Contents lists available at ScienceDirect

Expert Systems With Applications

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

Highlights

SAM-Net: Self-Attention based Feature Matching with Spatial Transformers and Knowledge Distillation

Expert Systems With Applications xxx (xxxx) xxx

Benjamin Kelenyi, Victor Domsa, Levente Tamas*

- Geometric key-point feature extraction for 2D vision with spatial transformers.
- Feature extraction with Knowledge distillation and self-attention.
- Robust position estimation from multi-view camera systems.

Contents lists available at ScienceDirect

Expert Systems With Applications

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

SAM-Net: Self-Attention based Feature Matching with Spatial Transformers and Knowledge Distillation

Benjamin Kelenyi, Victor Domsa, Levente Tamas *

Technical University of Cluj-Napoca, Memorandumului 28, 400114, Romania

ARTICLE INFO

Keywords:
Geometric features extraction
Self-attention
Knowledge-distillation
Spatial transformers
Pose estimation

ABSTRACT

In this research paper, we introduce a novel approach to enhance the performance of 2D feature matching and pose estimation through the integration of a hierarchical attention mechanism and knowledge distillation. Our proposed hierarchical attention mechanism operates at multiple scales, enabling both global context awareness and precise matching of 2D features, which is crucial for various computer vision tasks. To further improve our model's performance, we incorporate insights from an existing model PixLoc (Sarlin et al., 2021) through knowledge distillation, effectively acquiring its behavior and capabilities by ignoring dynamic objects. SAM-Net outperforms state-of-the-art methods, validated on both indoor and outdoor public datasets. For the indoor dataset, our approach achieves remarkable AUC (5°/10°/20°) scores of 55.31/71.70/83.37. Similarly, for the outdoor dataset, we demonstrate outstanding AUC values of 26.01/46.44/63.61. Furthermore, SAM-Net achieves top ranking among published methods in two public visual localization benchmarks, highlighting the real benefits of the proposed method. The code and test suite can be accessed at link.[1]

1. Introduction

Local feature matching (Kelenyi & Tamas, 2023) is a fundamental problem in computer vision and robotics, with applications in structure-from-motion (SfM) (Schonberger & Frahm, 2016), relative pose estimation (Frohlich, Tamas, & Kato, 2019), simultaneous localization and mapping (SLAM) (Domsa, Konievic, Kelenyi, & Tamas, 2023), and various other areas (Blaga, Militaru, Mezei, & Tamas, 2021; Farhat, Chaabouni-Chouayakh, & Ben-Hamadou, 2023; Molnár & Tamás, 2023; Pop & Tamas, 2022). The goal of local feature matching is to establish correspondences between image points across different views, which can later be used to recover the underlying 3D structure of the environment. This is typically achieved by detecting and describing distinctive features in images, such as corners, blobs, or edges, and then matching these features across different views.

In recent years, the integration of deep learning techniques has significantly enhanced local feature matching, leading to state-of-the-art performance across various applications (Szegedy et al., 2015). However, challenges persist due to factors like changes in illumination, viewpoint, scale, and occlusion. Illumination variations can distort object appearance, while alterations in viewpoint and scale introduce geometric transformations that traditional methods struggle to handle (Lowe, 2004a). Furthermore, occlusion, where objects are partially obstructed, complicates the identification of critical feature points for matching (see Fig. 1).

To address these challenges, researchers have proposed a variety of algorithms and techniques, such as hierarchical feature matching, deep learning-based methods, and real-time feature tracking. Some of the most influential papers in this field include SIFT (Lowe, 2004b) (Scale-Invariant Feature Transform), SURF (Speeded-Up Robust Features) (Bay, Tuytelaars, & Van Gool, 2006) and ORB (ORB: An efficient alternative to SIFT or SURF) (Rublee, Rabaud, Konolige, & Bradski, 2011). These methods were widely adopted in SLAM and SfM pipelines and significantly improved the accuracy and robustness of local feature matching. Another promising direction for improving local feature matching is knowledge distillation. Knowledge distillation (Gou, Yu, Maybank, & Tao, 2021) is a technique that involves transferring knowledge from a large, complex model, such as ResNet (Targ, Almeida, & Lyman, 2016) or VGG (Sengupta, Ye, Wang, Liu, & Roy, 2019), known as the *teacher*, to a smaller, simpler model (Kolodiazhnyi, 2022), known as the *student*. This approach was shown to be effective in reducing the computational complexity of local feature matching while maintaining or improving its accuracy and robustness.

The code (and data) in this article has been certified as Reproducible by Code Ocean: (https://codeocean.com/). More information on the Reproducibility Badge Initiative is available at https://www.elsevier.com/physical-sciences-and-engineering/computer-science/journals.

* Corresponding author.
E-mail addresses: benjamin.kelenyi@aut.utcluj.ro (B. Kelenyi), victor.domsa@campus.utcluj.ro (V. Domsa), levente.tamas@aut.utcluj.ro (L. Tamas).

[1] https://benjaminkelenyi.github.io/samnet/

https://doi.org/10.1016/j.eswa.2023.122804
Received 20 June 2023; Received in revised form 29 November 2023; Accepted 29 November 2023

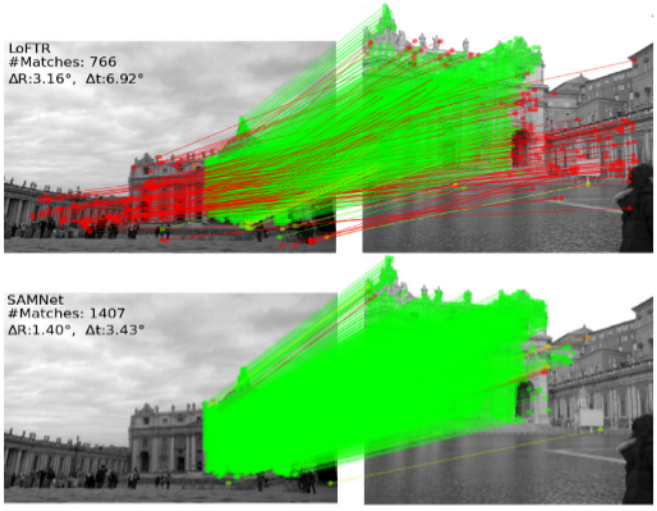

Fig. 1. This example demonstrates that SAM-Net is capable of finding more accurate correspondences than LoFTR (Sun, Shen, Wang, Bao, & Zhou, 2021) even in challenging conditions (green: correct feature matching, red: false feature matching).

Self-attention methods (Wang, Zhang, Yang, Peng, & Stiefelhagen, 2022), on the other hand, gained popularity for local feature matching due to their ability to capture long-range dependencies between the feature descriptors and improve their discriminability and robustness. The self-attention module is essential to achieving this objective and has the potential to increase the pipeline's accuracy and efficiency for feature matching. The ability to model long-range dependencies between feature descriptors, which is challenging with conventional feature matching methods that rely on local feature descriptors, is one of the main benefits of self-attention methods. The self-attention module can increase the discriminability and robustness of the feature descriptors by capturing the pairwise relationships between them, resulting in more precise and reliable feature matching.

1.1. Primary motivations

Our primary motivation is to advance local feature matching in computer vision and robotics. This process is critical for establishing correspondences between image points across different views in order to recover 3D environmental structures. The following challenges, prevent the creation of a deep local feature matcher for detector-free techniques:

(a) Standard detector-free methods typically begin with a Convolution Neural Network (CNN) for feature extraction, followed by the incorporation of Transformer layers to capture long-range context. However, there is a context coverage gap between the Transformer's global scope and the CNN's local focus, which may block deep feature interaction later on.
(b) In scenes with repetitive patterns or symmetrical structures, CNN's translation invariance can cause ambiguity. To handle this, current detector-free methods use absolute position encodings prior to Transformers, but the positional information reduces as the Transformer layers deepen. Humans, on the other hand, associate objects across observations using both absolute and relative positions.
(c) While network depth appears to be important for feature matching, LoFTR's linear Transformer finds it difficult to effectively aggregate context in deeper layers. This is because its context-independent self-attention method cannot sufficiently model the relevance of all keypoints.

Motivated by the aforementioned insights, we propose *Self-Attention based Feature Matching with Spatial transformers and Knowledge Distillation (SAM-Net)*, an efficient local feature matching technique.

1.2. Innovation aspects

Our proposed technique introduces novel aspects to the field of local feature matching in computer vision and robotics. To advance 2D feature matching, our proposed method combines the strengths of LoFTR (Sun et al., 2021), a state-of-the-art localization model, with knowledge distillation from PixLoc (Sarlin et al., 2021). By distilling knowledge from Pixloc at different levels, which ignores dynamic objects, we can focus on learning robust and discriminative features specific to static objects. The exclusion of dynamic objects during training improves the model's ability to accurately localize static objects even in the presence of dynamic elements (e.g. cars, bicycles, pedestrians), thereby improving overall performance.

1.3. Contributions

We summarize our contributions as follows:

- A hierarchical attention mechanism that performs attention operations at multiple scales, allowing for both global context awareness and precise matching of 2D features;
- A novel approach based on knowledge distillation to enhance our model's performance by integrating insights from an existing model, effectively acquiring its behavior and capabilities by ignoring dynamic objects;
- State-of-the-art performance in two-view pose estimation, surpassing baselines, and demonstrates potential for seamless integration into complex applications through challenging visual localization tasks.

Our goal is to provide a comprehensive understanding of local feature matching and its applications to localize pose estimation using self-attention and transformers, as well as to facilitate the development of more robust and accurate algorithms for this task by focusing only on the relevant parts of the scene by ignoring dynamic objects.

1.4. Sections of the manuscript

The paper is structured as follows: the Introduction contains the main motivation, problem description and a summary of the proposed method. Section 2 presents the related methods from the state-of-the-art including the classical and learning-based methods, while Section 3 presents in detail the proposed method. In Section 4 the experimental parts are described and the paper concludes with the last Conclusion section.

2. Related works

In this section, we briefly survey the existing literature on detector-based, detector-free local feature matching, and transformer-based methods. We start by presenting the early work in this area, which laid the foundation for many of the current techniques. Furthermore, we focus on the most recent methods, including deep learning-based approaches. Table 1 contains a comprehensive list of the top related papers on feature matching.

2.1. Detector-based local feature matching

Detector-based local feature matching is a widely used technique in computer vision that involves detecting distinctive features in an image and then matching them across different images to perform tasks such as object recognition, image retrieval, and 3D reconstruction (Kang, Yang, Yang, & Cheng, 2020). In recent years, significant progress has been made in this area, with several new algorithms being developed to improve the performance and robustness of local feature matching.

One of the earliest methods for local feature matching is the *SIFT* (The Scale Invariant Feature Transform) algorithm (Lowe, 2004a). SIFT is one of the earliest methods for local feature matching. It extracts scale-invariant keypoints by detecting local extrema and describes them using a histogram of gradient orientations. SIFT employs a nearest-neighbor algorithm and a ratio test for matching keypoints across images.

Another popular method is the *SURF* (Bay et al., 2006) (Speeded Up Robust Features) algorithm. *SURF* is a faster and more efficient alternative to SIFT that uses an approximate Laplacian of Gaussian filter to detect scale-invariant keypoints and a modified Haar wavelet descriptor to describe them. SURF also uses a fast approximate nearest-neighbor algorithm called the *kd-tree* to match keypoints across images.

ORB (Rublee et al., 2011) (Oriented FAST and Rotated BRIEF): is another method proposed by Rublee et al. in 2011. ORB combines the FAST corner detector with the BRIEF descriptor and introduces an orientation component to handle rotation invariance. It is designed to be faster and more efficient than SIFT (Lowe, 2004b) and SURF (Bay et al., 2006).

DenseGAP (Kuang, Li, He, Wang, & Zhao, 2022) (Graph-Structured Dense Correspondence Learning with Anchor Points): leverages a graph structure and anchor points to provide reliable prior information for context modeling. It employs a graph-structured network to efficiently propagate multi-level contexts, generating high-resolution feature maps. A coarse-to-fine framework is used for accurate correspondence prediction.

ClusterGNN (Shi et al., 2022) (Cluster-based Coarse-to-Fine Graph Neural Network for Efficient Feature Matching): addresses feature matching using an attentional Graph Neural Network (GNN) architecture. It dynamically separates keypoints into distinct subgraphs through a progressive clustering module, minimizing unnecessary connections. A coarse-to-fine approach is employed to reduce misclassification within images.

Recent research has focused on increasing the efficiency and robustness of detector-based local feature matching. *LIFT* (Learned Invariant Feature Transform) (Yi, Trulls, Lepetit, & Fua, 2016) is a deep neural network-based algorithm that learns invariant local feature descriptors, providing resilience to common image transformations like rotation, scaling, and lighting changes. These learned descriptors can be used for tasks such as image matching, object recognition, and 3D reconstruction, outperforming traditional hand-crafted feature descriptors. Another notable method is *BRISK* (Binary Robust Invariant Scalable Keypoints) (Leutenegger, Chli, & Siegwart, 2011), which uses a binary descriptor to detect and describe local features, exhibiting robustness and computational efficiency. *FREAK* (Fast Retina Keypoint) (Alahi, Ortiz, & Vandergheynst, 2012) is a fast and efficient method using a binary descriptor, suitable for various computer vision applications, including object recognition and tracking.

Deep learning advances in recent years have resulted in the development of excellent algorithms for local feature matching. *SuperPoint* (DeTone, Malisiewicz, & Rabinovich, 2018) is a fast and accurate deep neural network that detects keypoints and computes descriptors. *LF-Net* (Ono, Trulls, Fua, & Yi, 2018) utilizes a Siamese network for matching local features and achieves state-of-the-art performance. *SuperGlue* (Sarlin, DeTone, Malisiewicz, & Rabinovich, 2020) introduces a machine learning-based approach using a graph neural network (GNN), demonstrating significant improvements in matching accuracy. *D2-Net* (Dusmanu et al., 2019) and *R2D2* (Revaud, De Souza, Humenberger, & Weinzaepfel, 2019) are recent methods that have also shown outstanding performance on benchmark datasets.

2.2. Detector-free local feature matching

Detector-free local feature matching is an emerging technique in computer vision that eliminates the need for keypoint detectors. Traditional methods like SIFT (Lowe, 2004b) and SURF (Bay et al., 2006) rely on detectors, but they can be computationally expensive and less effective in certain image conditions. In contrast, detector-free methods directly operate on image pixels, extracting features at regular intervals. This approach is computationally efficient and works well in challenging regions.

An example of a detector-free local feature matching method is Dense SIFT. This method extracts SIFT descriptors at regularly spaced pixels in an image rather than at keypoints. Another example is DAISY (Tola, Lepetit, & Fua, 2010) (Dense, Invariant, and Spatially Augmented Descriptors), which extracts dense features in a grid pattern and uses an adaptive histogram equalization technique to enhance local contrast.

More recently, *ASpanFormer* (Chen et al., 2022) (Detector-Free Image Matching with Adaptive Span Transformer) was proposed: it makes use of a novel attention operation to adjust attention span in a self-adaptive manner. It regresses flow maps to identify the search region, generates a sampling grid of adaptive size, and computes attention across two images in derived regions. ASpanFormer maintains long-range dependencies and fine-grained attention among relevant pixels, eliminating the need for object detection and achieving high accuracy in object matching. The CNN backbone extracts initial features, which are updated with iterative Global Local Attention (GLA) blocks, and a matching module determines final matches.

HTMatch (Cai, Li, Wang, Li, & Liu, 2023) (An efficient hybrid transformer-based graph neural network for local feature matching): uses a hybrid transformer-based GNN for local feature matching. It combines self- and cross-attention to condition feature descriptors between image pairs. This enables efficient attentional aggregation using a single transformer layer. It also introduces a spatial embedding module to enhance spatial constraints and utilizes a seeded GNN architecture for improved efficiency and effectiveness.

DeepMatcher (Xie, Dai, Wang, Li, & Zhao, 2023) (A Deep Transformer-based Network for Robust and Accurate Local Feature Matching): introduces a Slimming Transformer approach for dense pixel-wise matching. SlimFormer is employed to model relevance among all keypoints and achieve long-range context aggregation efficiently. Position encoding, layer-scale strategy, and the Feature Transition Module are introduced to improve performance. DeepMatcher uses deep-narrow transformer layers and a network-based refinement block for more precise matches.

DAN-SuperPoint (Li et al., 2022) (Self-Supervised Feature Point Detection Algorithm with Dual Attention Network): the paper presents a network that uses a feature pyramid structure for *multi-scale feature fusion*, followed by a position and channel attention module to obtain the feature dependency relationship of the spatial and channel dimensions. The resulting weighted feature maps are added to enhance the feature representation and are trained for detectors and descriptors.

MatchFormer (Wang et al., 2022) (Interleaving Attention in Transformers for Feature Matching): introduces a hierarchical encoder architecture that combines self-attention and cross-attention for improved matching robustness. This approach interweaves self-attention and cross-attention in feature extraction and matching, resulting in an intuitive extract-and-match scheme. The match-aware encoder enhances model efficiency and reduces the decoder workload. Incorporating self- and cross-attention on multi-scale features in a hierarchical structure improves matching robustness, especially in challenging indoor scenes or with limited outdoor training data.

DRC-Net (Li et al., 2020) (Dual-resolution correspondence networks): DRC-Net is a model that extracts both coarse- and fine-resolution feature maps for dense correspondences. It uses a two-step process, generating coarse maps and refining them with a consensus module. The fine-resolution maps guide the final correspondences

Table 1
A comprehensive table containing the most recent works in the field of feature matching.

Paper (Author, Year)	Concept
LightGlue (Lindenberger, Sarlin, & Pollefeys, 2023)	This framework uses introspection to determine whether further computation is necessary in addition to predicting correspondences after each computational block.
DeepMatcher (Xie et al., 2023)	Introduces a Slimming Transformer approach for dense pixel-wise matching. SlimFormer is used to efficiently model relevance among all keypoints and achieve long-term context aggregation.
HTMatch (Cai et al., 2023)	A hybrid transformer-based Graph Neural Network is used for local feature matching. It combines self- and cross-attention to condition feature descriptors in image pairs.
TopicFM (Giang, Song, & Jo, 2022)	Improves matching robustness by encoding high-level contexts in images using topic modeling and probabilistic feature matching.
ASpanFormer (Chen et al., 2022)	A novel attention operation to adjust attention span in a self-adaptive manner. It regresses flow maps to identify the search region, generates an adaptive sampling grid, and computes attention across two images in derived regions.
ClusterGNN (Shi et al., 2022)	Is addressed using an attentional Graph Neural Network (GNN) architecture. It dynamically divides keypoints into distinct subgraphs using a progressive clustering module, reducing unnecessary connections.
DAN-SuperPoint (Li et al., 2022)	The paper presents a network that uses a feature pyramid structure for *multi-scale feature fusion*, followed by a position and channel attention module to obtain the feature dependency relationship of the spatial and channel dimensions.
DenseGAP (Kuang et al., 2022)	Uses a graph structure and anchor points to provide reliable prior information for context modeling.
MatchFormer (Wang et al., 2022)	Incorporates self- and cross-attention on multi-scale features in a hierarchical structure that improves matching robustness, especially in challenging indoor scenes or with limited outdoor training data.
LoFTR (Sun et al., 2021)	To improve cross-view features, a combination of self and cross attention blocks is used. LoFTR replaces global full attention with the Linear Transformer (Katharopoulos, Vyas, Pappas, & Fleuret, 2020) to make computations more manageable.
Patch2Pix (Zhou, Sattler, & Leal-Taixe, 2021)	The proposed architecture extracts features from a correspondence network using an adapted ResNet34 backbone. Patch2Pix then refines the proposals at image resolution by employing two levels of regressors with the same architecture.
DRC-Net (Li, Han, Li, & Prisacariu, 2020)	It extracts both coarse- and fine-resolution feature maps. It works in two steps, first generating coarse maps and then refining them with the agreement module.

based on the refined coarse tensor. By selecting matching scores at the coarse resolution, the model improves reliability and accuracy without the need for expensive computations on the fine-resolution features. DRC-Net enhances matching performance efficiently.

TopicFM (Giang et al., 2022) (Robust and Interpretable Feature Matching with Topic-assisted): this paper proposes a novel image-matching method that improves *matching robustness by encoding high-level contexts* in images through topic modeling and probabilistic feature matching. The method trains latent semantic instances called topics, explicitly modeling an image as a topic distribution, and focuses on semantic areas for improved matching. The architecture finds coarse matches from low-resolution features and refines coordinates at high resolution, providing a promising solution for improving image-matching performance in various computer vision applications.

Patch2Pix (Zhou et al., 2021) (Epipolar-guided pixel-level correspondences): proposes a new approach to estimate correspondences in a detect-to-refine manner, using patch-level match proposals followed by refinement with a novel network called Patch2Pix. The proposed architecture extracts features using an adapted ResNet34 backbone and feed them into a correspondence network to detect match proposals. Patch2Pix then refines the proposals using two levels of regressors with the same architecture to progressively refine the match proposals at image resolution. For each match proposal, the *mid-level regressor* outputs a confidence score and a pixel-level local match, updating the search space accordingly. The *fine-level regressor* outputs the final confidence score and pixel-accurate match.

LightGlue (Lindenberger et al., 2023): by leveraging predictive analysis and introspection, LightGlue introduces a paradigm shift in correspondence computation. This framework not only predicts correspondences after each computational block, but it also uses introspection to determine whether additional computation is required. LightGlue's early elimination of non-matchable points, which directs its attention to the visible area and thus improves accuracy, is a standout feature. LightGlue redefines how correspondences are established in visual data by combining predictive capabilities and selective focus, promising efficient and precise results.

Detector-free local feature matching has demonstrated promising outcomes in diverse computer vision tasks, including image matching, object recognition, and tracking. These methods may not perform optimally in certain scenarios, such as instances involving significant transformations or occlusions. However, the removal of keypoint detection from the local feature-matching process has introduced fresh prospects for the development of computer vision algorithms that are both more efficient and precise.

2.3. Optimized transformers

The combination of transformers and local feature matching is a powerful approach to solving problems in computer vision (Jarvis, 1983), such as image recognition (Dosovitskiy et al., 2020), general part assembly (Li, Zeng, & Song, 2023) and object detection (Carion et al., 2020; Dai, Cai, Lin, & Chen, 2021). Transformers are a custom type of neural network architecture that has been primarily used in natural language processing (Chowdhary & Chowdhary, 2020). They excel at modeling long-range dependencies between input sequences and have shown impressive results in tasks such as language translation, language modeling, and text classification.

The vanilla transformer's memory cost grows quadratically with the sequence length, limiting its efficiency for longer sequences. Recently, various approaches (Katharopoulos et al., 2020) suggested to improve transformers' efficiency, such as *sparse attention*, *hierarchical transformers*, *linear transformers* (Shen, Zhang, Zhao, Yi, & Li, 2021) and those with fixed receptive fields.

3. Materials and methodology

3.1. Preliminaries

In this section, we provide more context and detail on the LoFTR (Sun et al., 2021) and PixLoc (Sarlin et al., 2021) algorithm, as it forms the basis of our method. As such, we describe the LoFTR, PixLoc in detail, and the knowledge distillation mechanism as it is necessary to understand the underlying principles and techniques in order to highlight the novelty of the proposed method.

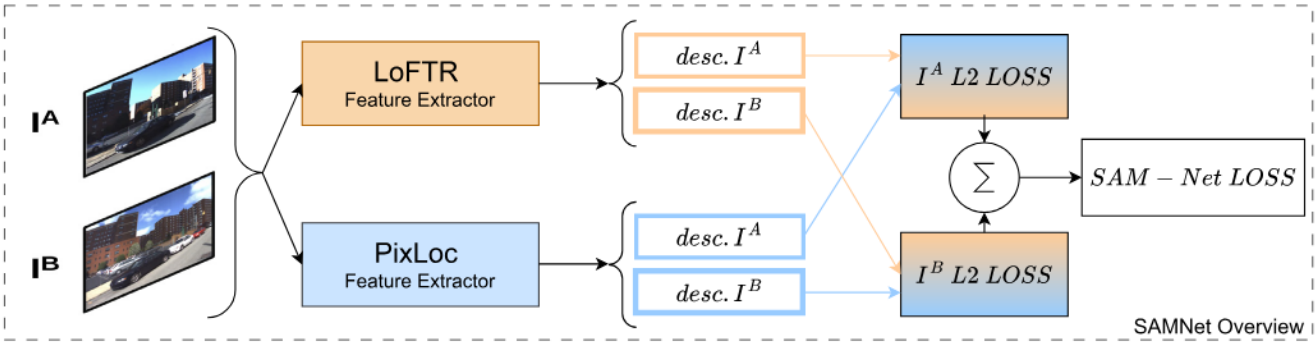

Fig. 2. A high-level overview of our knowledge distillation process with the image I^A and I^B at the input and $SAM - Net\ LOSS$ as output.

3.1.1. LoFTR

The work entitled Detector-Free Local Feature Matching with Transformers (Sun et al., 2021) proposes a novel approach for local feature matching in computer vision that does not require the use of handcrafted detectors. Instead, the authors use a transformer network to extract features from input images and match them across different scales. LoFTR (Sun et al., 2021) employs a combination of self and cross attention blocks to enhance cross-view features. To make computations more manageable, LoFTR replaces global full attention with the Linear Transformer (Tang, Zhang, Zhu, & Tan, 2022). While this approach has demonstrated its effectiveness, there is a valid concern regarding the absence of detailed local interaction among pixel tokens at the fine level. This limitation restricts LoFTR's ability to accurately extract well-defined correspondences. This concern gains further weight due to the observations made by Tang et al. (Rocco, Arandjelović, & Sivic, 2020), who found that the cross attention map generated by LoFTR's Linear Transformer tends to spread across larger areas rather than precisely concentrating on the actual corresponding regions.

To overcome this challenge, we present a Transformer-based detector-free matcher with a hierarchical attention framework to capture both global context and local details. Our foundation processing blocks, known as Global-Local Attention (GLA) blocks, perform coarse-level global attention at low resolution to acquire long-range dependencies, while fine-level local attention at high resolution is performed within only a concentrated region around a current correspondence discovered through dense flow prediction.

3.1.2. Knowledge distillation

Knowledge distillation is shown to have a number of advantages over traditional model training methods (Wang & Yoon, 2021). One of the most significant advantages is that it allows for the creation of smaller, more efficient models that can be deployed in resource-constrained environments. This is particularly important in the field of deep learning, where the size and complexity of models can be a major limiting factor in their applicability.

In our approach, we employ knowledge distillation to mimic the behavior of PixLoc (Sarlin et al., 2021), specifically focusing on ignoring dynamic objects. Knowledge distillation involves training a smaller model, known as the *student model*, to mimic the behavior of a larger and more complex model, known as the *teacher model*.

One challenge in this task is handling dynamic objects, which are objects that change appearance or location over time. These objects can introduce noise or errors in the camera pose estimation process.

To overcome this challenge, we utilize knowledge distillation to train our *student model*. We use the teacher model, which is PixLoc (Sarlin et al., 2021) in our case, as the source of knowledge.

During the training phase, we present the same input data to the teacher and student models. The teacher model produces its predictions, which serve as targets for the student model. The student model then tries to mimic the behavior of the teacher model by producing similar predictions for the given inputs.

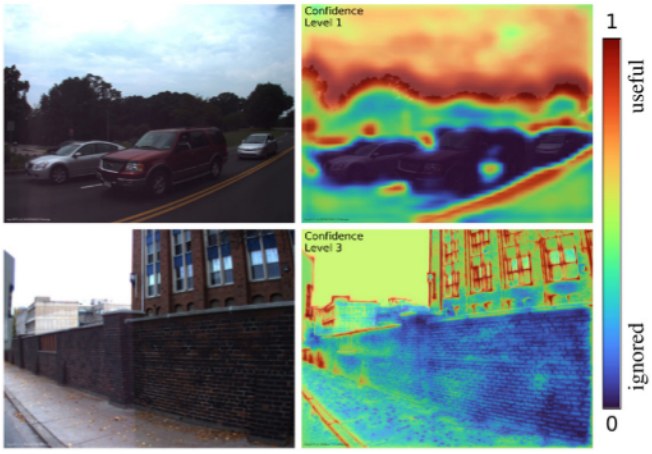

Fig. 3. Visualization of how PixLoc (Sarlin et al., 2021) ignore dynamic objects.

By training the student model in this manner, we aim to transfer the knowledge of handling dynamic objects from the teacher model to the student model. This means that the student model learns to ignore or appropriately account for dynamic objects in the camera pose estimation process.

Overall, our approach leverages knowledge distillation to mimic the behavior of PixLoc, specifically focusing on addressing the challenges posed by dynamic objects in camera pose estimation. This enables us to improve the robustness and accuracy of our model in real-world scenarios where dynamic objects are present. An overview of our knowledge distillation process can be seen in Fig. 2.

3.1.3. PixLoc

PixLoc (Sarlin et al., 2021), is a neural network designed to estimate the precise 6-DoF (Degree of Freedom) pose of an image with respect to a 3D model, regardless of the scene. The approach makes use of the direct alignment of multiscale deep features, treating camera localization as a metric learning task. By training PixLoc end-to-end from pixel-level information to pose estimation, it learns robust data priors. This enables PixLoc to demonstrate exceptional generalization capabilities when applied to new scenes, achieved by separating model parameters and scene geometry. PixLoc effectively ignores dynamic objects such as cars or fallen leaves, as well as repetitive patterns like brick walls. Instead, it focuses on salient features such as road markings, tree silhouettes, and prominent structures on buildings, as can be seen in Fig. 3.

3.2. Proposed model

Our method, called SAM-Net introduces a detector-free design. Instead of relying on a feature detector, SAM-Net directly extracts local

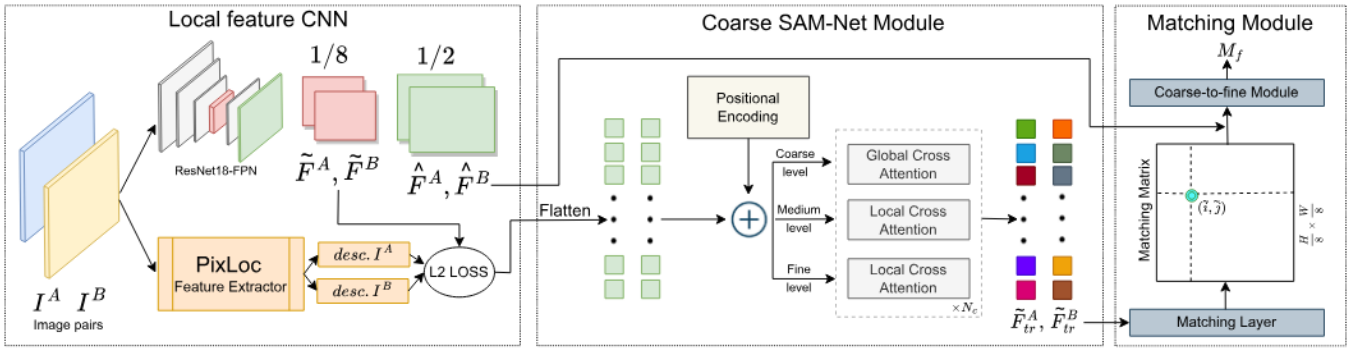

Fig. 4. Our architecture consists of two models, PixLoc (Sarlin et al., 2021) and SAM-Net, which operate concurrently. The PixLoc module provides us with the descriptors ($desc.\,I^A, desc.\,I^B$). The architecture is composed of the following components: **Local Feature CNN:** This component extracts coarse-level feature maps ($\hat{F}_A$ and $\hat{F}_B$) and fine-level feature maps ($\tilde{F}_A$ and $\tilde{F}_B$) from the image pair I^A and I^B. **Flattening and Positional Encoding:** The coarse feature maps are flattened into 1-D vectors and combined with positional encoding. These augmented features are then passed through the Local Feature Transformer (SAM-Net) module, which consists of N_c global and local attention layers. **Matching Layer:** To establish correspondences between the transformed features, a differentiable matching layer is employed. This layer produces a confidence matrix. We select matches based on a confidence threshold and the mutual-nearest-neighbor criteria. **Final Match Prediction:** M_f is obtained by refining the coarse matches within a local window, cropped from the fine-level feature map, for each selected coarse prediction.

features from the images. This design eliminates the repeatability issue associated with feature detectors, where the same interest points may not be consistently detected across different images. Furthermore, by distilling information from PixLoc (Sarlin et al., 2021), it emulates its behavior, namely ignoring dynamic objects. Ignoring dynamic objects in localization can enhance performance by reducing background noise, improving feature extraction, minimizing temporal inconsistencies, enhancing generalization, and enabling more efficient inference. Fig. 4 provides an overview of the SAM-Net method, illustrating the key steps and components involved in extracting robust and repeatable local features without the need for a feature detector.

3.2.1. Student model (Light FPN)

In the SAM-Net framework, the first step involves using a standard CNN with a Feature Pyramid Network (FPN) to extract features from a pair of images, denoted as I^A and I^B. The extracted features are divided into coarse-level and fine-level features.

The coarse-level features, denoted as $\tilde{F}^A$ and $\tilde{F}^B$, are extracted using the CNN with FPN. These features are represented as tensors of shape $\mathbb{R}^{\hat{C}\times H/8\times W/8}$, where $\hat{C}$ is the feature dimension, and H and W are the height and width of the original images divided by 8. This downsampling is performed to reduce the spatial dimensionality of the features.

On the other hand, the fine-level features, denoted as $\hat{F}^A$ and $\hat{F}^B$, are also extracted using the same CNN with FPN. These features are represented as tensors of shape $\mathbb{R}^{\hat{C}\times H/2\times W/2}$, where $\hat{C}$ is the feature dimension, and H and W are the height and width of the original images divided by 2. The fine-level features capture more detailed information compared to the coarse-level features.

Once the local features $\tilde{F}^A$ and $\tilde{F}^B$ are extracted, they undergo the SAM-Net module, which extracts local features dependent on position and context. The main purpose of the SAM-Net module is to convert the features into representations that are more suitable for matching. The transformed features are labeled as $\tilde{F}^A_{tr}$ and $\tilde{F}^B_{tr}$.

3.2.2. Positional encoding

Positional encoding plays a crucial role in preserving spatial information for flattened tokens, as demonstrated in transformer networks. In order to encode position information, 2D sinusoidal signals of various frequencies are employed and added to the initial features, following the same approach used in LoFTR (Sun et al., 2021). When the position encoding is applied to $\tilde{F}^A$ and $\tilde{F}^B$, the transformed features become dependent on their respective positions. This positional dependency is essential for the SAM-Net module to generate accurate matches, especially in regions that lack distinctive features. When the testing resolution differs from the training resolution, we apply normalization to ensure consistency.

3.2.3. PixLoc teacher model

In our knowledge distillation process from PixLoc (Sarlin et al., 2021), we employed the L2 Loss as a method to transfer information. The process involves two main phases: during the initial stage, the distilled knowledge is derived from both the teacher and student models, employing the L2 loss as the distillation criterion to instruct the student model's training. In the second stage, the already pre-trained student model will be trained again with the retained weights. The mathematical representation is presented in the following equation:

$$\mathcal{L}_{distill} = \sum MSE(K_S - K_T) \tag{1}$$

where:

- MSE represents the *mean square error*;
- K_S, K_T represents the knowledge of teachers and students, respectively, during the process of knowledge transfer.

PixLoc relies on a UNet feature extractor that is built upon the VGG19 architecture. It extracts 3 feature maps with different strides (1, 4, and 16) and dimensions (32, 128, and 128). To distill the knowledge from PixLoc, we specifically focused on utilizing the last feature maps. Additionally, we conducted experiments that involved incorporating all three feature maps. For more in-depth information on these experiments, please refer to Section 4 of the study.

3.2.4. Global-local attention block

In our approach, we make use of iterative global-local attention (GLA) blocks with a hierarchical structure. Each GLA block regresses auxiliary flow maps that describe the correspondence between coordinates and uncertainty. Rather than using these flow maps as our correspondence output, we use them to guide local cross-attention. In order to enable fine-grained attention without significant cost, we adopt a local attention mechanism on medium and fine level feature maps. This is particularly advantageous as it allows for an adaptive adjustment of the local attention span based on the uncertainty inherent in the matching process. Some work (Truong, Danelljan, Van Gool, & Timofte, 2021) suggests using a probabilistic model to jointly explain both flow estimations and their confidence as an elegant framework for uncertainty prediction (Tutsoy & Tanrikulu, 2022). Inspired by the above works, incorporating information about uncertainty (Zhou et al., 2020), the model gains the ability to dynamically modify the range of elements it attends to, thereby enhancing its accuracy in correspondence matching.

In the context of the iterative GLA block, we begin by predicting flow maps using an MLP based on input features. Simultaneously, the flow maps at a medium level are acquired through strided average

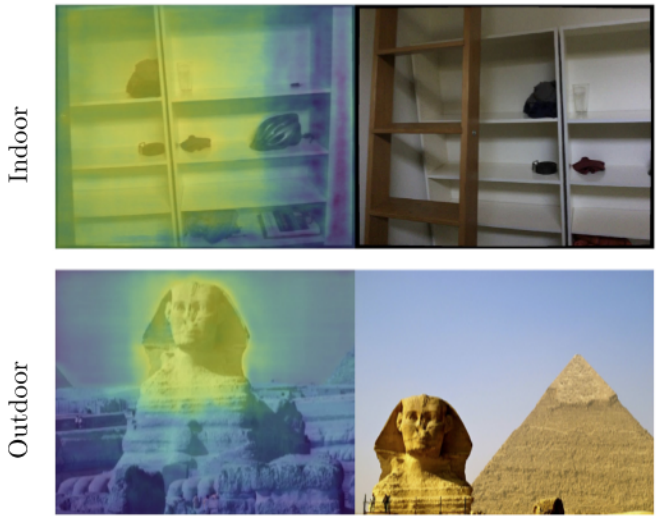

Fig. 5. The uncertainty map, predicted alongside the flow information. With warmer colors indicating lower levels of uncertainty.

pooling. At each scale level, we divide the associated query map Q into cells. Within each cell, we use average flow estimation to create a rectangular area on the K and V maps. Then, attention is applied to each cell and the sampled tokens.

Our method effectively captures both global and local information in the data. Improves the accuracy and robustness of correspondence estimation by adapting the attention span according to the uncertainty of the matches.

3.2.5. Uncertainty estimation

In uncertainty estimation, we categorize all pixels into two groups: matchable pixels and unmatchable pixels. These groups are determined based on ground-truth camera poses and depths. We then calculate the average standard deviation (σ) for each group. Over the course of iterations, the average (σ) for matchable pixels decreases as the network gains more confidence in its flow predictions during later stages. Conversely, the uncertainty values for unmatchable pixels are gradually increased by the network to avoid becoming excessively confident in a particular region.

In Fig. 5, we present a visual representation of the uncertainty map for flow prediction. In the first phase, overlapping and non-overlapping regions are differentiated. It is notable that uncertainty tends to be greater in textureless areas, suggesting that a broader context is necessary for cross-attention in these regions.

3.2.6. Matching layer

We adopt the approach utilized in *LoFTR* (Sun et al., 2021) for generating the final correspondences. This approach consists of two stages: a coarse matching stage and a sub-pixel refinement stage.

After undergoing N iterations of GLA (Global Local Attention) blocks, the updated feature maps are flattened, and a correlation matrix is then constructed. By applying dual-direction softmax along both the column and row dimensions, a score matrix is obtained Coarse-level matches are extracted using the mutual nearest neighbor (MNN) technique and by filtering out scores below a threshold. These coarse matches are subsequently passed through a correlation-based refinement block, which follows the same procedure as in LoFTR (Sun et al., 2021), to obtain the final matching results M_f.

3.2.7. Loss formulation

The final loss is composed of 4 distinct components: the coarse-level loss, the fine-level loss, the flow estimation loss, and the distillation loss. These individual loss terms collectively contribute to guiding the global loss.

$$\mathcal{L} = \mathcal{L}_c + \mathcal{L}_f + \mathcal{L}_{flow} + \mathcal{L}_{distill} \tag{2}$$

Coarse-level loss: To compute the coarse-level loss ($\mathcal{L}_c$), we use ground truth matches (M_{gt}) obtained through reprojection using depth and camera poses from datasets. The dual-softmax score matrix (S) is supervised by applying a cross-entropy loss for any location (i, j) in an image.

$$\mathcal{L}_c = -\frac{1}{\left|M_{gt}\right|} \sum_{(i,j)\in M_{gt}} \log(S(i,j)) \tag{3}$$

Fine-level loss: The fine-level loss is supervised by directly comparing the refined coordinates (M_f) with the ground truth reprojection coordinates using the L_2 distance metric. By individually comparing each coordinate $M_f(i, j)$ with its corresponding ground truth coordinate, the loss considers the Euclidean distance between them, ensuring the model is trained to accurately estimate and refine the coordinates of the correspondences.

$$\mathcal{L}_f = \frac{1}{\left|\mathcal{M}_f\right|} \sum_{(\hat{i},\hat{j}')\in\mathcal{M}_f} \frac{1}{\sigma^2(\hat{i})} \left\| \hat{j}' - \hat{j}'_{gt} \right\|_2 \tag{4}$$

where: $\hat{j}_{gt}$ is computed by warping each $\hat{i}$ from $\hat{F}^A_{tr}(\hat{i})$ to $\hat{F}^B_{tr}(\hat{j})$ with the ground-truth camera pose and depth. The corresponding heatmap is denoted with $\sigma^2(\hat{i})$.

Flow-level loss: In our approach, a Multi Layer Perceptron (MLP) is utilized to predict auxiliary flow maps within each GLA block. To train this MLP, a loss function is employed, which comprises a weighted sum of the L2-distance between the estimated flows and the ground truth flows.

The supervision of flow estimation involves minimizing the log-likelihood for each estimated distribution. Specifically, this entails comparing the flow estimation (θ) obtained from each layer of the model with the corresponding ground truth flow (D^{gt}). This process ensures that the model is trained to accurately estimate the flow maps by aligning them with the ground truth flows.

$$\mathcal{L}_{flow} = -\frac{1}{|D^{gt}|} \sum_{ij} \log\left(\mathcal{N}\left(D^{gt}_{ij} \mid \Phi_{ij}\right)\right) \tag{5}$$

where D^{gt}_{ij} is the ground truth flow, $\Phi_{ij} = \left(u^{ij}_x, u^{ij}_y, \sigma^{ij}_x, \sigma^{ij}_y\right)$ are predicted parameters at location (i, j), $[u_x, u_y] = Sigmoid(f[:2]) * [H, W]$, $[\sigma_x, \sigma_y] = \exp(f[2:])$ and f is 4-dimensional feature.

3.2.8. Implementation details

We implemented our model based on the LoFTR (Sun et al., 2021) framework, using ResNet-18 (Targ et al., 2016) as the initial feature extractor. ResNet-18 outputs feature maps at two resolutions: 1/8 and 1/2. The 1/8 feature map is fed into a transformer-based network for updating, while the 1/2 resolution is used for fine-match coordinates refinement. We adopted a dual-softmax approach for coarse matching, applying a learnable temperature initialized as 10. The transformer network was trained using backpropagation and Adam with a *learning rate* $= 10^{-3}$ and *weight decay* $= 0.1$. In the fine matches coordinates refinement step, we use the 1/2 resolution feature map to improve the accuracy of coordinate estimation. The model's parameters were determined based on experimentation and evaluation of the results.

We trained two models specifically for indoor and outdoor scenes. The indoor model was trained on the ScanNet (Dai et al., 2017) dataset, while the outdoor model was trained on the MegaDepth (Li & Snavely, 2018) dataset. Both models underwent 30 epochs of training on 4 × A100 GPUs (more details about the software and hardware specifications can be found in the Appendix D). The training process involved

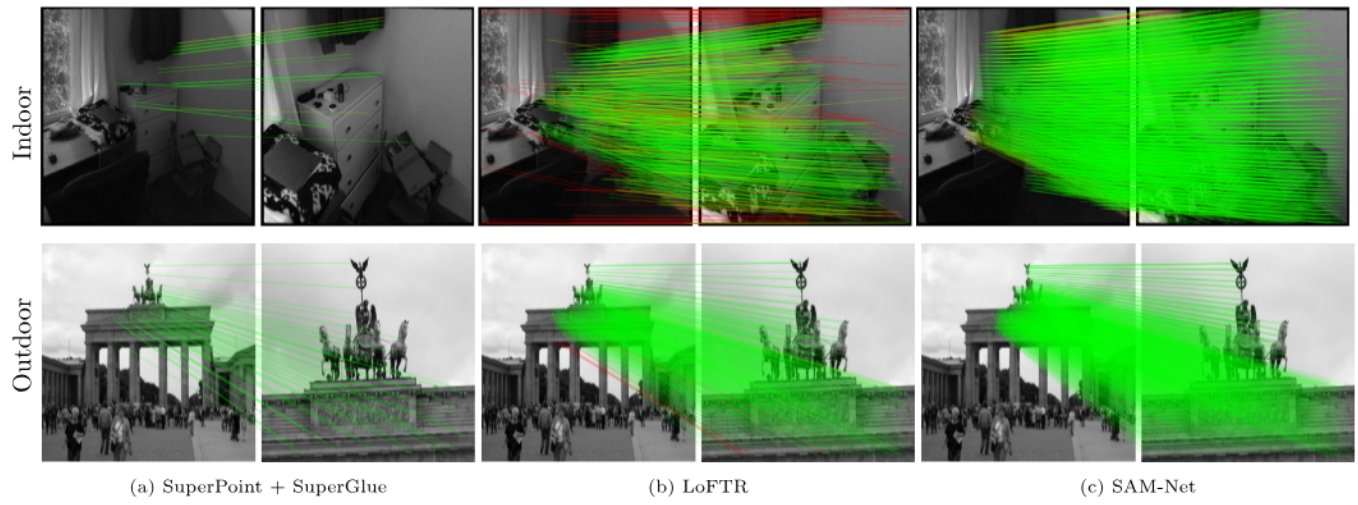

Fig. 6. Qualitative analysis. The results showed that SAM-Net achieved a higher number of correct matches and a lower number of mismatches compared to the other two methods (SuperGlue Sarlin et al., 2020 and LoFTR Sun et al., 2021). This indicates that SAM-Net is more effective in accurately establishing correspondences between images. Comparison between the correct matches/total matches. **Outdoor:** SuperPoint + SuperGlue (41/42), LoFTR (432/438), SAM-Net (797/797). **Indoor:** SuperPoint + SuperGlue (24/26), LoFTR (723/799), SAM-Net (1819/1838).

optimizing model parameters using techniques such as gradient descent and backpropagation. This approach allows the models to capture and understand scene-specific patterns, resulting in improved performance and robustness in scene-understanding tasks for indoor and outdoor environments.

4. Experiments

This section of our work highlights the performance of our proposed method. We evaluate the effectiveness of our method on both indoor and outdoor scenes to ensure its generalizability and robustness in diverse environments. To demonstrate the effectiveness of our method in both indoor and outdoor scenes, we rely on two popular public datasets: ScanNet (Dai et al., 2017) and MegaDepth (Li & Snavely, 2018).

4.1. Pose estimation

4.1.1. Indoor dataset

The ScanNet dataset (Dai et al., 2017) is a benchmark dataset commonly used for evaluating the performance of visual localization algorithms. It consists of 1613 sequences, with each sequence containing RGB images that expose large view changes and repetitive or textureless patterns. Ground-truth depth maps and camera poses are provided for each sequence, making it a valuable resource for developing and testing visual localization approaches.

To ensure fair comparison with other state-of-the-art methods, we follow the same training and testing protocols used by LoFTR (Sun et al., 2021). Specifically, we sample 230M image pairs for training and 1.5K image pairs for testing. The training and testing sets are carefully selected to cover a wide range of scenes and camera poses, providing a comprehensive evaluation of our approach.

In line with LoFTR, we resize all test images to 640 x 480 to ensure consistency in the input size across different methods. This allows for a fair comparison of the performance of different approaches on the same dataset.

4.1.2. Outdoor dataset

MegaDepth (Li & Snavely, 2018) is a challenging dataset that consists of 196 3D reconstructions from 1M internet images. The dataset provides a diverse range of outdoor scenes with varying lighting, weather conditions, and textures, making it an ideal benchmark for evaluating visual localization approaches in outdoor environments.

To evaluate our method's performance on this dataset, we perform two-view pose estimation on 1.5K testing pairs. For each pair, we use our method to estimate the relative camera pose between the two images. The ground-truth camera poses and depth maps are initially computed using COLMAP (Schonberger & Frahm, 2016), and then refined to provide accurate ground-truth data for evaluation. We resize all test images to have a longest dimension of 1152 pixels in order to provide a fair comparison with other state-of-the-art methods.

4.1.3. Evaluation metrics

To train and evaluate our method, we follow the standard protocols used in previous work (Sarlin et al., 2020). We train and evaluate our method separately on the two datasets, ScanNet and MegaDepth. For two-view pose estimation, we recover the essential matrix from correspondences produced by our method. The essential matrix encodes the relative position and orientation between two views and can be used to estimate the camera pose. We then use the estimated camera pose to measure pose accuracy by computing the Area Under the Curve (AUC) at multiple error thresholds (5°, 10° and 20°). To be considered accurate, a pose's angular rotation error and translation error must be less than a certain threshold when compared to ground-truth poses. The angular rotation error measures the difference between the estimated and ground-truth rotation angles, whereas the translation error measures the difference between the estimated and ground-truth translation vectors.

4.1.4. Results

Tables 2 and 3 highlight the performance of our method for both indoor and outdoor scenes. When compared to other methods, it consistently achieves the highest accuracy. Furthermore, Fig. 6 provides a visual representation of our method's superior performance when compared to other matches. This combination of quantitative and qualitative results demonstrates the efficacy of our approach.

4.2. Visual localization

In addition to evaluating our network's performance on the two-view pose estimation task, we extend its capabilities by integrating it into a visual localization pipeline. To validate its effectiveness in handling multi-view matching in different environments, we employ two widely used datasets: InLoc (Taira et al., 2018) for indoor scenes and Aachen Day-Night v1.1 (Zhang, Sattler, & Scaramuzza, 2021) for outdoor scenes. Through these experiments, we demonstrate the robustness and accuracy of the method in achieving reliable visual localization results across diverse indoor and outdoor settings.

Table 2
The performance of two-view pose estimation on outdoor scenes in the MegaDepth dataset (Li & Snavely, 2018) was evaluated. The results indicate the accuracy and effectiveness of the pose estimation algorithm in determining the relative camera positions in outdoor environments.

Local features	Matcher	Pose estimation AUC		
		@5°	@10°	@20°
	Detector-based methods			
SuperPoint	SuperGlue (Sarlin et al., 2020)	42.18	61.16	75.96
	DenseGAP (Kuang et al., 2022)	41.17	56.87	70.22
	ClusterGNN (Shi et al., 2022)	44.19	58.54	70.33
	LightGlue (Lindenberger et al., 2023)	49.9	67.0	80.1
	Detector-free methods			
–	DRC-Net (Li et al., 2020)	27.01	42.96	58.31
	Patch2Pix (Zhou et al., 2021)	41.40	56.32	68.31
	LoFTR (Sun et al., 2021)	52.80	69.19	81.18
	TopicFM (Giang et al., 2022)	54.10	70.10	81.60
	QuadTree (Tang et al., 2022)	54.60	70.50	82.20
	MatchFormer (Wang et al., 2022)	52.91	69.74	82.00
	ASpanFormer (Chen et al., 2022)	55.30	71.50	83.10
	Ours	**55.31**	**71.70**	**83.37**

Table 3
The performance of two-view pose estimation on indoor scenes in the ScanNet dataset (Dai et al., 2017) was evaluated. The results indicate the accuracy and effectiveness of the pose estimation algorithm in determining the relative camera positions in outdoor environments.

Local features	Matcher	Pose estimation AUC		
		@5°	@10°	@20°
	Detector-based methods			
D2-Net	NN	5.25	14.53	27.96
ContextDesc	Ratio test (Lowe, 2004a)	6.64	15.01	25.75
	NN	9.43	21.53	36.40
	NN + OANet (Zhang et al., 2019)	11.76	26.90	43.85
SuperPoint	SuperGlue (Sarlin et al., 2020)	16.16	33.81	51.84
	SGMNet (Chen et al., 2021)	15.40	32.06	48.32
	DenseGAP (Kuang et al., 2022)	17.01	36.07	55.66
	HTMatch (Cai et al., 2023)	15.11	31.42	48.23
	Detector-free methods			
–	LoFTR (Sun et al., 2021)	22.06	40.80	57.62
	QuadTree (Tang et al., 2022)	24.90	44.70	61.80
	MatchFormer (Wang et al., 2022)	24.31	43.90	61.41
	ASpanFormer (Chen et al., 2022)	25.60	46.00	63.30
	Ours	**26.01**	**46.44**	**63.61**

4.2.1. Indoor dataset

We evaluate the performance of our network on the InLoc dataset (Taira et al., 2018), which consists of a comprehensive collection of 9972 RGBD indoor images. These images are precisely aligned to create a reference scene model, while 329 RGB query images are included for visual localization, with ground truth camera poses manually verified. The dataset presents a significant challenge due to the presence of textureless or repetitive patterns and large perspective differences, making accurate matching a difficult task. By testing our network on this dataset, we aim to assess its ability to handle such challenges and provide robust visual localization results in complex indoor environments.

4.2.2. Outdoor dataset

The Aachen Day-Night v1.1 dataset (Zhang et al., 2021) demonstrates a city environment by constructing a reference scene model from 6697 day-time images. The dataset includes 824 additional day-time images and 191 night-time images as query images for visual localization. One of the most difficult challenges in this dataset is accurately identifying correspondences, especially in night-time images with significant and sudden changes in illumination. When dealing with night-time images and the significant variations in lighting conditions they present, the task of matching features and finding similarities becomes especially difficult. To overcome this challenge, robust algorithms that can handle the complexities of matching features under extreme lighting changes in night-time scenes are required.

Table 4
Visual localization evaluation on the InLoc (Taira et al., 2018) benchmark.

Method	DUC1	DUC2
	(0.25 m, 10°) / (0.5 m, 10°) / (1.0 m, 10°)	
KAPTURE + R2D2	41.4/60.1/73.7	47.3/67.2/73.3
KAPTURE + R2D2	41.4/60.1/73.7	47.3/67.2/73.3
LoFTR	47.5/72.2/84.8	54.2/74.8/**85.5**
SP + LighGlue	49.0/68.2/79.3	55.0/74.8/79.4
SP + SuperGlue	49.0/68.7/80.8	53.4/**77.1**/82.4
ASpanFormer	51.5/73.7/86.4	55.0/75.7/82.5
Ours	**51.8/73.9/87.8**	**56.0**/75.8/83.1

Table 5
Visual localization results on Aachen V1.1 (Zhang et al., 2021) dataset.

Method	Day	Night
	(0.25 m, 10°) / (0.5 m, 10°) / (1.0 m, 10°)	
Localization with matching pairs provided in dataset		
R2D2 + NN	–	71.2/86.9/98.9
ASLFeat + NN	–	72.3/86.4/97.9
SP + SuperGlue	–	73.3/88.0/98.4
SP + SGMNet	–	72.3/85.3/97.9
Localization with matching pairs generated by HLoc		
LoFTR	88.7/95.6/99.0	78.5/90.6/99.0
SP + SuperGlue	89.8/**96.1**/**99.4**	77.0/90.6/**100**
LightGlue	89.2/95.4/98.5	**87.8**/**93.9**/**100**
ASpanFormer	89.4/95.6/99.0	77.5/91.6/99.5
LightGlue	**90.2**/96.0/99.4	77.0/91.1/**100**
Ours	89.7/95.8/99.0	78.6/91.8/**100**

4.2.3. Evaluation metrics

To compute query poses, we follow the Long-Term Visual Localization Benchmark (Toft et al., 2020) guidelines. To find potential candidate pairs in both datasets, we use the pre-trained HLoc (Toft et al., 2020). To recover camera poses, we use the trained model on the MegaDepth dataset using SuperGlue (Sarlin et al., 2020) and LoFTR (Sun et al., 2021).

4.2.4. Results

According to the findings presented in Table 4, our methods outperform multiple comparative methods on the InLoc dataset (Taira et al., 2018). The results show that our approach achieves the best overall results for DUC1 when compared to the other methods. Similarly, on the Aachen V1.1 dataset, as shown in Table 5, our method outperforms all other approaches at night, with the exception of *SuperGlue*. These findings highlight our method's effectiveness and competitiveness on both datasets, and confirm our method's potential for a variety of applications in the field.

4.3. Ablation study

In order to assess the effectiveness of different design components in our method, we carried out ablation experiments. These experiments specifically involved comparing the timing of SAM-Net method with LoFTR and evaluating the impact of incorporating all three feature maps from PixLoc (Sarlin et al., 2021). By conducting these experiments, we aimed to determine the relative performance and efficiency of our approach compared to LoFTR.

We conducted ablation experiments to evaluate the effectiveness of various design components in our method. These experiments compared the timing of the SAM-Net method with LoFTR and evaluated the impact of incorporating all three feature maps from PixLoc (Sarlin et al., 2021). The goal of these experiments was to determine the relative

Table 6
Comparison of runtime speed on 640 × 480 images.

Stage	Runtime (ms)	
	LoFTR	Ours
Local feature CNN	28.85	28.85
Attention module	23.10	46.99
Matching	37.20	37.12
Total	89.15	112.96

Table 7
The evaluation of two-view pose estimation on outdoor scenes in the MegaDepth dataset (Li & Snavely, 2018). We compared our method, by using the last feature map from PixLoc (Sarlin et al., 2021), and the use of all 3 feature maps.

Feature maps	Pose estimation AUC		
	@5°	@10°	@20°
SAM-Net with 3 feature maps from PixLoc	25.87	46.43	63.57
SAM-Net with last feature map from PixLoc	**26.01**	**46.44**	**63.61**

performance and efficiency of our approach compared to LoFTR (Sun et al., 2021).

4.3.1. Timing

We evaluate the speed of SAM-Net by measuring its performance on 100 randomly selected ScanNet image pairs, each with a resolution of 640 × 480. The timing is conducted using 4 × A100 GPU's, and we present the averaged results. The time required for processing an image pair of 640 × 480 on SAM-Net is 112.96 ms, whereas on LoFTR (Sun et al., 2021), it is 89.15 ms.

As presented in Table 6, our proposed method has marginally slower performance compared to LoFTR (Sun et al., 2021). This variance in speed can be attributed to the more complex attention operation in our approach. Despite this slight difference in execution speed, our method offers valuable advantages in various aspects.

4.3.2. Feature maps

We explored the effects of utilizing different feature maps and variations in the loss function. We evaluated the knowledge transfer by focusing on the last feature maps extracted from the UNet feature extractor, which is based on the VGG19 architecture used in PixLoc.

Furthermore, we conducted experiments incorporating all three feature maps (strides 1, 4, and 16 with dimensions 32, 128, and 128, respectively). This allowed us to examine the influence of incorporating information from multiple scales. The results are presented in Table 7.

4.3.3. Limitations

Table 5 illustrates that our method outperforms all alternative methods except SuperGlue (Sarlin et al., 2020). This achievement is attributed, in part, to our use of coarse matches only for database reconstruction. To elaborate further, when dealing with the Aachen Day-Night (Zhang et al., 2021) dataset, in our approach, we first triangulate reference models using coarse matches between images. Afterward, we establish fine-level matches between the query images and the database images, with the database images considered as left images. However, this strategy introduces a localization error that has a negative impact on the accuracy of position estimation.

5. Conclusions

This paper describes a detector-free matching method based on transformers and knowledge distillation. The proposed SAM-Net module transforms local features to be context- and position-dependent using cross-attention layers in Transformers, which is critical for obtaining high-quality matches. Furthermore, to ignore noise introduced by dynamic objects, we incorporated PixLoc behavior using knowledge distillation. In localization, ignoring dynamic objects reduces noise and interference caused by their presence, allowing the model to focus on relevant static objects for accurate localization. The effectiveness of our method is validated by state of the art results.

6. Future work

A possible future research direction is to explore the applicability of knowledge distillation and transformer-based models to a broader array of computer vision tasks, extending beyond object detection, semantic segmentation, and image recognition. Furthermore, we place significant emphasis on enhancing the resilience of the proposed approach against adversarial attacks and refining its real-time implementation for visual localization.

CRediT authorship contribution statement

Benjamin Kelenyi: Conceptualization, Methodology, Software, Writing. **Victor Domsa:** Data curation, Visualization, Investigation. **Levente Tamas:** Supervision, Reviewing and editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Levente Tamas reports financial support was provided by Executive Unit for Financing Higher Education Research Development and Innovation.

Data availability

Data will be made available on request.

Acknowledgments

The authors thank Analog Devices GMBH Romania for the equipment list, and NVidia for the Deep learning GPU Accelerator (DGX) grade server offered as support for this work. This work was financially supported by the Romanian National Authority for Scientific Research, project nr. PN-III-P2-2.1-PED-2021-3120 and by the Project "Network of excellence in applied research and innovation for doctoral and post-doctoral programs/InoHubDoc", project co-funded by the European Social Fund financing agreement no. POCU/993/6/13/ 153437 and by project SeaClear2.0, which received funding from the European Union's Horizon Europe innovation programme under grant agreement No 101093822 and project nr. 38 PFE in the frame of the programme PDI-PFE-CDI 2021.

Appendix A. Abbreviations

AUC	Area Under the Curve
CNN	Convolution Neural Network
DGX	Deep learning GPU Accelerator
FPN	Feature Pyramid Network
GLA	Global Local Attention
GPU	Graphics Processing Unit
MLP	Multi Layer Perceptron
GNN	Graph Neural Network

Appendix B. Variables

Symbol	Description
$\hat{C}$	Image feature dimension
$\tilde{F}^A$, $\tilde{F}^B$	Coarse-level features
$\hat{F}^A$, $\hat{F}^B$	Fine-level features
$\tilde{F}^A_{tr}$, $\tilde{F}^B_{tr}$	Transformed features
H, W	Image height, width
i, j	Location in an image
I^A, I^B	Input image pairs
K_S, K_T	Knowledge of teachers and students
$\mathcal{L}_c$	Coarse level loss
$\mathcal{L}_f$	Fine level loss
$\mathcal{L}_{flow}$	Flow level loss
$\mathcal{L}$	Set of knowledge pairs between the teacher and student model
M_f	Final matching results
M_{gt}	Ground truth matches
$\mathcal{N}$	Gaussian distribution
Q, K, V	Query, key, values for self-attention
S	Dual-softmax score matrix
σ	Average standard deviation
θ	Flow estimation

Appendix C. Adjustable control parameters

Parameter	Value
Learnable temperature	10
Learning rate	10^{-3}
Nr. of epochs	30
Weight decay	0.1

All the additional model parameters can be located in our Git[1] repository, specifically in the "scr/config" folder.

Appendix D. Software and hardware specifications

D.1. Software specifications

The software specifications, including libraries, tools, and dependencies, are available on our Git[2] in the *"environment.yaml"* and *"requirements.txt"* files.

D.2. Hardware specifications

The training and testing process was done using the following hardware configurations: Intel(R) Xeon(R) Gold 6226R CPU @ 2.90 GHz, 768 GB System Memory, 4 × A100 Graphics Processing Unit (GPU). More details can be found in our Git[1] in the *"specs.txt"* file.

References

Alahi, A., Ortiz, R., & Vandergheynst, P. (2012). FREAK: Fast retina keypoint. In *2012 IEEE conference on computer vision and pattern recognition* (pp. 510–517). IEEE.

Bay, H., Tuytelaars, T., & Van Gool, L. (2006). SURF: Speeded up robust features. *Lecture notes in Computer Science, 3951*, 404–417.

Blaga, A., Militaru, C., Mezei, A.-D., & Tamas, L. (2021). Augmented reality integration into mes for connected workers. *Robotics and Computer-Integrated Manufacturing, 68*, Article 102057.

Cai, Y., Li, L., Wang, D., Li, X., & Liu, X. (2023). HTMatch: An efficient hybrid transformer based graph neural network for local feature matching. *Signal Processing, 204*, Article 108859.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In *Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, part I 16* (pp. 213–229). Springer.

Chen, H., Luo, Z., Zhang, J., Zhou, L., Bai, X., Hu, Z., et al. (2021). Learning to match features with seeded graph matching network. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 6301–6310).

Chen, H., Luo, Z., Zhou, L., Tian, Y., Zhen, M., Fang, T., et al. (2022). ASpanFormer: Detector-free image matching with adaptive span transformer. In *Computer vision–ECCV 2022: 17th European conference, Tel Aviv, Israel, October 23–27, 2022, proceedings, part XXXII* (pp. 20–36). Springer.

Chowdhary, K., & Chowdhary, K. (2020). Natural language processing. *Fundamentals of Artificial Intelligence*, 603–649.

Dai, Z., Cai, B., Lin, Y., & Chen, J. (2021). Up-DERT: Unsupervised pre-training for object detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1601–1610).

Dai, A., Chang, A. X., Savva, M., Halber, M., Funkhouser, T., & Nießner, M. (2017). Scannet: Richly-annotated 3D reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5828–5839).

DeTone, D., Malisiewicz, T., & Rabinovich, A. (2018). Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 224–236).

Domsa, V., Konievic, R., Kelenyi, B., & Tamas, L. (2023). Local image feature extraction in the context of automated valet parking based on simultaneous localization and mapping. In *2023 European control conference (ECC)* (pp. 1–6). IEEE.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An image is worth 16 × 16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Dusmanu, M., Rocco, I., Pajdla, T., Pollefeys, M., Sivic, J., Torii, A., et al. (2019). D2-net: A trainable cnn for joint description and detection of local features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8092–8101).

Farhat, M., Chaabouni-Chouayakh, H., & Ben-Hamadou, A. (2023). Self-supervised endoscopic image key-points matching. *Expert Systems with Applications, 213*, Article 118696.

Frohlich, R., Tamas, L., & Kato, Z. (2019). Absolute pose estimation of central cameras using planar regions. *IEEE Transactions on Pattern analysis and Machine Intelligence, 43*, 377–391.

Giang, K. T., Song, S., & Jo, S. (2022). TopicFM: Robust and interpretable feature matching with topic-assisted. arXiv preprint arXiv:2207.00328.

Gou, J., Yu, B., Maybank, S. J., & Tao, D. (2021). Knowledge distillation: A survey. *International Journal of Computer Vision, 129*, 1789–1819.

Jarvis, R. A. (1983). A perspective on range finding techniques for computer vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-5*(2), 122–139.

Kang, Z., Yang, J., Yang, Z., & Cheng, S. (2020). A review of techniques for 3D reconstruction of indoor environments. *ISPRS International Journal of Geo-Information, 9*(5), 330.

Katharopoulos, A., Vyas, A., Pappas, N., & Fleuret, F. (2020). Transformers are RNNS: Fast autoregressive transformers with linear attention. In *International conference on machine learning* (pp. 5156–5165). PMLR.

Kelenyi, B., & Tamas, L. (2023). D3GATTEN: Dense 3D geometric features extraction and pose estimation using self-attention. *IEEE Access, 11*, 7947–7958.

Kolodiazhnyi, K. (2022). Local feature matching with transformers for low-end devices. arXiv preprint arXiv:2202.00770.

Kuang, Z., Li, J., He, M., Wang, T., & Zhao, Y. (2022). DenseGAP: graph-structured dense correspondence learning with anchor points. In *2022 26th international conference on pattern recognition (ICPR)* (pp. 542–549). IEEE.

Leutenegger, S., Chli, M., & Siegwart, R. Y. (2011). BRISK: Binary robust invariant scalable keypoints. In *2011 international conference on computer vision* (pp. 2548–2555). IEEE.

Li, Z., Cao, J., Hao, Q., Zhao, X., Ning, Y., & Li, D. (2022). DAN-SuperPoint: Self-supervised feature point detection algorithm with dual attention network. *Sensors, 22*(5), 1940.

Li, X., Han, K., Li, S., & Prisacariu, V. (2020). Dual-resolution correspondence networks. *Advances in Neural Information Processing Systems, 33*, 17346–17357.

Li, Z., & Snavely, N. (2018). Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2041–2050).

Li, Y., Zeng, A., & Song, S. (2023). General part assembly planning. arXiv preprint arXiv:2307.00206.

Lindenberger, P., Sarlin, P.-E., & Pollefeys, M. (2023). Lightglue: Local feature matching at light speed. arXiv preprint arXiv:2306.13643.

Lowe, D. G. (2004a). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision, 60*, 91–110.

Lowe, G. (2004b). SIFT-the scale invariant feature transform. *International Journal, 2*(91–110), 2.

Molnár, S., & Tamás, L. (2023). Representation learning for point clouds with variational autoencoders. In *European conference on computer vision* (pp. 727–737). Cham: Springer.

Ono, Y., Trulls, E., Fua, P., & Yi, K. M. (2018). LF-net: Learning local features from images. *Advances in Neural Information Processing Systems, 31*.

[2] https://benjaminkelenyi.github.io/samnet/.

Pop, A., & Tamas, L. (2022). Next best view estimation for volumetric information gain. *55*, (pp. 160–165).

Revaud, J., De Souza, C., Humenberger, M., & Weinzaepfel, P. (2019). R2d2: Reliable and repeatable detector and descriptor. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems, Vol. 32* (p. 11). Curran Associates, Inc..

Rocco, I., Arandjelović, R., & Sivic, J. (2020). Efficient neighbourhood consensus networks via submanifold sparse convolutions. In *Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, part IX 16* (pp. 605–621). Springer.

Rublee, E., Rabaud, V., Konolige, K., & Bradski, G. (2011). ORB: An efficient alternative to SIFT or SURF. In *2011 international conference on computer vision* (pp. 2564–2571). IEEE.

Sarlin, P.-E., DeTone, D., Malisiewicz, T., & Rabinovich, A. (2020). Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4938–4947).

Sarlin, P.-E., Unagar, A., Larsson, M., Germain, H., Toft, C., Larsson, V., et al. (2021). Back to the feature: Learning robust camera localization from pixels to pose. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3247–3257).

Schonberger, J. L., & Frahm, J.-M. (2016). Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4104–4113).

Sengupta, A., Ye, Y., Wang, R., Liu, C., & Roy, K. (2019). Going deeper in spiking neural networks: VGG and residual architectures. *Frontiers in Neuroscience*, *13*, 95.

Shen, Z., Zhang, M., Zhao, H., Yi, S., & Li, H. (2021). Efficient attention: Attention with linear complexities. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 3531–3539).

Shi, Y., Cai, J.-X., Shavit, Y., Mu, T.-J., Feng, W., & Zhang, K. (2022). Clustergnn: Cluster-based coarse-to-fine graph neural network for efficient feature matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12517–12526).

Sun, J., Shen, Z., Wang, Y., Bao, H., & Zhou, X. (2021). LoFTR: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8922–8931).

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1–9).

Taira, H., Okutomi, M., Sattler, T., Cimpoi, M., Pollefeys, M., Sivic, J., et al. (2018). InLoc: Indoor visual localization with dense matching and view synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7199–7209).

Tang, S., Zhang, J., Zhu, S., & Tan, P. (2022). Quadtree attention for vision transformers. In *International conference on learning representations*.

Targ, S., Almeida, D., & Lyman, K. (2016). Resnet in resnet: Generalizing residual architectures. arXiv preprint arXiv:1603.08029.

Toft, C., Maddern, W., Torii, A., Hammarstrand, L., Stenborg, E., Safari, D., et al. (2020). Long-term visual localization revisited. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(4), 2074–2088.

Tola, E., Lepetit, V., & Fua, P. (2010). DAISY: An efficient dense descriptor applied to wide-baseline stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(5), 815–830.

Truong, P., Danelljan, M., Van Gool, L., & Timofte, R. (2021). Learning accurate dense correspondences and when to trust them. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5714–5724).

Tutsoy, O., & Tanrikulu, M. Y. (2022). Priority and age specific vaccination algorithm for the pandemic diseases: a comprehensive parametric prediction model. *BMC Medical Informatics and Decision Making*, *22*(1), 4.

Wang, L., & Yoon, K.-J. (2021). Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Wang, Q., Zhang, J., Yang, K., Peng, K., & Stiefelhagen, R. (2022). Matchformer: Interleaving attention in transformers for feature matching. In *Proceedings of the Asian conference on computer vision* (pp. 2746–2762).

Xie, T., Dai, K., Wang, K., Li, R., & Zhao, L. (2023). DeepMatcher: A deep transformer-based network for robust and accurate local feature matching. arXiv preprint arXiv:2301.02993.

Yi, K. M., Trulls, E., Lepetit, V., & Fua, P. (2016). LIFT: Learned invariant feature transform. In *Computer vision ECCV 2016: 14th European conference, Amsterdam, the Netherlands, October 11-14, 2016, proceedings, part VI 14* (pp. 467–483). Springer.

Zhang, Z., Sattler, T., & Scaramuzza, D. (2021). Reference pose generation for long-term visual localization via learned features and view synthesis. *International Journal of Computer Vision*, *129*, 821–844.

Zhang, J., Sun, D., Luo, Z., Yao, A., Zhou, L., Shen, T., et al. (2019). Learning two-view correspondences and geometry using order-aware network. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 5845–5854).

Zhou, L., Luo, Z., Shen, T., Zhang, J., Zhen, M., Yao, Y., et al. (2020). Kfnet: Learning temporal camera relocalization using kalman filtering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4919–4928).

Zhou, Q., Sattler, T., & Leal-Taixe, L. (2021). Patch2pix: Epipolar-guided pixel-level correspondences. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4669–4678).