Solar Energy 318 (2026) 115058

Contents lists available at ScienceDirect

Solar Energy

journal homepage: www.elsevier.com/locate/solener

Geometry transfer of deep neural networks for heliostat detection

Rafal Broda [a, d, *], Alexander Schnerring [a, d], Michael Nieslony [a], Tobias Wagner [a], Dominik Schnaus [e], Niels Algner [a], Marc Röger [a], Sonja Kallio [a], Rudolph Triebel [c, f], Robert Pitz-Paal [b, d]

[a] German Aerospace Center (DLR), Institute of Solar Research, Calle Doctor Carracido 44, Almería, 04005, Spain
[b] German Aerospace Center (DLR), Institute of Solar Research, Linder Höhe, Cologne, 51147, Germany
[c] German Aerospace Center (DLR), Institute of Robotics and Mechatronics, Münchener Str. 20, Wessling, 82234, Germany
[d] RWTH Aachen University, Chair of Solar Technology, Linder Höhe, Cologne, 51147, Germany
[e] Technical University of Munich (TUM), Chair of Computer Vision and Artificial Intelligence, Boltzmannstr. 3, Garching, 85748, Germany
[f] Karlsruhe Institute of Technology (KIT), Institute for Anthropomatics and Robotics, Haid-und-Neu-Str. 7, Karlsruhe, 76131, Germany

HIGHLIGHTS

- Synthetic dataset generated to represent existing heliostat collector geometries.
- Real-world test sets compiled for three distinct geometries.
- Baseline, universal, and fine-tuned models compared across geometries.
- Universal model is inconsistent, while fine-tuning adapts reliably with minimal effort.

ARTICLE INFO

Keywords:
Heliostat
Deep learning
Object detection
Keypoint detection
Transfer
Geometry

ABSTRACT

Reliable object and keypoint detection on heliostats is essential for advancing automation and practical deployment of airborne condition-monitoring methods in concentrated solar thermal (CST) tower plants. Yet the wide variety of heliostat collector geometries in real systems limits the applicability of deep-learning approaches, and their transferability between plants remains unexplored. To address this gap, we compiled a database of representative real-world heliostat geometries and generated a large synthetic dataset using our previously published rendering framework. With this data, we evaluated three training strategies: geometry-specific baseline models trained from scratch, a universal model intended to generalize across all geometries, and fine-tuning approaches initialized from either the baseline or the universal model. Baseline models perform well on their respective geometries, while the universal model shows inconsistent performance and may not be sufficient as a stand-alone solution. Fine-tuning, however, consistently adapts the model to new geometries and achieves performance comparable to geometry-specific baselines, as shown by suitable metrics on real-world test datasets of three distinct collector types. In practice, an effective model for a new geometry can be obtained by rendering as few as 100 synthetic images and fine-tuning an available baseline or universal model for 2000 iterations. This procedure requires 20 h on a single GPU and scales efficiently with additional computational resources. Overall, the results demonstrate that fine-tuning an existing heliostat detection model to the target domain provides a practical and reliable strategy for deploying deep models across the diverse heliostat collector geometries present in CST plants.

1. Introduction

Concentrated solar thermal (CST) plants, especially those using towers, demonstrate substantial potential for providing sustainable and dispatchable energy [1] and fuel [2]. These systems employ up to thousands of individual mirror collectors, so-called heliostats, which reflect the sunlight onto a tower receiver. Ensuring efficient and reliable operation of these plants requires both a precise initial setup upon construction and regular condition monitoring of the heliostat field during operation. Conventional approaches for inspection such as the so-called camera-target method [3] are labor- and time-intensive, motivating the

* Corresponding author at: German Aerospace Center (DLR), Institute of Solar Research, Calle Doctor Carracido 44, Almería, 04005, Spain.
Email address: rafal.broda@dlr.de (R. Broda).

https://doi.org/10.1016/j.solener.2026.115058
Received 11 December 2025; Received in revised form 10 June 2026; Accepted 24 August 2026
Available online 4 September 2026

Acronyms

AP	average precision
CST	concentrated solar thermal
DNN	deep neural network
GPU	graphics processing unit
HDRI	high dynamic range image
HPC	high performance computing
IoU	intersection over union
PCK	percentage of correct keypoints
PSA	Plataforma Solar de Almería
STJ	Solar Tower Jülich
VRAM	video random access memory

use of drones to acquire aerial imagery that can be analyzed to extract operation-relevant information [4–7].

Image processing techniques have been proposed for this task, but existing pipelines still lack the accuracy, robustness, and scalability required in practice. In contrast, deep learning has demonstrated exactly these advantages in many other domains, leading to first attempts to apply deep neural networks (DNNs) to heliostat detection [8–12]. However, these studies are limited to specific collector geometries, with detection performance tied to the facet arrangement and size of the heliostats under consideration. This geometry dependence is problematic because a wide range of heliostat designs has been developed over time for different applications. Based on available literature [13,14] at least 33 different heliostat collector geometries can be identified, as shown in Table 1. It remains unclear whether models trained on one geometry can generalize to others.

In previous work [12], we demonstrated that transfer learning enables adaptation across plants, but this approach was only evaluated qualitatively and still required additional rendering and retraining, preventing the deployment of a ready-to-use model. Preliminary evidence for a working geometry transfer model was also provided in a master's thesis by Wagner [15], which investigated simulation-to-simulation generalization through a cluster-based study. While this work showed that transfer is possible in certain cases, it was restricted to synthetic data and did not incorporate the aspect of simulation-to-reality transfer settings.

Building on these insights, the present study provides the first systematic investigation of heliostat collector geometry transfer using real-world image test datasets and a simulation framework designed to minimize the simulation-to-reality gap. By disentangling geometry dependence, this study provides critical insights toward developing robust, transferable deep learning solutions for solar tower plant monitoring.

2. Methodology

The objective of this study is to systematically evaluate the transfer capabilities of deep learning models for heliostat detection across different collector geometries. To achieve this, we conduct experiments in three stages: (1) baseline training to determine the requirements and effort involved in developing a model for a single geometry and to serve as a reference for comparison, (2) universal training to assess the feasibility of a single, general-purpose model applicable across all geometries, and (3) fine-tuning as an alternative strategy for transferring models between geometries. For this purpose, we construct a synthetic dataset which can incorporate – depending on the case (1) through (3) – a wide range of heliostat designs, train representative deep learning models, and evaluate their performances using real-world data. This section first describes the dataset generation process, then introduces the model architectures and training setup, and finally explains the evaluation framework used to assess transferability.

2.1. Synthetic dataset generation

At the core of the dataset generation process is the parametric *Blender* model of a heliostat, presented in [11]. The 3D model enables flexible adaptation of the collector geometry by adjusting parameters such as the number and size of facets, inter-facet gaps (including an optional middle gap), and the pylon's height and diameter. The corresponding parameters are illustrated in Fig. 1. This flexibility enables the simulation of a wide variety of heliostat designs observed in practice while maintaining a consistent rendering framework. To ensure compatibility with automated rendering and annotation, the model currently supports only the simulation of rectangular facet geometries. Consequently, only rectangular heliostats from Table 1 are included in the analysis, while geometry 33 – being a pentagonal collector – is excluded. Despite this limitation, the parametric approach covers the vast majority of heliostat types deployed in commercial and experimental plants.

Using the compiled data and making a reasonable 0.01 m gap assumption between mirror facets where information is unavailable, the

Table 1
Overview of heliostat collector geometries used for this study, sorted by mirror-facet count. Note that given size measurements are assumed and rounded and do not necessarily represent exact dimensions.

id		model name ○ manufacturer △ plant/operator □	facet number [-]		facet size [m]		middle gap [m]	ref
			x	y	x	y		
1	○	eSolar ST3	1	1	1.1†	1.1†	0.00	[13]
2	○	eSolar SCS5	1	1	1.5†	1.5†	0.00	[13]
3	△	SUPCON China	1	1	1.6†	1.2†	0.00	[13]
4	□	CSIRO	1	1	2.4	1.8	0.00	[14]
5	○	BrightSource Energy LH-2.2 Wireless	2	1	2.3	3.3	0.50†	[13]
6	□	Greenway	2	1	2.4†	3.0†	0.50†	[13]
7	△	Heliosystems	2	1	1.8†	3.0†	0.00	[14]
8	○	BrightSource Energy LH-2.5 Wireless	2	1	2.0†	2.6†	0.00	[13]
9	△	SUPCON China	2	2	2.7†	1.8†	0.40†	[13]
10	□	Solar Tower Jülich	2	2	1.6	1.3	0.00	TW
11	△	AORA Solar	2	2	1.0†	1.0†	0.00	[14]
12	△	Cosin Solar	2	4	3.0†	1.3†	0.00	[13]
13	□	CESA-1 (PSA)	2	6	3.0	1.0	0.75	TW
14	□	Vast Solar	3	1	0.8†	1.5†	0.00	[13]
15	△	Hengji Nengmai China	3	3	1.8†	1.8†	0.00	[13]
16	△	AORA Solar	3	3	1.0†	1.0†	0.00	[14]
17	○	Jiangsu Xinchen P3	4	4	1.0	1.0	0.40†	[13]
18	△	AORA Solar	4	4	1.0†	1.0†	0.00	[14]
19	□	CESA-1 (PSA)	4	6	1.5	1.0	0.75	TW
20	○	Abengoa Spain Solucar 120	4	7	3.3†	1.3†	0.00	[13]
21	○	Abengoa Spain ASUP 140V3	4	8	3.3†	1.3†	0.00	[13]
22	□	Sandia	5	5	1.2†	1.2†	0.00	[13]
23	△	Shouhang China	5	7	2.3†	1.4†	0.00	[13]
24	□	Abengoa CRS Sales	6	6	1.8†	1.8†	1.00	[13]
25	○	Sener HE35	7	5	1.4†	2.5†	0.00	[13]
26	△	Shouhang China	7	5	1.5†	2.2†	0.00	[13]
27	○	SolarReserve Rocketdyne	7	5	1.8	1.8	0.00	[13]
28	○	Titan Tracker STD-150 HE	8	5	2.2†	1.8†	1.00†	[14]
29	○	Titan Tracker STD-264 HE	8	5	2.9†	2.3†	1.00†	[14]
30	△	Himin Solar China	8	8	1.3	1.3	0.00	[13]
31	□	Solar Two	8	8	1.0†	1.0†	0.50†	[13]
32	○	Sener HE54	9	6	1.5†	2.2†	0.00	[13]
33	○	sbp sonne Stellio	10		n.a.		n.a.	[13]

† Estimated using available information or image data. TW: this work

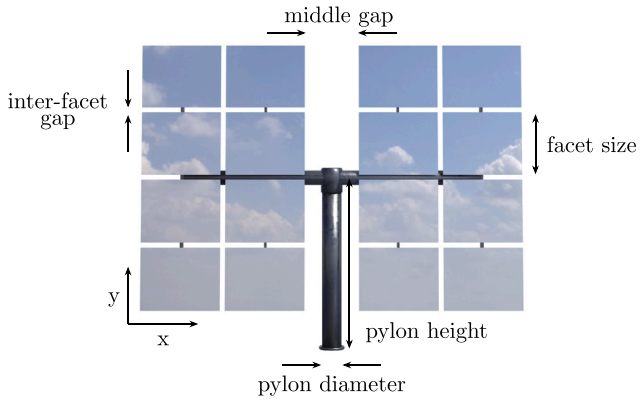

Fig. 1. Schematic of the heliostat model, whose geometry is determined by adjustable parameters including facet size, inter-facet gap, middle gap, pylon height, and pylon diameter in the parametric Blender implementation.

total collector area is determined and subsequently used to estimate the heliostat pylon height and diameter for modeling. Note that the given overview only provides a rough estimate of existing collector geometries and does not represent the exact dimensions: due to incomplete data, part of the information is calculated or inferred from image data. Additionally, for simplicity, facet dimensions are rounded to one decimal.

Training data is generated using the rendering framework introduced in previous studies [11,12]. Scene parameters are chosen according to the recommendations of [12], which systematically investigated the simulation-to-reality gap and evaluated the influence of scene parameters on transfer to real-world data. Therefore, we use procedural ground textures, soiling on mirror facets, a realistic clear-sky lighting high dynamic range image (HDRI) resembling the test cases, distractor objects, random object positions and heliostat orientations that track an assumed sun position.

To balance the heliostat density across different sizes and to ensure they fit within the scene, we fix the overall scene dimensions and adjust the number of heliostats based on the total collector area. Consequently, smaller geometries result in more placed heliostats, while larger collector geometries lead to fewer heliostats. The scene settings for collector geometry 13 derived in previous work are used as the reference case for scaling the number of heliostats for the other geometries, using an empirically derived mathematical relationship. In total, three different field configurations are sampled to provide variation in field size and heliostat density. Appendix A presents a comprehensive summary of the data utilized in the rendering process for all geometries.

In line with [12], we create 10 scenes per geometry and render photorealistic images with a resolution of 6000×4000 px along with labels. To balance dataset diversity with computational effort, we first render 2 000 images for geometries 10, 13 and 25, which constitute the test geometries with available real-world image data and perform a dataset size study as part of the experiments. The goal of this analysis is to determine the number of images per geometry required to achieve stable model performance. The details and outcomes of this study are presented later in Section 3. Based on the determined optimal dataset size, the data for the remaining geometries is generated. The sampling of camera poses is performed as described in [12].

The combination of the parametric heliostat model and the automated rendering pipeline enables the replication of the previously identified heliostat collector geometries in a consistent manner. This setup ensures efficient large-scale rendering across varied geometries and scene settings, producing a dataset designed to train models with strong simulation-to-reality transfer capabilities across a wide range of geometries. By doing so, we effectively compensate for the limited availability of real-world image data and lay a robust foundation for the subsequent investigation.

2.2. Model and training

To enable continuity with earlier research, we employ an encoder-decoder-based model architecture. In line with previous work [12], we combine a ResNet50 network [16] pre-trained on ImageNet [17] with a decoder part along with skip connections and heatmap-based output heads. The model output stride is two. This model has demonstrated strong performance in heliostat detection and serves as a representative, established foundation.

While transformer-based detection architectures have recently demonstrated strong performance on object detection benchmarks, these approaches are primarily designed for bounding-box-based object localization and generally rely on large-scale pre-training and substantial computational resources during training to achieve competitive performance [18]. In contrast, the heatmap-based encoder-decoder formulation employed in this work provides dense spatial predictions and is therefore well suited for the simultaneous estimation of heliostat bounding boxes and precise corner locations. For completeness, a later ablation study additionally investigates the replacement of the ResNet50 backbone with a recent vision transformer (ViT) architecture initialized from large-scale pre-training and discusses its impact on the obtained results.

For training, as in previous work, we largely adopt the settings described by Zhou et al. [19]. We use the same training objective, and for both objects and keypoints we scale the Gaussian kernel size based on the camera–object distance. Accordingly, for data augmentation, we employ a random scale-and-crop transformation. The input image is first resized by a random factor between 0.5 and 2.0 to promote scale invariance, and then a 2048×2048 px patch is cropped at random position to maximize GPU efficiency. The model is optimized using the Adam optimizer [20] with a one-cycle learning rate schedule [21] and a peak learning rate of 1×10^{-3}. Training from scratch is performed for 32 000 iterations on eight *NVIDIA A100 SXM4 80 GB* GPUs with an effective batch size of 64, requiring approximately 20 h. For fine-tuning an existing model to adapt it to another geometry, preliminary tests showed that 4000 iterations are sufficient, taking roughly 2.5 h. To reduce variance due to random initialization and optimization dynamics, the models are trained with three different random seeds and the mean and standard deviation of the results are reported.

We distinguish three distinct training dataset configurations. In the *specialized* setting, training is performed exclusively on data from the geometry used in testing. Both the *baseline* and *fine-tuning* stages correspond to this configuration. For the *universal* training stage, we design two complementary datasets: an *inclusive* dataset that combines data from all available geometries to develop a broadly representative model, and an *exclusive* dataset that excludes the geometries (or closely related ones) used for testing, specifically geometries 8, 9, 10, 11, 13, 19, 25, 26, and 27, in order to evaluate the model's ability to generalize to previously unseen geometries.

2.3. Evaluation framework

To evaluate geometry transfer, we employ three real-world test datasets. Each dataset contains six aerial images of a solar field, representing three distinct heliostat collector geometries. One of these geometries, employed at Plataforma Solar de Almería (PSA; owned and operated by CIEMAT), corresponds to geometry 13 in the compiled overview and the test data used in a previous study [12], ensuring comparability with earlier results. The second geometry is geometry 10 from Solar Tower Jülich (STJ) and its test data is newly introduced here to broaden the evaluation scope. The two test datasets are shown in Fig. 2. Given that the two geometries belong to small (geometry 10: $\sim 8\,\mathrm{m}^2$) and medium-sized (geometry 13: $\sim 40\,\mathrm{m}^2$) heliostats in terms of facet number and collector size, we also take into account a further six real-world test images from a solar field with geometry 25, which can be classified as large heliostats ($\sim 120\,\mathrm{m}^2$). Fig. 3 illustrates this statement. For confidentiality reasons, these images cannot be shown, but they cover roughly the same perspectives and object scales as the

Fig. 2. Top row: Images recorded during a measurement flight over the STJ plant, showing heliostat collector geometry 10 and supplementing the evaluation in this study. Bottom row: Masked images from the measurement flight by Jessen et al. [6] at the PSA (owned and operated by CIEMAT), showing heliostat collector geometry 13 and used in the evaluation of both previous work and the present study.

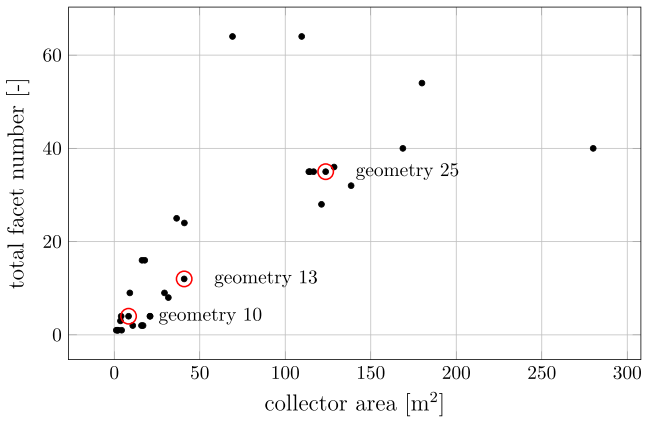

Fig. 3. Relationship between collector area and total facet number for available collector designs. The test geometries highlighted in red cover small, medium, and large collector sizes. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Summary of the real-world evaluation datasets. Gimbal pitch is reported relative to nadir (0°). Camera parameters are image width w, image height h, and focal length $f = f_x = f_y$. All images were resized to 6000×4000 px for evaluation.

Geom.	Plant	n_{img}	n_{hel}	n_{cor}	Flight height [m]	Pitch range [°]	Cam. params. [px]
10	STJ	6	901	3323	15–80	0–45	6000×4000 $f \approx 4000$
13	CESA-1 (PSA)	6	380	1391	30–120	0–55	6000×4000 $f \approx 4000$
25	Conf.	6	332	1191	45–90	50–70	8192×5460 $f \approx 7980$

images shown in Fig. 2. All test images are carefully annotated by hand, with bounding boxes drawn around each collector and precise sub-pixel localization of the four outer collector corners. A summary of the evaluation datasets, including acquisition parameters and annotation statistics, is provided in Table 2. We assess model performance using two complementary tasks: object detection and keypoint detection. For object detection we report the average precision (AP) based on intersection over union (IoU) overlap, following standard detection practices. Since different geometries are not distinguished but instead assigned to a single heliostat class, no classification step is performed. For keypoint evaluation, we measure the detection rate of the four collector outer corner points as percentage of correct keypoints (PCK). A detection is considered correct if (i) the predicted keypoint lies within 3 px of the ground truth location and (ii) the model confidence exceeds a value of 0.5. For the geometry 13 evaluation dataset, this threshold approximately corresponds to a physical distance of 2 cm to 12 cm, depending on whether the furthest or closest visible heliostat is considered, respectively. Similar ranges are observed for geometries 10 and 25. It should be noted that this value represents the maximum matching threshold used to classify a keypoint prediction as correct and not the actual localization error of the model, which is typically lower. Together, these evaluation tasks ensure that the model is assessed not only for its detection capability across different geometries but also for its ability to capture fine-grained geometric details essential for reliable heliostat monitoring.

3. Results and discussion

This section presents the experimental results following the three-stage structure introduced in the methodology. First, baseline training results are reported to identify ideal boundary conditions for training specialized models and to establish reference performance. Next, the outcomes of the universal training experiments are discussed to assess the generalization capability of a single model across all collector geometries. Subsequently, we analyze the fine-tuning results to evaluate the benefits of adapting pre-trained models to specific geometries. The final part discusses practical considerations concerning training efficiency, highlighting its relevance for real-world deployment.

3.1. Baseline training

A dataset size analysis in [12] showed that model performance remained largely unaffected when the training dataset was reduced from 20 000 to 2000 images. This observation suggests that the training dataset size may be further reduced, offering potential savings in computational resources and processing time. Therefore, in this study, we investigate the effect of progressively reducing the dataset size from 2000 down to 20 images for the three test geometries. During preliminary tests, we observed that images showing strongly occluded scenarios can negatively affect model performance, particularly when the training dataset is small. To mitigate this effect, such images were excluded from the training data in the present study. Affected samples can be identified, for instance, by computing the mean IoU per image and automatically removing those images below a defined threshold. Training is performed using synthetic data and evaluation is based on the three real-world test datasets. The aim of this analysis is twofold: first, to identify an optimal dataset size that minimizes rendering effort for subsequent experiments; and second, to establish a consistent reference for comparison.

The results of this experiment, shown in Fig. 4(a), reveal distinct outcomes across the three test geometries. Geometry 10 (small) reaches good performance with as few as 20 images and shows a slight decline when the dataset size exceeds 200 images. Geometry 13 (medium) stabilizes around 100 images, while geometry 25 (large) exhibits the strongest dependence on dataset size, with lowest performance at 20 images. Across all cases, performance converges between 0.5 and 0.7 AP and

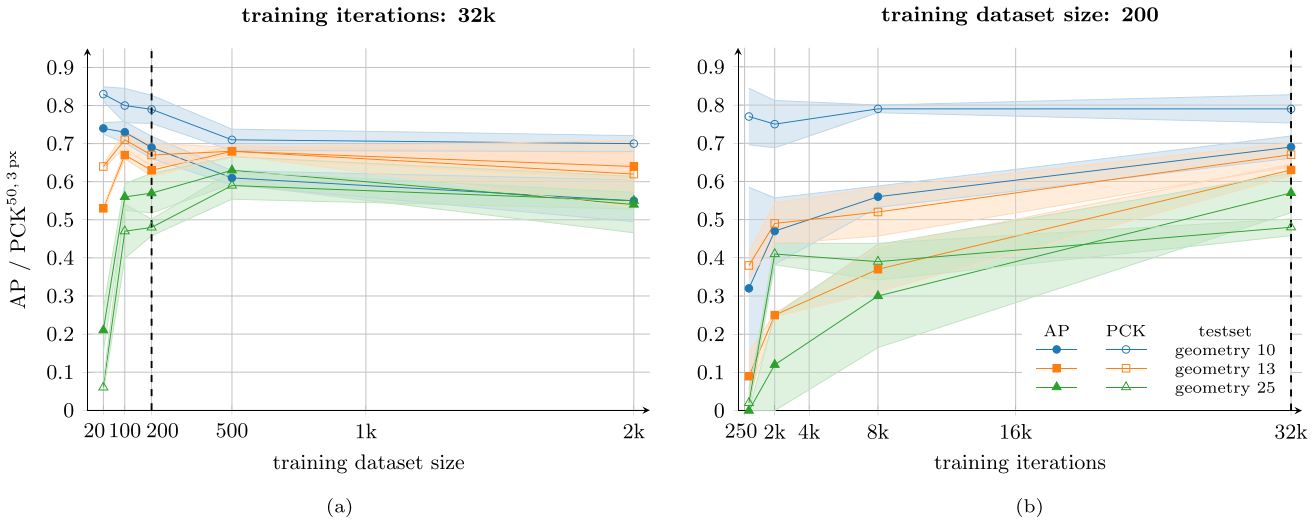

Fig. 4. Results of the baseline model training for (a) different training dataset sizes and (b) varying numbers of training iterations. The plots show AP and PCK for each geometry, with the lines representing the mean over three random seeds and the shaded regions indicating the corresponding standard deviation. The dashed vertical line denotes corresponding experiments between (a) and (b).

PCK from roughly 500 training images onward. The observed differences can be attributed to the varying number of collectors visible per image: smaller geometries show many collectors per frame, meaning fewer images are needed, whereas larger geometries expose fewer collectors, requiring more images to reach comparable coverage. This hypothesis is supported by an alternative representation provided in Appendix B, where the same metrics are plotted against the total number of visible collectors in the training dataset, revealing an optimal range of approximately 5000–50,000 collectors. Beyond this point, adding more training images does not improve performance and may even reduce it, likely because the additional samples from the same scene offer little new visual information. Moreover, because the training dataset consists solely of synthetic images, increasing the number of samples may cause the model to overfit to synthetic-specific artifacts that do not generalize well to real-world data.

An additional observation is that for geometries 10 and 13, PCK follows the same trend as AP but at consistently higher values. In contrast, for geometry 25, PCK initially lags behind and only converges at around 500 images. This difference can be explained by the fact that the test dataset for geometry 25 contains more situations in which objects overlap. This is directly related to the larger dimensions of the collectors, which lead to a higher number of occluded corners that are more difficult to detect and therefore appear to require more training data. A visual inspection of the model predictions supports this interpretation.

Although the optimal dataset sizes differ across geometries (20 images for geometry 10, 100 for geometry 13, and 500 for geometry 25, aligning with the recommended number of collectors for each case), for the subsequent experiments, we select 200 images as a consistent training dataset size, representing a practical compromise between performance and rendering cost across all three geometries.

A reduced dataset size may also allow for a reduction in the number of training iterations required for convergence. Therefore and to provide a point of reference for the fine-tuning scenario, we conduct an additional investigation into the effect of the number of training iterations using the previously identified optimal dataset size of 200 images. The results in Fig. 4(b) indicate that shortening the training duration from 32 000 to less iterations is not appropriate. In fact, the AP metric in particular suggests that longer training could yield further improvements, so additional computational effort appears promising in principle. However, investigating this in detail lies beyond the scope and resource budget of this study. The results further show that, while the number of required training samples saturates at a certain point, maintaining a sufficiently high training duration remains important. A plausible explanation is that, once enough variability in camera poses and collector orientations is present in the data, further improvements depend more strongly on variation in object scales, which is introduced primarily through data augmentation and becomes more effective with longer training. The role of this data augmentation operation is considered further in the next section. Overall, we recommend at least 32 000 training iterations for baseline training as a reasonable and efficient choice.

3.2. Universal training

For the universal training setup, we rendered 200 synthetic training images for each heliostat collector geometry documented earlier. The corresponding datasets were merged, and the universal model variants were trained for 32 000 iterations. The comparison of baseline at 200 images and universal models for each test geometry is shown in Fig. 5.

When evaluating the universal inclusive model, we observe a notable drop in performance for geometry 10, with AP decreasing from 0.69 to 0.39 and PCK from 0.79 to 0.59 when compared to the baseline model. Geometry 13 shows only a modest reduction (AP: 0.63 → 0.57,

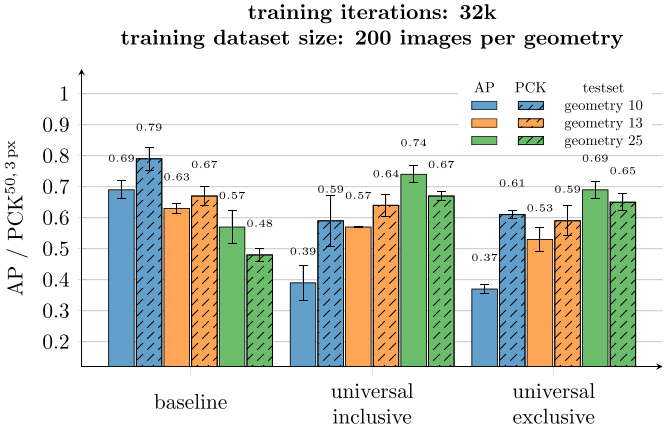

Fig. 5. Comparison of universal model training results with the specialized baseline models at a training dataset size of 200 for the three test geometries. Solid bars show AP and hatched bars show PCK, with mean values displayed above the bars and whiskers indicating standard deviations.

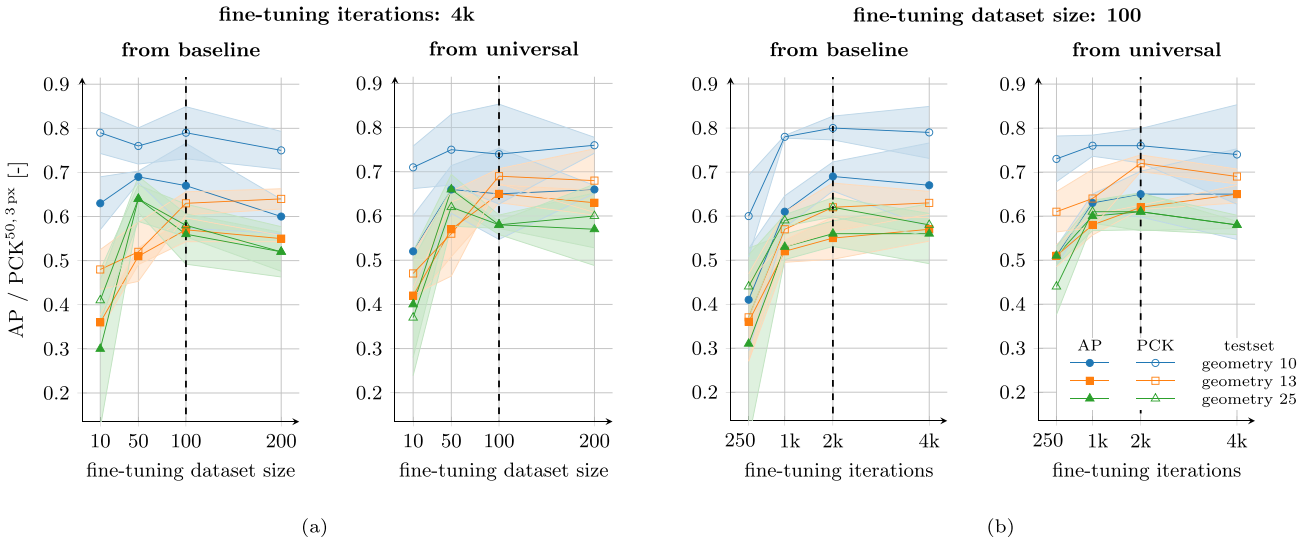

Fig. 6. Results of the fine-tuning experiments for (a) different training-dataset sizes and (b) varying numbers of training iterations, when starting from either a baseline or a universal model. The plots show AP and PCK for each geometry, with the lines representing the mean over three random seeds and the shaded regions indicating the corresponding standard deviation. The dashed vertical lines denote the final recommended fine-tuning strategy.

PCK: 0.67 → 0.64), suggesting that a universal model remains a viable option for this medium-size collector. In contrast, for geometry 25 the universal model even outperforms the baseline (AP: 0.57 → 0.74, PCK: 0.48 → 0.67), indicating potential advantages of a unified approach for larger collectors. While baseline training exhibits decreasing performance with increasing collector size, the universal training shows the opposite trend, with larger geometries benefiting more from the shared model. Part of this trend may be influenced by limitations of the available test data, as each test dataset is relatively small. For geometry 10 in particular, the universal model performs poorly on a few specific images that contain a high number of visible collectors (e.g., the images at the far left and far right of the top row in Fig. 2), which disproportionately lowers its reported performance. However, the fact that the baseline model handles these images well suggests that the issue is not primarily caused by inadequate data for geometry 10 but may instead result from the changed data balance and shared representation introduced by universal training. These observations also align with earlier findings that increased variance in poses, orientations, and object appearance can disproportionately affect smaller geometries. To rule out potential confounding factors related to the training setup, we conducted ablation experiments on ground truth Gaussian kernel size, augmentation settings, and training-data balancing. Although these experiments did not produce significant improvements, they underline the relevance of the random resize augmentation (see Appendix C).

For the universal exclusive case, performance remains similar to the inclusive case, albeit with slightly reduced detection rates for both objects and keypoints, indicating that generalization to unseen geometries is possible but comes with a performance penalty. Overall, while universal training appears feasible for geometries 13 and 25, the current evidence is insufficient to make a firm recommendation on when universal versus baseline training should be preferred. More test geometries and larger test datasets would be required for a definitive assessment. Nonetheless, for applications that do not require the highest reliability across all geometries, the universal model remains a practical choice. It still offers a relatively promising keypoint detection rate even for geometry 10 and can even outperform the baseline for larger collectors, such as geometry 25, without additional training effort. However, when consistently high reliability is required, a geometry-specific baseline model or a fine-tuned universal model may be the preferable approach. The potential of such fine-tuning is examined in the following section.

3.3. Fine-tuning

To assess whether fine-tuning is a viable alternative to baseline or universal training, and whether the universal model can provide a suitable starting point, we conducted additional experiments focusing on the required training dataset size and the number of training iterations. As outlined in Section 2, all fine-tuning experiments begin at a training iteration count of 4000. When fine-tuning from baseline models, for geometries 10 and 25 we load the baseline model trained on geometry 13 with 200 images, and for geometry 13 we start from the baseline model trained on geometry 10. When fine-tuning from the universal model, we select the universal-exclusive case to emulate transfer to an unseen geometry. The results of the fine-tuning experiments are presented in Fig. 6.

When comparing the baseline and the universal model as starting points, we observe that fine-tuning from the universal model yields slightly higher and more stable performance for geometries 13 and 25. This indicates that universal pre-training can be effectively leveraged for downstream geometry adaptation. Nonetheless, in the absence of a universal model, fine-tuning a baseline model remains a practical alternative capable of achieving comparable performance, especially as it offers a slight advantage for geometry 10.

Focusing on fine-tuning from the universal model, noting that the same trends appear for the fine-tuning from baseline, we observe in Fig. 6(a) that its performance saturates with a lower amount of data than the baseline training. Whereas the baseline case required between 200 and 500 images to reach its plateau, and a joint dataset size for all three geometries was difficult to determine without performance losses, fine-tuning from the universal model saturates at 100 images for all three test geometries. For a larger fine-tuning dataset size of 200 images, we observe a decrease or stagnation in performance similar to the baseline case, indicating that the same underlying mechanism is present: additional samples from the same synthetic scene introduce little new visual information and may even enhance synthetic-specific artifacts that do not contribute to real-world transfer.

Examining the fine-tuning duration shown in Fig. 6(b), we find that 2000 iterations constitute an adequate choice, as extending to 4000 iterations yields no meaningful improvement in object or keypoint detection and therefore does not justify the additional computational cost.

Table 3
Overview of the key metrics (AP and PCK) for all investigated transfer approaches across the three geometries. For the baseline models, the configuration at 200 images is reported. For the universal model, the inclusive variant is used, and for the fine-tuning approaches, the metrics obtained using the recommended strategy (100 images, 2000 iterations) are presented. Averages are computed over the individual performance values.

approach	geometry 10		geometry 13		geometry 25		average	
	AP	PCK	AP	PCK	AP	PCK	AP	PCK
baseline	0.69 ± 0.029	0.79 ± 0.037	0.63 ± 0.015	0.67 ± 0.031	0.57 ± 0.052	0.48 ± 0.022	0.63 ± 0.062	0.65 ± 0.138
universal	0.39 ± 0.056	0.59 ± 0.082	0.57 ± 0.003	0.64 ± 0.035	0.74 ± 0.028	0.67 ± 0.013	0.56 ± 0.155	0.63 ± 0.057
fine-tuning (base)	0.69 ± 0.033	0.80 ± 0.027	0.55 ± 0.048	0.62 ± 0.055	0.56 ± 0.029	0.62 ± 0.022	0.60 ± 0.076	0.68 ± 0.096
fine-tuning (univ)	0.65 ± 0.052	0.76 ± 0.039	0.62 ± 0.005	0.72 ± 0.019	0.61 ± 0.041	0.61 ± 0.043	**0.63**± 0.037	**0.69**± 0.074

Table 4
Compute time and VRAM usage comparison for rendering and training. The reported values reflect computations executed on a single *NVIDIA A100 SXM4 80 GB*, which served as the hardware platform for the experiments in this study.

operation	GPU-hours	peak VRAM usage [GB]
rendering (1 image)	0.1	26
training (1000 iterations)	5	80

Overall, the optimal fine-tuning configuration therefore uses the universal model as a starting point and requires 100 images and 2000 iterations. Using this setup, fine-tuning for geometry 10 achieves an AP of 0.65 and a PCK of 0.76, clearly surpassing the universal-inclusive model and reaching detection accuracy comparable to the baseline configuration (AP: 0.69, PCK: 0.79). For geometry 13, fine-tuning yields an AP of 0.62 and a PCK of 0.72, slightly above the universal-inclusive training and comparable to the baseline (AP: 0.63, PCK: 0.67). For geometry 25, we obtain an AP of 0.61 and a PCK of 0.61, surpassing the performance of the baseline model (AP: 0.57, PCK: 0.48) but still lagging behind the superior universal-inclusive model (AP: 0.74, PCK: 0.67). Table 3 presents an overview of the key metrics. In summary, these results indicate that fine-tuning, preferably starting from a universal model, is a robust strategy for achieving competitive performance across the tested geometries and offers a reliable and consistent procedure for model adaptation.

3.4. Time required for model transfer

In this section, we discuss the time and computational resources required for rendering and training to offer an additional criterion for selecting an optimal geometry transfer strategy. Table 4 summarizes the GPU-hours and peak VRAM usage for both operations. Training utilizes the full VRAM capacity of the employed hardware, which implies that smaller GPUs would lead to longer processing times. In contrast, rendering is far less demanding and peaks at 26 GB, enabling parallelization and the use of less powerful hardware without any expected time penalties.

Considering baseline training using 200 images and 32 000 iterations, rendering requires approximately 20 h and training 160 h, resulting in a total processing time of 180 h on a single GPU. This total scales down when using a multi-GPU setup as in this work. For the universal-inclusive model, 6400 images (32 geometries × 200 images each) were rendered, requiring approximately 640 GPU-hours. Training for 32 000 iterations again required 160 GPU-hours, summing to a total of 800 GPU-hours. For the fine-tuning strategy with 100 images and 2000 iterations, rendering consumes 10 GPU-hours and training requires another 10 GPU-hours, totaling 20 GPU-hours. When fine-tuning from a baseline model, a one-time cost of 180 GPU-hours must be considered. When fine-tuning from the universal-exclusive model as in this study (23 geometries × 200 images = 4600 images), the corresponding one-time overhead is 620 GPU-hours.

From a detection-performance perspective, the previous section identified fine-tuning from the universal model as the preferred approach for geometry transfer. While fine-tuning from a baseline model also performs well and remains a reasonable choice when a universal model is not yet available, the computational analysis presented here shows that rendering is not the primary cost driver in terms of time and memory usage. Therefore, it is reasonable to generate a comprehensive synthetic database covering a diverse set of geometries and train a universal model, which can be applied directly and, if necessary, fine-tuned for specific use cases, making this our final recommended strategy.

4. Conclusion and outlook

The present study provides a systematic evaluation of strategies for transferring deep learning models across the wide range of heliostat collector geometries used in existing CST tower plants. Based on a compiled database of representative real-world geometries and an extensive synthetic dataset generated with our rendering framework, we evaluated geometry-specific baseline models, universal models trained jointly across all geometries, and fine-tuning approaches starting from either the baseline or a universal model.

The results show that baseline models perform reliably on their respective geometries, while universal training suffers from reduced and geometry-dependent performance. Fine-tuning consistently provides robust adaptation and achieves detection rates comparable to geometry-specific baselines, making it the most practical strategy for transferring DNNs to new heliostat designs.

These findings highlight the potential of synthetic data and targeted fine-tuning to support scalable and efficient deployment of such models in CST condition monitoring systems. A functional model for a new geometry can be obtained by rendering as few as 100 synthetic images and fine-tuning an already available model for 2000 iterations, requiring a total of approximately 20 h on a single GPU. This enables rapid, flexible adaptation of the system to the large design diversity found in practice. We expect that similar transfer may also apply to aspects such as lighting conditions, sky cover, and site-specific features, in case a trained model does not generalize well, although this remains to be examined in future work.

Despite these strengths, several limitations remain. The available real-world test datasets are small, which affects the reliability of the performance estimates. Furthermore, the reasons for the instability of universal training, particularly for small collectors, could not be fully identified and may involve factors not captured in the training dataset creation process. Future work may therefore involve expanding the real-world evaluation dataset and investigating whether any techniques can improve the stability of universal training. Finally, integrating the approach into real-time or onboard drone systems would enable fully automated, large-scale heliostat condition monitoring in operational environments.

In summary, while universal training alone is not yet reliable, fine-tuning emerges as a practical, scalable, and effective approach for deploying deep learning models across the diverse heliostat collector geometries found in CST plants.

CRediT authorship contribution statement

Rafal Broda: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Alexander Schnerring:** Writing – review & editing, Methodology. **Michael Nieslony:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Tobias Wagner:** Investigation, Data curation. **Dominik Schnaus:** Software, Methodology. **Niels Algner:** Writing – review & editing, Data curation. **Marc Röger:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization. **Sonja Kallio:** Supervision, Resources. **Rudolph Triebel:** Supervision, Funding acquisition. **Robert Pitz-Paal:** Supervision, Funding acquisition.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT (GPT-4o and GPT 5, OpenAI) to improve the readability and language of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was partly funded by the German Federal Ministry for the Environment, Climate Action, Nature Conservation and Nuclear Safety through the AuSeSol-AI project (Grant Agreement No. 67KI21007A), based on a decision by the German Bundestag. Another part was funded by the European Union under Grant Agreement No. 101234781 within the Sun-DT project. Additionally, we gratefully acknowledge the computational and data resources provided through the joint high-performance data analytics (HPDA) project “terrabyte” of the German Aerospace Center (DLR) and the Leibniz Supercomputing Center (LRZ). Lastly, we extend our thanks for the use of image data from the Plataforma Solar de Almería (owned and operated by CIEMAT) and the Solar Tower Jülich (owned and operated by the German Aerospace Center).

Appendix A. Heliostat modeling parameters

See Table A.1.

Table A.1
Parameters used to model heliostats and build scenes in the rendering environment. Note that given dimensions are estimated and rounded and do not necessarily represent exact dimensions of real-world collector geometries.

id	facet number [-]		facet size [m]		inter-facet gap [m]		middle gap [m]	collector area [m^2]	pylon height [m]	pylon diameter [m]	number of heliostats in scene [-]		
	x	y	x	y	x	y					field 1	field 2	field 3
1	1	1	1.1	1.1	0.01	0.01	0.00	1.21	1.0	0.10	10 668	10 034	8695
2	1	1	1.5	1.5	0.01	0.01	0.00	2.25	1.2	0.12	8441	7940	6881
3	1	1	1.6	1.2	0.01	0.01	0.00	1.92	1.2	0.12	8856	8329	7219
4	1	1	2.4	1.8	0.01	0.01	0.00	4.32	1.2	0.13	6150	5784	5013
5	2	1	2.3	3.3	0.01	0.01	0.50	16.86	2.0	0.22	2946	2771	2401
6	2	1	2.4	3.0	0.01	0.01	0.50	15.93	1.8	0.22	2946	2771	2401
7	2	1	1.8	3.0	0.01	0.01	0.00	10.83	1.8	0.20	3740	3518	3049
8	2	2	2.0	2.6	0.01	0.01	0.00	20.89	3.0	0.25	2415	2272	1969
9	2	2	2.7	1.8	0.01	0.01	0.40	20.97	2.5	0.25	2451	2305	1998
10	2	2	1.6	1.3	0.01	0.01	0.00	8.38	1.7	0.20	4346	4088	3543
11	2	2	1.0	1.0	0.01	0.01	0.00	4.04	1.5	0.20	6486	6100	5287
12	2	4	3.0	1.3	0.05	0.01	0.00	31.64	3.0	0.25	1831	1722	1492
13	2	6	3.0	1.0	0.01	0.01	0.75	40.90	3.0	0.50	1666	1567	1358
14	3	1	0.8	1.5	0.01	0.01	0.00	3.63	1.3	0.12	6475	6091	5278
15	3	3	1.8	1.8	0.01	0.01	0.00	29.38	3.2	0.28	1946	1831	1587
16	3	3	1.0	1.0	0.01	0.01	0.00	9.12	1.9	0.25	4188	3939	3414
17	4	4	1.0	1.0	0.01	0.01	0.40	17.85	2.5	0.25	2925	2752	2385
18	4	4	1.0	1.0	0.01	0.01	0.00	16.24	2.5	0.28	2925	2752	2385
19	4	6	1.5	1.0	0.01	0.01	0.75	41.02	3.0	0.50	1666	1567	1358
20	4	7	3.3	1.3	0.01	0.01	0.00	121.19	5.5	0.40	607	571	495
21	4	8	3.3	1.3	0.01	0.01	0.00	138.52	6.5	0.50	561	527	457
22	5	5	1.2	1.2	0.01	0.01	0.00	36.48	4.0	0.30	1662	1563	1355
23	5	7	2.3	1.4	0.01	0.01	0.00	113.78	6.0	0.50	670	630	546
24	6	6	1.8	1.8	0.01	0.01	1.00	128.57	6.5	0.50	658	619	536
25	7	5	1.4	2.5	0.01	0.01	0.00	123.64	7.0	0.50	616	580	502
26	7	5	1.5	2.2	0.01	0.01	0.00	116.58	6.5	0.50	663	623	540
27	7	5	1.8	1.8	0.01	0.01	0.00	114.45	6.0	0.50	643	605	524
28	8	5	2.2	1.8	0.01	0.01	1.00	168.78	6.0	0.60	424	399	346
29	8	5	2.9	2.3	0.01	0.01	1.00	280.08	7.0	0.70	264	248	215
30	8	8	1.3	1.3	0.01	0.01	0.00	109.62	6.5	0.60	698	657	569
31	8	8	1.0	1.0	0.01	0.01	0.50	69.16	5.0	0.50	1066	1002	869
32	9	6	1.5	2.2	0.01	0.01	0.00	179.94	8.0	0.70	459	431	374

Appendix B. Baseline training: number of collectors

See Fig. B.1.

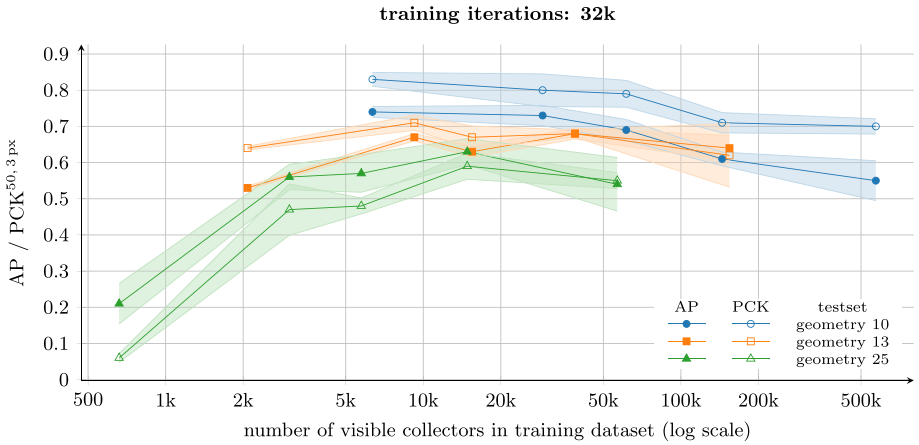

Fig. B.1. Baseline model performance versus number of visible collectors in the training dataset. The plot shows AP and PCK for each geometry, with the lines representing the mean over three random seeds and the shaded regions indicating the corresponding standard deviation.

Appendix C. Universal training: ablation study

To rule out training-setup-related causes for the reduced performance of the universal approach, we conducted a series of ablation experiments focusing on parameters that could contribute to the differences between the baseline and universal settings. All experiments were run with a single seed to keep computational effort manageable. Based on the variability observed throughout this work, we consider this sufficient for the scope of the present analysis.

We first examined whether extending the training time by doubling the number of iterations would improve performance, since a more diverse dataset could benefit from longer optimization. This modification did not produce any measurable improvements. Motivated by the comparatively low AP values of the universal model, we then investigated the influence of the Gaussian kernel used to generate the ground-truth heatmap for object detection. In addition to the distance-scaled kernel, we tested a fixed kernel size of 16 px and a kernel derived from bounding-box dimensions as described by Law and Deng [22] and implemented by Zhou et al. [19] in the work of the original model. The fixed kernel size slightly increased keypoint-detection performance, but object detection continued to underperform. Similarly, an adapted kernel size did not show any improvement.

Next cale-and-crop data augmentation. Avoiding downscaling before cropping resulted in a small improvement for geometry 10 but still fell short of baseline performance. Excluding upscaling led to lower detection rates for geometries 10 and 13, and removing both downscaling and upscaling resulted in an even stronger decline. These findings highlight the importance of this data augmentation operation. Interestingly, geometry 25 showed little sensitivity to these changes.

To explore imbalances in the joint training dataset, we generated 20 additional **synthetic** test images per geometry and evaluated AP performance (see Fig. C.3). A dependence on collector size was visible even in the synthetic domain. Interestingly, geometry 10, although consistently underperforming on real-world data, did not exhibit reduced performance in the synthetic tests. This observation suggests that, beyond object scale, additional factors not captured by the synthetic setup may influence universal training. It also points to reduced robustness of the universal model, indicating that it may fail for other small or large geometries beyond those examined in this study. Finally, for geometries with AP values below 0.8 (with geometries 10 and 25 among them), we increased their representation in the training data by a factor of two. As shown in Fig. C.2, this adjustment did not lead to improvements.

Finally, motivated by recent advances in transformer-based vision models, we investigated the replacement of the ResNet50 encoder with a ViT-B/16 [23] backbone initialized from the large-scale DINOv3 pre-training framework [24]. Two configurations were evaluated: end-to-end training and training with frozen DINOv3 weights. The end-to-end configuration achieved promising keypoint detection performance for the three geometries,

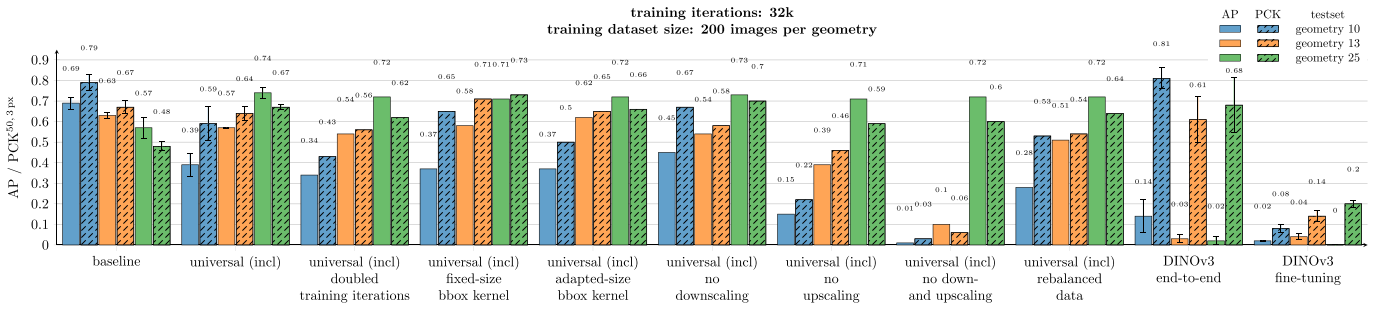

Fig. C.2. Evaluation of alternative training settings for the universal-inclusive scenario. The experiments show no measurable improvement across the three test geometries, confirming the findings reported in the main section.

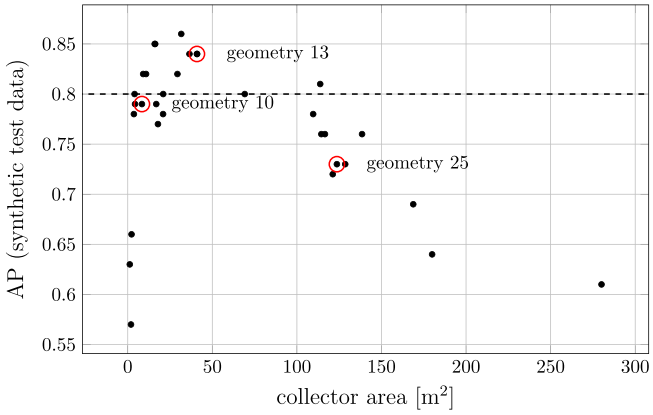

Fig. C.3. AP of the universal-inclusive model as a function of collector area, computed from additional synthetic validation data consisting of 20 images for each documented geometry. Test geometries are highlighted in red. Data rebalancing is applied to all geometries with AP values below 0.8, indicated by the dashed horizontal line. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

exceeding the performance of the universal model, but failed to achieve competitive object detection performance. Inspection of the predictions revealed that heliostats were generally detected, yet the predicted bounding boxes were systematically smaller than the ground truth. This behavior may indicate incomplete convergence, which is consistent with reports that transformer-based architectures often require substantially longer training schedules. Although longer training could potentially improve object detection performance and further exploit the architecture's strong keypoint detection capabilities, the ViT backbone incurs a substantially higher computational cost than CNN-based alternatives when processing high-resolution imagery. As a result, inference time increased from 0.44 s to 17.87 s per 6000×4000 image, making the architecture unattractive for practical use despite its promising performance. In contrast, training with frozen DINOv3 weights resulted in poor performance for both object and keypoint detection, suggesting that the learned representations do not transfer effectively to heliostat imagery without task-specific adaptation. Overall, the results indicate that, despite their strong performance on large-scale vision benchmarks, transformer-based backbones do not provide a clear advantage for the considered task and training setup.

Overall, these ablations show that universal training remains challenging, and the reduced performance, particularly for geometry 10, cannot be attributed to the training settings examined in this study.

Data availability

The data that support the findings of this study are available from the corresponding author upon request.

References

[1] F. Schöniger, R. Thonig, G. Resch, J. Lilliestam, Making the sun shine at night: comparing the cost of dispatchable concentrating solar power and photovoltaics with storage, Energy Sources B Econ. Plan. Policy 16 (1) (2021) 55–74.
[2] R. Schäppi, D. Rutz, F. Dähler, A. Muroyama, P. Haueter, J. Lilliestam, A. Patt, P. Furler, A. Steinfeld, Drop-in fuels from sunlight and air, Nature 601 (7891) (2022) 63–68.
[3] S.A. Jones, K.W. Stone, Analysis of Strategies to Improve Heliostat Tracking at Solar Two, Tech. Rep. SAND99-0092C, Sandia National Laboratories, 1999.
[4] R.A. Mitchell, G. Zhu, A non-intrusive optical (NIO) approach to characterize heliostats in utility-scale power tower plants: methodology and in-situ validation, Sol. Energy 209 (2020) 431–445.
[5] J. Yellowhair, P.A. Apostolopoulos, D.E. Small, D. Novick, M. Mann, Development of an aerial imaging system for heliostat canting assessments, AIP Conf. Proc. 2445 (1) (2022) 120024.
[6] W. Jessen, M. Röger, C. Prahl, R. Pitz-Paal, A two-stage method for measuring the heliostat offset, AIP Conf. Proc. 2445 (1) (2022) 070005.
[7] J.J. Krauth, C. Happich, N. Algner, R. Broda, A. Kämpgen, A. Schnerring, S. Ulmer, M. Röger, HelioPoint – a fast airborne calibration method for heliostat fields, J. Sol. Energy Eng. 146 (6) (2024) 061005.
[8] J.A. Carballo, J. Bonilla, M. Berenguel, J. Fernández-Reche, G. García, New approach for solar tracking systems based on computer vision, low cost hardware and deep learning, Renew. Energy 133 (2019) 1158–1166.
[9] G. Zhu, C. Augustine, R. Mitchell, et al., HelioCon: a roadmap for advanced heliostat technologies for concentrating solar power, Sol. Energy 264 (2023) 111917.
[10] B. Liu, A. Sonn, A. Roy, B. Brewington, Deep learning method for heliostat instance segmentation, in: SolarPACES Conference Proceedings, vol. 1, 2024, https://doi.org/10.52825/solarpaces.v1i.735
[11] R. Broda, A. Schnerring, J.J. Krauth, M. Röger, R. Pitz-Paal, Towards deep learning based airborne monitoring methods for heliostats in solar tower power plants, in: Advances in Solar Energy: Heliostat Systems Design, Implementation, and Operation, vol. 12671, SPIE, 2023, p. 1267108.
[12] R. Broda, A. Schnerring, D. Schnaus, M. Nieslony, J.J. Krauth, M. Röger, S. Kallio, R. Triebel, R. Pitz-Paal, Bridging the sim2real gap: training deep neural networks for heliostat detection with purely synthetic data, Sol. Energy 300 (2025) 113728.
[13] Heliostat Consortium for Concentrating Solar-Thermal Power, HelioCon database: power tower plant database, heliostat database https://heliocon.org/resources/plant_information_overview.html (accessed: 28 September 2025).
[14] A. Pfahl, J. Coventry, M. Röger, F. Wolfertstetter, J.F. Vásquez-Arango, F. Gross, M. Arjomandi, P. Schwarzbözl, M. Geiger, P. Liedke, Progress in heliostat development, Sol. Energy 152 (2017) 3–37, progress in Solar Energy Special Issue: Concentrating Solar Power (CSP).
[15] T. Wagner, Development of a Geometry-Agnostic Deep Learning Model for Detection of Heliostats in Solar Power Tower Plants (Master's thesis), RWTH Aachen, 2025.
[16] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 770–778.
[17] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: a large-scale hierarchical image database, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.
[18] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in: European Conference on Computer Vision (ECCV), 2020, pp. 213–229.
[19] X. Zhou, D. Wang, P. Krähenbühl, Objects as points, arXiv preprint arXiv:1904.07850, 2019.
[20] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, in: International Conference on Learning Representations (ICLR), 2015.
[21] L.N. Smith, N. Topin, Super-convergence: very fast training of neural networks using large learning rates, in: Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications, vol. 11006, SPIE, 2019, p. 1100612.
[22] H. Law, J. Deng, Cornernet: detecting objects as paired keypoints, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 734–750.
[23] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: transformers for image recognition at scale, in: International Conference on Learning Representations (ICLR), 2021.
[24] O. Siméoni, H.V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. Yi, M. Ramamonjisoa, et al., Dinov3, arXiv preprint arXiv:2508.10104, 2025.