priors, that is required to build reliable cross-view shadow correspondences. Consequently, shadow-removal methods developed for close-range or UAV imagery do not transfer to satellite imagery, where parallax, seasonal variability, and radiometric inconsistencies dominate. This motivates the need for a methodology that can transform detection-oriented EO datasets into paired resources for deshadowing through principled geometric filtering and reproducible data processing.

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-3-2026-11-2026 | © Author(s) 2026. CC BY 4.0 License.

To address these limitations, we introduce *deSEO*, a data processing methodology that leverages multi-temporal views of a scene to translate shadow-detection datasets into shadow-removal datasets. Minimally shadowed images are selected as proxy references and paired with shadowed counterparts, enabling weakly supervised training despite the technical impossibility of perfectly shadow-free ground truth. The pairing enforces weak supervision under explicit constraints (*e.g.*, seasonal proximity, footprint overlap, and view-geometry similarity) to reduce bias from seasonal and illumination variations while preserving sufficient shadow contrast for learning. In doing so, deSEO reframes existing detection-focused resources into datasets engineered for weakly supervised deshadowing in high-resolution EO imagery.

Our contributions are threefold, *i.e.*,

i) a *geometry-aware pipeline* for constructing paired training samples for satellite deshadowing from a multi-acquisition EO shadow detection dataset through two-stage filtering, orientation normalisation, feature-based registration, and validity-aware pairing,
ii) the first paired, *geometry-consistent* dataset for high resolution satellite shadow removal, derived from S-EO using the deSEO pipeline, as well as
iii) a *deshadowing model* and *training strategy* tailored to high-resolution satellite imagery, which uses registration-driven validity masks to restrict supervision to reliable correspondences and remain robust to residual misalignment.

2. Methodology

We propose a preprocessing and training pipeline that makes the S-EO dataset (Masquil et al., 2025) suitable for weakly supervised deep learning of shadow removal in high-resolution satellite imagery. The pipeline leverages S-EO's geometry-aware design and multi-temporal, multi-angle acquisitions to create paired samples that support data-driven deshadowing under realistic acquisition variability. In the second part of this study, we propose a single-stage, weakly supervised network to perform deshadowing, leveraging the physical priors incorporated into S-EO. A diagram of the shadow removal dataset creation deSEO is presented in Figure 1.

2.1 Dataset Overview: S-EO

The S-EO dataset consists of 20 000 georeferenced WorldView-3 images covering 702 tiles of $500\,\mathrm{m} \times 500\,\mathrm{m}$ across the three cities of San Diego (UCSD), Omaha (OMA), and Jacksonville (JAX), acquired over several years to provide diverse solar and viewing geometries. Available modalities include pan-sharpened RGB and panchromatic imagery ($30\,\mathrm{cm/px}$), LiDAR-derived minimal and maximal DSM heights ($50\,\mathrm{cm/px}$), physically generated shadow masks, NDVI vegetation masks, and bundle-adjusted RPC camera models. These study areas span diverse urban and suburban morphologies, enabling evaluation of geographic generalisation. While the S-EO dataset provides both physically simulated shadow masks and corresponding uncertainty maps identifying unprojected or geometrically unreliable pixels, the latter were not used in the deSEO pairing process. This choice follows from the fundamental difference in supervision design: our goal is to construct paired, image-level correspondences rather than to refine pixel-wise shadow annotations. Because shadow detection is not the target task, incorporating the uncertainty masks would have excluded large regions of otherwise valid imagery, reducing the diversity of spatial and radiometric contexts available for pairing. Instead, deSEO introduces registration-driven validity masks that are derived directly from the multi-temporal alignment stage. These masks delimit reliable correspondences after geometric harmonisation, ensuring that losses are applied only where cross-view registration is confident, thereby serving the same purpose as the original uncertainty maps but in the context of weakly supervised deshadowing. Figure 2 shows an example of a pansharpened RGB image, a max shadow mask, and a DSM required for the deSEO pipeline. In the remainder, the shadow masks generated from the DSM are referred to as the *shadow masks* for short.

2.2 Deshadowing Dataset Creation: deSEO

It is possible to obtain a shadow-removal dataset from a shadow-detection dataset by exploiting paired crops from multi-temporal, high-resolution satellite imagery that contains a shadowy scene and a counterpart presenting less pronounced shadows. For each tile, the least-shadowed acquisition is selected as a weak reference, and all other acquisitions of that tile are paired with this reference to form shadowed–clean pairs. The DSM-based shadow masks and the DSMs themselves serve as auxiliary inputs, providing physically grounded cues about scene structure and illumination.

In the following, we describe the proposed pipeline consisting of i) standardise metadata, ii) sample fixed-size windows, iii) select a target date per window based on shadow content, iv) pair it with a geometrically and temporally compatible input date, v) link elevation data, vi) record results in per-scene IDs, and vii) produce train, validation, test splits from scenes with valid pairs.

i) Data Organisation and Matching Each scene is associated with RGB images, binary shadow masks, and acquisition metadata. A common index is created, linked to the image, masks, and metadata triplet, keeping only indices that are present in all required modalities. Acquisition timestamps from heterogeneous formats are converted to timezone-aware *coordinated universal time (UTC)*. The scene footprint is summarised by a bounding polygon or bounding box, and key viewing and solar geometry (*i.e.*, off-nadir angle, look azimuth, sun elevation, sun azimuth) is derived for further downstream filtering.

ii) Window Sampling For each scene, a fixed number of K random windows of size $C \times C\,\mathrm{px}$ (*e.g.*, $K = 10, C = 576$) is drawn using a reproducible seed. These windows define the crops used to evaluate local shadow content and to specify exact crop coordinates in the input-target data pair.

iii) Per-window Target Selection and Pairing For each window $w \subset \mathbb{R}^2$ and acquisition date t, let $S_t(x, y) \in \{0, 1\}$ be the binary S-EO shadow mask ($1 =$ non-shadow, $0 =$ shadow). We measure the *clear* (non-shadow) fraction

$$\rho(w, t) = \frac{1}{|w|} \sum_{(x,y)\in w} S_t(x, y).$$

in w. The target date

$$t^\star = \arg\max_t \rho(w, t) = \arg\min_t \left(1 - \rho(w, t)\right),$$

is chosen as the least-shadowed (*i.e.*, most clear) observation, which is equivalent to minimising the shadow fraction, since it corresponds to $1 - \rho(w, t)$.

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-3-2026-11-2026 | © Author(s) 2026. CC BY 4.0 License.

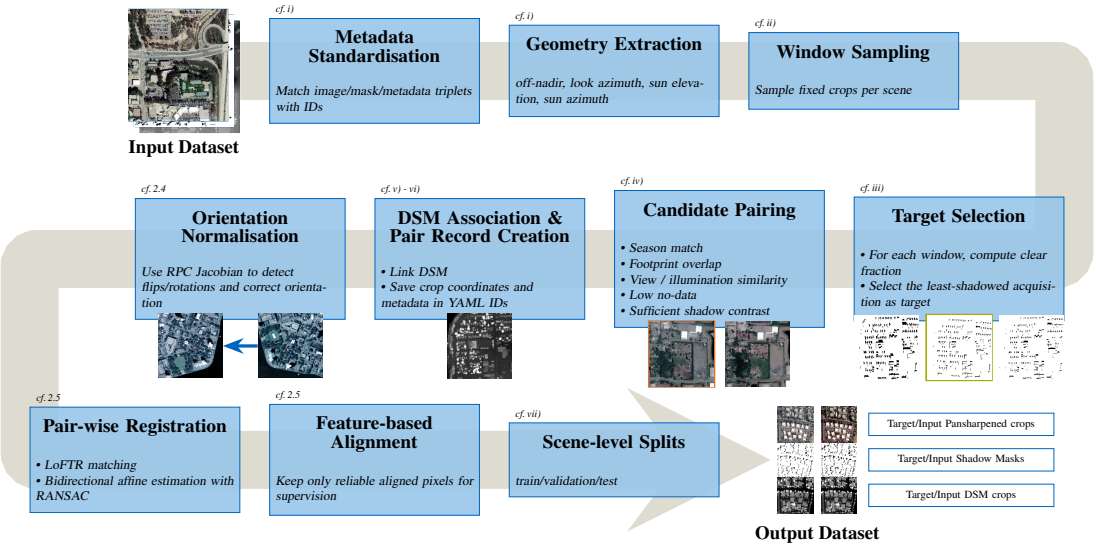
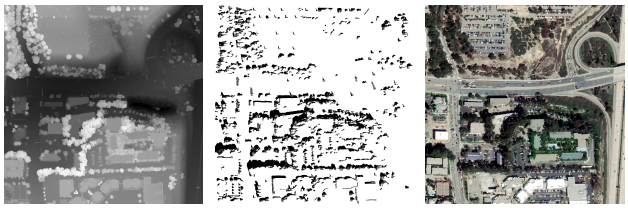


Figure 1. Diagram of deSEO, the proposed framework for shadow removal dataset creation. The processing pipeline uses multitemporal acquisitions paired with a digital surface model and the relative shadow masks as input and outputs a machine-learning-ready dataset.



(a) DSM (b) shadow mask (c) pan-sharpened RGB

Figure 2. Acquisition from the UCSD tile: (a) LiDAR-derived DSM (maximum aggregation), (b) the physically simulated shadow mask, and (c) the corresponding WorldView-3 RGB image. The DSM footprint is larger than the RGB and mask because it is designed to cover all the offsets in the time series.

iv) Data-quality Constraints Among dates $i \neq t^\star$, a single input time step is selected if it satisfies temporal, geometric, and data-quality constraints (*cf.* Table 2) and exhibits sufficient shadow contrast, *i.e.*,

$$\Delta s(w; t^\star, i) = \left| s(w, t^\star) - s(w, i) \right| \geq \tau_{\text{shadow}} \ .$$

At most one input is kept per target per window to avoid redundancy. In this study, $\tau_{\text{shadow}} = 0.1$ is selected. Furthermore, a temporal filtering is applied. A season-aware wrap-around gap is enforced through

$$\Delta_d(t^\star, i) = \min\big(|d(t^\star) - d(i)|,\ 365 - |d(t^\star) - d(i)|\big) \leq \tau_{\text{season}}$$

with $d(i)$ denoting the be day-of-year. A stricter limit τ_{winter} may be applied for winter months November – February to limit the scene changes due to tree canopies and vegetation.

To further ensure high data quality in the target shadow-free reference, a candidate i is accepted only if all of the following conditions hold (*cf.* Table 2):

Footprint overlap: The *intersection-over-union* between target and candidate footprints exceeds τ_{IoU}.

View-geometry proximity: The differences in off-nadir angle, look azimuth, and sun elevation are below $\tau_{\text{off}}, \tau_{\text{az}}, \tau_{\text{sun}}$, respectively. For circular angle differences, we use

$$\Delta\theta(\theta_1, \theta_2) = \pi - |\pi - |\theta_1 - \theta_2|| \, .$$

No-data check: The fraction of no-data pixels in the candidate RGB crop does not exceed τ_{nodata}.

If no candidate passes, the scene is skipped due to poor alignment with the reference acquisition.

v) Digital Surface Model Association If a DSM is available, its path is linked to each accepted pair, enabling downstream data processing to incorporate surface height information.

vi) Data Pair and IDs Creation For every accepted pair (input, target) and window, a YAML entry records the file paths to the input-target RGB and masks, the associated metadata paths, the crop coordinates (y, x, h, w), the scene IDs, and the DSM file path. These IDs are used to generate data pairs during training/evaluation.

vii) Dataset Splits A scene-level split is adopted to avoid spatial data leakage between the training, validation, and test sets. Each scene tile is assigned to exactly one split, so that no overlapping footprints or near-duplicate acquisitions appear across sets. The final configuration comprises 37 training scenes, 7 validation scenes, and 9 test scenes (*cf.* Table 1). In particular, each city is represented in all three splits, and distinct scenes from every city are held out for validation and testing, enabling both within-city and cross-city generalisation to be assessed.

Splits are reported as scene-ID lists and recorded in YAML files in the accompanying code repository. All thresholds (*cf.* Table 2), sampling hyperparameters (K, C), file-naming patterns, and dataset paths are defined in a structured configuration, yielding a deterministic process given a fixed seed. The outcome consists of i) paired, crop-level samples with strong inter-date shadow contrast and controlled geometric differences, ii) per-scene YAML IDs that reproduce every crop, and iii) consistent scene-level train, validation, and test splits for fair evaluation.

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-3-2026-11-2026 | © Author(s) 2026. CC BY 4.0 License.

Table 1. Dataset splits grouped by city. The number of samples refers to paired image crops per split.

Split	# Scenes	# Samples	Scene IDs (grouped by city)
Train	37	255	**UCSD**: 75, 721, 741, 744; **OMA**: 290, 385, 387, 478, 734, 766, 798, 831, 835, 927; **JAX**: 170, 203, 204, 275, 276, 295, 296, 313, 557, 584, 618, 652, 725, 726, 727, 728, 762, 763, 799, 800, 801, 802, 837
Val	7	44	**UCSD**: 742; **OMA**: 291, 423, 799; **JAX**: 172, 230, 764
Test	9	63	**UCSD**: 189; **OMA**: 760; **JAX**: 171, 231, 254, 255, 314, 530, 560

Table 2. Thresholds for constructing noisy–ground-truth pairs. The ‘Refined’ column indicates any overrides or additional constraints imposed by the refined filtering procedure.

Parameter	Pretraining	Refined	Explanation
τ_{IoU}	0.7	0.7	Minimum bounding box overlap
τ_{off}	90°	7°	Maximum allowed difference in camera viewing angle
τ_{az}	100°	10°	Maximum allowed difference in azimuth angle
τ_{sun}	90°	90°	Maximum allowed difference in sun elevation angle
τ_{nodata}	0.03	0.03	Maximum allowed fraction of missing pixels
τ_{season}	60 days	30 days	Maximum allowed difference between acquisition seasonal drift
τ_{winter}	60 days	10 days	Tighter limit applied during winter period
τ_{shadow}	0	0.1	Minimum fraction of difference between shadowy pixels

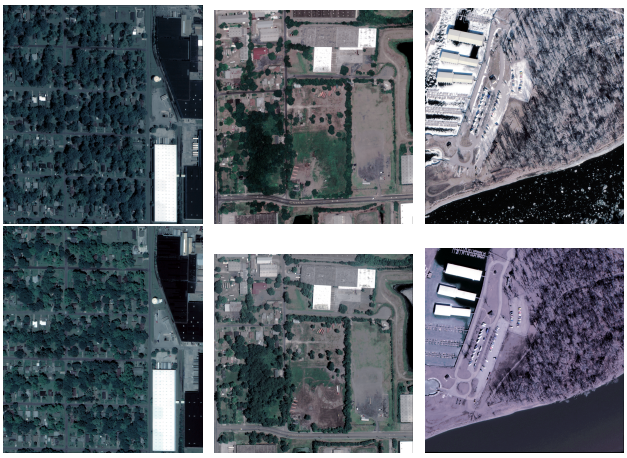
2.3 Geometric Harmonisation

A practical challenge when working with the S-EO dataset is the prevalence of off-nadir acquisitions. Differences in viewing geometry are typical in satellite imagery and can weaken alignment between paired dates. The effect is most pronounced in dense urban areas, where occlusions and parallax distort the image projection despite unchanged scene geometry. Our pipeline addresses this by using RPC metadata to align and normalise geometry across dates and viewing directions in sensor space through the use of feature-based aligners. Pairs that remain outside predefined geometric tolerances after harmonisation are discarded, since they do not contribute to good reconstructions for the model. This treatment allows the use of a broader range of imagery, including opportunistic acquisitions that would otherwise be difficult to employ. It also supports applications where controlled acquisition is limited, such as disaster response or resource-constrained settings, and may extend to contexts with weak calibration. Pair selection proceeds in two passes:

Pretraining: broad thresholds (*cf.* ‘pretraining’ column in Table 2) to maximise coverage; residual viewpoint differences are handled by feature-based alignment (*cf.* Section 2.5).

Refinement: tighter thresholds (*cf.* ‘refined’ column in Table 2) for comparisons with methods assuming strong alignment (*e.g.*, UAV-based pipelines).

We regulate seasonal proximity with τ_{season} using day-of-year, so cross-year matches are allowed within the same tolerance, *e.g.*, 22nd Nov 2015 *vs.* 18th Nov 2016. Because our dataset is restricted to the northern hemisphere, thresholds τ_{winter} are further tightened from November to March to prevent leaf-off/leaf-on canopy changes and snow coverage that would otherwise promote blurred outputs.



(a) (b) (c)

Figure 3. Rejected examples for (a) azimuth, (b) off-nadir, and (c) seasonal/temporal constraints following the corresponding threshold in Table 2.

Table 3. Orientation diagnosis from Jacobian derivatives (Equation (1)) and their corrections.

Case	$\partial\text{row}/\partial\text{lon}$	$\partial\text{row}/\partial\text{lat}$	$\partial\text{col}/\partial\text{lon}$	$\partial\text{col}/\partial\text{lat}$	Orientation
(A)	>0	— otherwise	—	>0	-90° 90°
(B)	—	<0	>0	—	correct
	—	>0	>0	—	V. flip
	—	<0	<0	—	H. flip
	—	>0	<0	—	180° rotation

(A) $|\partial\text{row}/\partial\text{lon}| \gg |\partial\text{row}/\partial\text{lat}|$, $|\partial\text{col}/\partial\text{lat}| \gg |\partial\text{col}/\partial\text{lon}|$
(B) $|\partial\text{row}/\partial\text{lat}| \geq |\partial\text{row}/\partial\text{lon}|$, $|\partial\text{col}/\partial\text{lon}| \geq |\partial\text{col}/\partial\text{lat}|$
H: horizontally, V: vertically

Figures 3a to 3c show representative rejections across three filters to visualise the effect of our geometric and temporal pairing constraints—*i.e.*, azimuth difference, off-nadir angle, and seasonal/temporal proximity, respectively—in accordance with the corresponding thresholds summarised in Table 2. These filters, tailored to the S-EO multi-date, multi-angle setting, are the basis of our deSEO pairing pipeline and prevent geometrically or radiometrically inconsistent matches from entering training.

2.4 Orientation Normalisation

Additionally, due to variations in WorldView 30 satellite orbits, georeferenced imagery often differs in axis orientation. To correct for this, we approximate the Jacobian

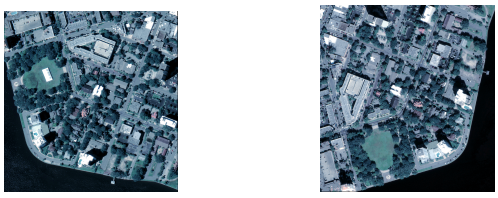
$$J : (\text{lon}, \text{lat}) \mapsto (\text{row}, \text{col}) \quad = \begin{bmatrix} \frac{\partial\,\text{row}}{\partial\,\text{lon}} & \frac{\partial\,\text{row}}{\partial\,\text{lat}} \\ \frac{\partial\,\text{col}}{\partial\,\text{lon}} & \frac{\partial\,\text{col}}{\partial\,\text{lat}} \end{bmatrix}, \qquad (1)$$

of the mapping from world coordinates (lon, lat) to image coordinates (row, col) which we estimate using central finite differences. The image orientation is determined by the relative magnitudes of derivatives (to detect axis swaps) and their signs (to detect flips/rotations). The cases are summarised in Table 3.

2.5 Pairwise Registration and Validity Mask

Since a residual perspective distortion persists in the acquired data due to the 10° tolerance in the off-nadir angle, we perform a feature-based registration between shadowy and non-shadowy image pairs. Preliminary experiments with lightweight

https://doi.org/10.5194/isprs-annals-XI-3-2026-11-2026 | © Author(s) 2026. CC BY 4.0 License.



(a) before preprocessing

(b) after preprocessing

Figure 4. Examples for (a) the unprocessed acquisition with incorrect rotation and (b) the same area after orientation correction during preprocessing, as described in Table 3.

descriptors such as *oriented FAST and rotated BRIEF (ORB)* (Rublee et al., 2011) and *accelerated features (XFeat)* (Potje et al., 2024) yield insufficient alignment quality. We therefore adopt the detector-free *local feature transformer (LoFTR)* framework (Sun et al., 2021), which has demonstrated strong performance in remote sensing contexts (Jovhari et al., 2023). In our pipeline, grayscale input and target images are processed with the outdoor-pretrained LoFTR model, using a confidence threshold of 0.20 and a cap of 8000 correspondences, to obtain dense keypoint matches. The images input into LoFTR are downscaled to have a maximum width of $W = 1024\,\mathrm{px}$ and height H scaled accordingly. Affine transformations are then estimated bidirectionally using *random sample consensus (RANSAC)* (Fischler and Bolles, 1981), and the model with the highest inlier count and lowest median reprojection error is selected. Finally, to ensure supervision is applied only where cross-view correspondence is well-defined, we propagate a binary validity mask $M \in \{0,1\}^{H\times W}$ alongside each training pair after warping. The L_1 loss of the *pix2pix* generator and the discriminator inputs are restricted to the valid subset by masking and renormalisation.

2.6 Neural Network Architecture

We frame shadow removal as a conditional image-to-image translation task driven by multi-date, multi-geometry supervision. The model consists of a U-Net generator (Ronneberger et al., 2015) that predicts a residual deshadowing correction and a PatchGAN-style discriminator (Isola et al., 2017) enhanced by a soft, mask-driven attention prior. Together, they leverage geometry-aware cues and validity masking to learn illumination-consistent corrections while preserving radiometry outside shadow regions.

In all the following runs, optimisation uses Adam (Kingma and Ba, 2014), a cosine learning-rate schedule (Loshchilov and Hutter, 2016) over 100 epochs, and weights initialisations drawn from a normal distribution with a gain of 0.02. Training is performed with a batch size of 576×576 tiles, using random crops of size $256 \times 256\,\mathrm{px}$, and employs two discriminator updates per generator update during the first 20 epochs. Applying self-attention at intermediate feature resolutions best preserves fine-grained structure in shadow removal. This placement sharpens edges and micro-textures while avoiding over-smoothing at high resolutions and detail loss at coarse scales. It balances local detail with contextual consistency, preserving high-frequency content. By contrast, using attention only in the generator weakens adversarial supervision and produces overly smooth results. Ablation results support this design choice (cf. Table 4).

2.6.1 Generator The proposed model follows a pix2pix-style *conditional generative adversarial network (cGAN)* architecture (Isola et al., 2017) in which the generator G is realised by a U-Net with skip connections. Unlike *embedding space consistency networks (ESCNet; Luo et al., 2023)*, where shadow removal using a *style-guided re-deshadow network (SRNet; Wan et al., 2022)* and radiometric correction through a *radiation adjustment network (RANet)* are separated, we employ a single end-to-end translation model. This unified formulation reflects the deSEO setup, where multi-date, multi-geometry satellite supervision and DSM-derived priors favour a geometry-aware treatment. Optional spectral normalisation (Miyato et al., 2018) and non-local self-attention layers, following the *self-attention generative adversarial network (SAGAN)* design (Zhang et al., 2019), allow each spatial location to attend globally for improved context modelling. When available, DSMs are concatenated with RGB inputs to form a 4-channel condition for both G and D.

All reconstruction and perceptual objectives are computed only on valid RGB pixels, excluding DSM and padded regions according to the masks introduced in Section 2.5. The reconstruction term can be configured as L_1, SSIM, VGG-perceptual (Johnson et al., 2016), or LPIPS with VGG backbone (Zhang et al., 2018)

$$\begin{aligned}\mathcal{L}_{\text{rec}}^{L_1}(\hat{B}, B) &= \|\hat{B} - B\|_1,\\ \mathcal{L}_{\text{rec}}^{\text{VGG}}(\hat{B}_{\text{rgb}}, B_{\text{rgb}}) &= \sum_{\ell} w_\ell \|\phi_\ell(\hat{B}_{\text{rgb}}) - \phi_\ell(B_{\text{rgb}})\|_1, \text{or}\\ \mathcal{L}_{\text{rec}}^{\text{LPIPS}}(\hat{B}_{\text{rgb}}, B_{\text{rgb}}) &= \text{LPIPS}_{\text{VGG}}(\hat{B}_{\text{rgb}}, B_{\text{rgb}}),\end{aligned}$$

respectively. Perceptual losses are evaluated on RGB tensors normalised to ImageNet mean and standard deviation, and are preferred to mitigate blur under imperfect alignment. To maintain colour and hue stability in non-shadowed regions, we apply a colour-consistency loss and an HSV-based regularization

$$\begin{aligned}\mathcal{L}_{\text{color}} &= \|(\hat{B}_{\text{rgb}} - A_{\text{rgb}}) \odot M_{\text{tar}}\|_1 \quad \text{and}\\ \mathcal{L}_{L_1\,\text{HS}} &= \|[\Delta H, |S_{\hat{B}} - S_A|] \odot M_{\text{tar}}\|_1 \quad ,\end{aligned}$$

respectively, with $\Delta H = \min(|H_{\hat{B}} - H_A|, 1 - |H_{\hat{B}} - H_A|)$.

In addition to the mask-restricted objectives, we include a lightly weighted global L_1 loss term $\mathcal{L}_{+L_1}$ to stabilise optimisation. This provides a uniform gradient across all valid pixels, especially in weakly aligned regions and prevents global colour drift without overriding the perceptual and colour-consistency losses.

The full generator objective

$$\begin{aligned}\mathcal{L}_G = \mathcal{L}_{\text{GAN}} + \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{col}}\mathcal{L}_{\text{color}} + \lambda_{L_1\,\text{HS}}\mathcal{L}_{\text{HS}}\\ + \lambda_{+L_1}\mathcal{L}_{+L_1}\end{aligned}$$

combines all components and is applied only on valid correspondences to ensure stability under weak alignment. Optionally, inputs or gradients corresponding to pixels that remain shadowed in both views can be zeroed out using the prior mask S.

2.6.2 Discriminator To capture high-frequency details, we focus the discriminator on local image neighbourhoods rather than the entire image, following the classical PatchGAN approach (Isola et al., 2017). The discriminator classifies whether each overlapping $N \times N$ patch appears real or generated, operating fully convolutional across the image. The resulting patch-wise responses are then aggregated (*e.g.*, averaged) to produce the final output of the discriminator D.

We enhance this design with a *soft shadow attention mask (SSAM)* that focuses the adversarial signal on regions transitioning from shadow to non-shadow. Given binary input and target

masks $M_{\text{in}}, M_{\text{tar}} \in \{0,1\}^{W\times H}$, respectively, the attention prior

$$\text{SSAM}(x,y) = (1 - M_{\text{in}}(x,y))M_{\text{tar}}(x,y),$$

emphasises areas of shadow disappearance. For each layer l, the feature map

$$\phi_l \leftarrow \phi_l \odot (1 + \gamma_l \widehat{S}_l) \qquad \text{with} \qquad \widehat{S}_l = \tfrac{S_l}{\text{mean}(S_l)+\varepsilon}$$

is modulated by a normalised attention map, where γ_l is a learnable gating parameter. The optional suppression of gradients from fully shadowed pixels can be enforced via the relative flag. All convolutional layers use spectral normalisation and a dropout layer with $p = 0.1$ follows each downsampling block. Training supports vanilla, *least-squares generative adversarial network (LSGAN; Mao et al., 2017)*, or hinge adversarial losses, with their optional relativistic variants. One-sided label smoothing (*e.g.*, $y_{\text{real}} \in [0.9, 1]$) and an additional penalty $\lambda_{R_1} \cdot \mathbb{E}\|\nabla_x D(x)\|_2^2$ on real samples are available for regularisation. Early training may use $k_D > 1$ discriminator updates per generator step to stabilise the adversarial game. This geometry-aware discriminator emphasises shadow transitions while preserving radiometric stability elsewhere, complementing G's reconstruction and perceptual losses.

3. Results

Before developing our architecture, we trained the original SRNet of Luo et al. (2023) on deSEO-generated paired data. Training quickly became unstable, with non-convergent adversarial loss and severe generator artefacts. We attribute this to SRNet's reliance on tightly aligned UAV imagery with near-pixel correspondence, assumptions violated by our multi-temporal, multi-geometry satellite pairs, which contain residual misalignment and scene changes. We therefore use SRNet as the closest transfer baseline. Its failure indicates that UAV-oriented shadow-removal models do not readily transfer to high-resolution satellite data, motivating the geometry-aware design of deSEO and our model. This negative result motivated the development of a dedicated architecture specifically designed for high-resolution satellite imagery. In particular, we adopt a geometry-aware formulation that emphasises perceptual and feature-space losses, rather than purely pixel-wise objectives, in order to improve robustness under imperfect correspondences and illumination variability. The following experiments evaluate this model on the deSEO dataset, analyse the impact of key design choices, and quantify the contribution of geometry-aware inputs and perceptual supervision.

We first explore a range of training configurations to identify models that can reconstruct high-resolution satellite imagery and learn useful features for shadow removal. Using the pretraining dataset obtained with the filtering procedure in Table 2, this stage does not yet produce high-quality reconstructions but encourages the generator to learn structural and radiometric priors of satellite imagery. Representative validation examples are shown in Figure 5.

Pretraining prioritises optimisation stability over deshadowing quality. Low, balanced learning rates and a mild discriminator warm-up prevent early divergence, while the loss emphasises L_1 and HVS terms, disabling colour, perceptual, and attention losses to reduce adversarial sensitivity before meaningful spatial structure is learned. DSM input and relatively large model capacities are retained to enable early geometry-aware feature learning.

(a) Reconstruction (b) Input (c) Target

Figure 5. Examples from the validation set during pretraining.

(a) input (b) target (c) full model (d) L_1-only

Figure 6. Qualitative comparison of the full model and the L_1-only ablation.

Although pretraining alone does not yield accurate reconstructions or deshadowing, it establishes structural and radiometric priors. We then fine-tune on the refined dataset (*cf.* Table 2), containing higher-quality, more consistent pairs. The core optimisation setup is kept fixed, varying only capacity and regularisation. Finetuning uses longer training, smaller batches, spectral normalisation in the discriminator, dropout in the generator, and activation of geometry-aware modules (Gamma and shadow attention) with DSM input. These changes stabilise adversarial training and exploit geometric cues while limiting overfitting.

We validate these design choices through an ablation study (Table 4). The full model achieves the best balance across metrics, with the highest SSIM and the lowest VGG19 and LPIPS. Removing DSM causes the largest degradation, halving PSNR and SSIM and confirming its central role. Disabling self-attention or gamma features leads to moderate drops, mainly in structural fidelity and perceptual quality, while removing HV regularisation or spectral normalisation slightly increases reconstruction and perceptual errors. Overall, DSM is the dominant factor, with spectral normalisation and HV regularisation having smaller individual effects.

Some variants, such as the L_1-only model, achieve strong pixel-wise scores (highest PSNR/RMSE and low L_1; Table 4) but fail at deshadowing. They prioritise intensity matching over shadow removal, yielding worse perceptual metrics and visibly retaining shadows (*cf.* Figure 6). Reducing the channel capacity similarly induces a predictable decline, although the model remains relatively robust to such architectural compression. Overall, the results highlight that DSM input, perceptual supervision, and stabilising regularisation (HV, spectral norm, attention) contribute jointly to the model's ability to preserve both radiometric and perceptual fidelity while performing meaningful shadow reduction.

To fully assess the model's performance, we evaluate it on the test set, with results provided in Table 5. A qualitative example is presented in Figure 7. The final test run uses the same configuration as our baseline model (Exp. 4 in Table 4). The results on the held-out test set are presented in Table 5. Similarly to

https://doi.org/10.5194/isprs-annals-XI-3-2026-11-2026 | © Author(s) 2026. CC BY 4.0 License.

Table 4. Ablation study on the validation set (mean ± standard deviation). Standard deviations rounded to one significant digit; means adjusted accordingly. Metrics computed on RGB channels.

Configuration	PSNR ↑	RMSE ↓	SSIM ↑	L_1 ↓	VGG19 L_1 ↓	LPIPS ↓
Baseline	*18 ± 2*	32 ± 5	**0.6 ± 0.1**	**0.09 ± 0.01**	**0.44 ± 0.04**	**0.41 ± 0.07**
RGB only (remove DSM)	9 ± 2	90 ± 20	0.18 ± 0.06	0.30 ± 0.05	0.57 ± 0.06	0.71 ± 0.02
No G self-attention	17 ± 2	37 ± 8	0.5 ± 0.1	0.10 ± 0.02	0.47 ± 0.05	0.42 ± 0.07
Gamma features disabled	17 ± 2	35 ± 9	**0.6 ± 0.1**	0.10 ± 0.03	0.46 ± 0.04	0.42 ± 0.06
No HV regularisation	16 ± 2	39 ± 9	0.4 ± 0.1	0.11 ± 0.03	0.46 ± 0.05	0.45 ± 0.08
L_1-only (no perceptual)	**18 ± 1**	**31 ± 5**	**0.6 ± 0.1**	**0.09 ± 0.01**	0.55 ± 0.05	0.47 ± 0.04
Reduced latent channels	17 ± 2	35 ± 7	0.5 ± 0.1	0.10 ± 0.01	0.46 ± 0.04	0.44 ± 0.06
No spectral normalisation	16 ± 2	39 ± 7	0.5 ± 0.1	0.12 ± 0.02	0.49 ± 0.05	0.46 ± 0.07
Shadow attention disabled	**18 ± 1**	34 ± 3	**0.6 ± 0.1**	0.10 ± 0.01	0.45 ± 0.05	0.42 ± 0.06
No Pretraining	17 ± 2	36 ± 8	0.5 ± 0.1	0.10 ± 0.01	0.46 ± 0.05	0.42 ± 0.06
No dropout	17 ± 2	37 ± 7	**0.6 ± 0.1**	0.10 ± 0.02	0.47 ± 0.05	0.42 ± 0.06

Table 5. Final metrics evaluated on the test set.

Metric	Mean ± Std
PSNR (dB) ↑	18 ± 1
RMSE ↓	34 ± 8
SSIM ↑	0.49 ± 0.08
L_1 $[0, 1]$ ↓	0.09 ± 0.03
Perceptual (VGG19 L_1) ↓	0.42 ± 0.08
LPIPS (VGG) ↓	0.46 ± 0.05

(a) reconstruction (b) input (c) target

Figure 7. Examples from the test set on the baseline model.

validation (*cf.* Table 4), the model favours perceptual fidelity (low VGG19-L_1 and LPIPS, competitive L_1) over distortion metrics such as PSNR and RMSE. This mirrors why the baseline configuration performed best on validation: it optimises SSIM, L_1, and perceptual metrics rather than maximising PSNR/RMSE (Table 4). Moreover, this behaviour is not confined to the validation split—the same pattern appears in the test set, suggesting that it is a stable property of the model rather than an observation on a particular subset. The relatively large standard deviations on test (*e.g.*, SSIM ±0.08, RMSE ±8) suggest scene-dependent variance, likely tied to shadow coverage and illumination. This spread is partially attributable to the limited size of the validation and test splits. In future iterations, we plan to expand the dataset, both in scale and geographic diversity, to better quantify generalisation across a broader range of conditions. Because shadows occupy a small fraction of each image, purely global metrics can underweight improvements in shadowed regions. We therefore complement metrics with a blind manual review focused on cast-shadow boundaries and relit regions. Despite limited texture recovery, the method consistently identifies and reduces shadows (*cf.* Figure 7). Nevertheless, the major limitation remains that the difference between the shadow profiles is often not significant across the image, so the reconstructed image does not yield a completely shadow-free reference. One reason might be the lower cardinality of the dataset, potentially not enabling the model to learn general shadow features. Given the limited extent of most shadows, traditional metrics are more informative about fine-detail reconstruction than deshadowing itself, justifying our use of an additional visual assessment. The results are qualitatively encouraging: the model can reliably identify shadows in unseen images under different lighting conditions and off-nadir perspectives, providing a basis for further improvement.

4. Conclusion

We address the challenge of generating reliable paired data for deshadowing in high-resolution satellite imagery, where the lack of true shadow-free references and the variability of acquisition geometries make supervised learning particularly difficult. We introduced deSEO, a geometry-aware preprocessing and training pipeline that transforms shadow detection datasets into weakly supervised deshadowing datasets by exploiting multi-date acquisitions under explicit geometric, temporal, and radiometric constraints. The pipeline produces reproducible window-level pairings, registration-driven validity masks, and scene-level splits suitable for weakly supervised deshadowing.

Using S-EO as a case study, we demonstrate that the proposed pipeline can derive high-quality training pairs from a detection-oriented resource. We also propose a geometry-aware deshadowing model, inspired by SRNet but redesigned for satellite imagery, that can learn meaningful shadow reduction despite residual misalignment. While global metrics on the test set show only moderate improvements, they are consistent across scenes and align with qualitative inspection, which confirms that the model consistently reduces the visual impact of cast shadows under diverse illumination and viewing conditions. The ablation study further highlights the importance of DSM inputs, perceptual supervision, and stabilising regularisation in achieving perceptually coherent deshadowing (*cf.* Table 4).

The current size of the S-EO-derived dataset remains a critical limiting factor, contributing to the variance observed across scenes and constraining the model’s ability to learn generalisable shadow patterns. While the present results are best characterised as shadow reduction rather than complete shadow removal, we believe that deSEO provides a practical path toward full shadow removal in high-resolution satellite imagery. By enabling the systematic creation of weakly supervised training data from multi-temporal observations, deSEO establishes the data foundation needed for models to progressively learn stronger shadow compensation under diverse illumination, seasonal, and viewing conditions. We therefore view this work not as an endpoint, but as a first step toward fully shadow-free reconstruction.

Future work will focus on applying deSEO to additional datasets spanning different sensors and geographic regions, enriching the diversity of paired samples, and incorporating more advanced radiometric normalisation, uncertainty modelling, and modelling strategies that better disentangle shadow effects from intrinsic surface appearance. Improved temporal and geometric alignment will also be important for moving from shadow reduction

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-3-2026-11-2026 | © Author(s) 2026. CC BY 4.0 License.

towards more complete shadow removal. Finally, given the inherent difficulty of obtaining shadow-free ground truth, evaluating the effect of deshadowing on downstream tasks such as classification or change detection remains an important direction for future work and would enable a broader assessment of its practical impact.

Overall, deSEO establishes a first reproducible, physics-aware dataset creation methodology for weakly supervised deshadowing in high-resolution satellite imagery. This framework provides a foundation upon which more robust and generalisable deshadowing methods can be developed, particularly in complex observation conditions where shadows and illumination effects degrade image quality.

Acknowledgments

This work was carried out within the SAFIR research project funded by the Austrian Research Promotion Agency (FFG) as part of the Research, Technology & Innovation (RTI) initiative ‘Digitaler Zwilling Österreich’. Code and materials are available on GitHub[1].

References

Dare, P. M. (2005). ‘Shadow analysis in high-resolution satellite imagery of urban areas’. In: *Photogramm. Eng. Remote Sens.* 71.2, pp. 169–177.

Dong, X., Cao, J. and Zhao, W. (2023). ‘A review of research on remote sensing images shadow detection and application to building extraction’. In: *Eur. J. Remote Sens.* 57.1, p. 2293163. DOI: 10.1080/22797254.2023.2293163.

Fischler, M. A. and Bolles, R. C. (1981). ‘Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography’. In: *Communications of the ACM* 24.6, pp. 381–395. DOI: 10.1145/358669.358692.

Giles, P. T. (2001). ‘Remote sensing and cast shadows in mountainous terrain’. In: *Photogramm. Eng. Remote Sens.* 67.7, pp. 833–840.

Hu, X. et al. (2021). ‘Revisiting Shadow Detection: A New Benchmark Dataset for Complex World’. In: *IEEE Transactions on Image Processing* 30, pp. 1925–1934. DOI: 10.1109/tip.2021.3049331.

Isola, P. et al. (2017). ‘Image-to-image translation with conditional adversarial networks’. In: *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 5967–5976. DOI: 10.1109/cvpr.2017.632.

Johnson, J., Alahi, A. and Fei-Fei, L. (2016). ‘Perceptual Losses for Real-Time Style Transfer and Super-Resolution’. In: *Eur. Conf. Comput. Vis. (ECCV)*, pp. 694–711. DOI: 10.1007/978-3-319-46475-6_43.

Jovhari, N. et al. (2023). ‘Performance evaluation of learning-based methods for multispectral satellite image matching’. In: *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.* X-4/W1-2022, pp. 335–341. DOI: 10.5194/isprs-annals-x-4-w1-2022-335-2023.

Kingma, D. P. and Ba, J. L. (2014). ‘Adam: A Method for Stochastic Optimization’. In: DOI: https://doi.org/10.48550/arXiv.1412.6980.

Le, H. and Samaras, D. (2022). ‘Physics-Based Shadow Image Decomposition for Shadow Removal’. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 44.12, pp. 9088–9101. DOI: 10.1109/tpami.2021.3124934.

Li, H. et al. (2016). ‘A general variational framework considering cast shadows for the topographic correction of remote sensing imagery’. In: *ISPRS J. Photogramm. Remote Sens.* 117, pp. 161–171. DOI: 10.1016/j.isprsjprs.2016.03.021.

Loshchilov, I. and Hutter, F. (2016). ‘SGDR: Stochastic Gradient Descent with Warm Restarts’. In: *Int. Conf. Learn. Represent. (ICLR)*.

Luo, S., Li, H. and Shen, H. (2020). ‘Deeply supervised convolutional neural network for shadow detection based on a novel aerial shadow imagery dataset’. In: *ISPRS J. Photogramm. Remote Sens.* 167, pp. 443–457. DOI: 10.1016/j.isprsjprs.2020.07.016.

Luo, S. et al. (2023). ‘An Evolutionary Shadow Correction Network and a Benchmark UAV Dataset for Remote Sensing Images’. In: *IEEE Trans. Geosci. Remote Sens.* 61, pp. 1–14. DOI: 10.1109/tgrs.2023.3295450.

Mao, X. et al. (2017). ‘Least Squares Generative Adversarial Networks’. In: *International Conference on Computer Vision (ICCV)*, pp. 2813–2821. DOI: 10.1109/iccv.2017.304.

Masquil, E. et al. (2025). ‘S-EO: A Large-Scale Dataset for Geometry-Aware Shadow Detection in Remote Sensing Applications’. In: *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, pp. 2374–2384. DOI: 10.1109/cvprw67362.2025.00224.

Miyato, T. et al. (2018). ‘Spectral normalization for generative adversarial networks’. In: *Int. Conf. Learn. Represent. (ICLR)*.

Potje, G. et al. (2024). ‘XFeat: Accelerated Features for Lightweight Image Matching’. In: *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 2682–2691. DOI: 10.1109/cvpr52733.2024.00259.

Qu, L. et al. (2017). ‘DeshadowNet: A Multi-context Embedding Deep Network for Shadow Removal’. In: *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 2308–2316. DOI: 10.1109/cvpr.2017.248.

Ronneberger, O., Fischer, P. and Brox, T. (2015). ‘U-Net: Convolutional Networks for Biomedical Image Segmentation’. In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234–241. DOI: 10.1007/978-3-319-24574-4_28.

Rublee, E. et al. (2011). ‘ORB: An efficient alternative to SIFT or SURF’. In: *International Conference on Computer Vision (ICCV)*, pp. 2564–2571. DOI: 10.1109/iccv.2011.6126544.

Sun, J. et al. (2021). ‘LoFTR: Detector-Free Local Feature Matching with Transformers’. In: *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 8918–8927. DOI: 10.1109/cvpr46437.2021.00881.

Vasluianu, F.-A., Seizinger, T. and Timofte, R. (2023). ‘WSRD: A Novel Benchmark for High Resolution Image Shadow Removal’. In: *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, pp. 1826–1835. DOI: 10.1109/cvprw59228.2023.00181.

Wan, J. et al. (2022). ‘Style-Guided Shadow Removal’. In: *Eur. Conf. Comput. Vis. (ECCV)*, pp. 361–378. DOI: 10.1007/978-3-031-19800-7_21.

Wang, J., Li, X. and Yang, J. (2018). ‘Stacked Conditional Generative Adversarial Networks for Jointly Learning Shadow Detection and Shadow Removal’. In: *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 1788–1797. DOI: 10.1109/cvpr.2018.00192.

Wang, Q. et al. (2024). ‘Recreating Brightness From Remote Sensing Shadow Appearance’. In: *IEEE Trans. Geosci. Remote Sens.* 62, pp. 1–11. DOI: 10.1109/tgrs.2024.3398576.

Zhang, A. et al. (2025). ‘An impervious surfaces extraction method based on optical, ascending and descending SAR remote sensing imagery in high-density urban core areas’. In: *Int. J. Appl. Earth Obs. Geoinf.* 140, p. 104595. DOI: 10.1016/j.jag.2025.104595.

Zhang, H. et al. (2019). ‘Self-Attention Generative Adversarial Networks’. In: *Int. Conf. Mach. Learn. (ICML)*. PMLR, pp. 7354–7363.

Zhang, R. et al. (2018). ‘The Unreasonable Effectiveness of Deep Features as a Perceptual Metric’. In: *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 586–595. DOI: 10.1109/cvpr.2018.00068.

Zhu, X. et al. (2024). ‘Mitigating terrain shadows in very high-resolution satellite imagery for accurate evergreen conifer detection using bi-temporal image fusion’. In: *Int. J. Appl. Earth Obs. Geoinf.* 134, p. 104244. DOI: 10.1016/j.jag.2024.104244.

[1] https://github.com/AIT-Assistive-Autonomous-Systems/deSEO