Scale-aware Co-visible Region Detection for Image Matching

Xu Pan [a], Zimin Xia [b] and Xianwei Zheng [a,*]

[a]The State Key Lab. LIESMARS, Wuhan University, Wuhan, P.R. China
[b]The VITA Lab. École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

ARTICLE INFO

Keywords:
Image Matching
Scale Variation
Co-visible Region Detection
Structure from Motion
Correspondence Estimation

ABSTRACT

Matching images with significant scale differences remains a persistent challenge in photogrammetry and remote sensing. The scale discrepancy often degrades appearance consistency and introduces uncertainty in keypoint localization. While existing methods address scale variation through scale pyramids or scale-aware training, matching under significant scale differences remains an open challenge. To overcome this, we address the scale difference issue by detecting co-visible regions between image pairs and propose **SCoDe** (**S**cale-aware **Co**-visible region **De**tector), which both identifies co-visible regions and aligns their scales for highly robust, hierarchical point correspondence matching. Specifically, SCoDe employs a novel Scale Head Attention mechanism to map and correlate features across multiple scale subspaces, and uses a learnable query to aggregate scale-aware information of both images for co-visible region detection. In this way, correspondences can be established in a coarse-to-fine hierarchy, thereby mitigating semantic and localization uncertainties. Extensive experiments on three challenging datasets demonstrate that SCoDe outperforms state-of-the-art methods, improving the precision of a modern local feature matcher by 8.41%. Notably, SCoDe shows a clear advantage when handling images with drastic scale variations. Code will be made publicly available at github.com/Geo-Tell/SCoDe.

1. Introduction

Establishing reliable point correspondences between image pairs is a fundamental task in photogrammetry remote sensing, and computer vision, as it supports a wide range of applications, including visual localization (Lindenberger et al., 2023; Sun et al., 2021; Tang et al., 2019), Structure from Motion (Tu et al., 2023; Chen et al., 2020; Shen, 2013), and image registration (Zhang et al., 2023; Wang et al., 2023a; Ye et al., 2022; Li et al., 2021b). It aims to find 2D points in an image pair that are projected from the same 3D point in the scene. Typically, those 2D point matches are identified by matching local features extracted from the image pair. In practical scenarios such as oblique 3D reconstruction and aerial-ground joint reconstruction, images are captured from significantly different viewpoints, often using heterogeneous cameras. This results in substantial scale variations and perspective distortions, creating ambiguity in the semantics and spatial positioning of homologous points. The matching process thus faces considerable uncertainty, especially in the presence of repetitive structures, occlusions, and other challenges common in photogrammetry and remote sensing.

Given these challenges, extensive research has been conducted to improve correspondence estimation. Traditional image matching approaches often rely on handcrafted local feature descriptors for image matching (Rublee et al., 2011; Bay et al., 2006; Rosten and Drummond, 2006; Harris et al., 1988), with some specifically addressing scale variations (Lowe, 2004, 1999; Bay et al., 2006; Rublee et al., 2011). However, these methods typically rely on low-level features and address scale variations using scale pyramids (Rublee et al., 2011), where the image is processed at multiple resolutions to enable approximate scale invariance during keypoint detection and description. Despite such mechanisms, they still exhibit limited capability in capturing global context and establishing long-range associations, making them susceptible to large scale differences, especially when combined with lighting changes and motion blur. Recently, learning-based methods (Lindenberger et al., 2023; Sun et al., 2023; Jiang et al., 2021; Sarlin et al., 2020; Dosovitskiy et al., 2021; Wu et al., 2020; Peyré et al., 2019; Vaswani et al., 2017) have leveraged deep neural networks to extract discriminative features from images, thereby expanding the receptive field and enabling the capture of global context. These deep learning approaches enhance correspondence reliability and accuracy by increasing the density of valid matching point pairs while mitigating false matches through the constraints of global image context.

Corresponding author: Xianwei Zheng
panxurs@whu.edu.cn (Xu Pan); zimin.xia@epfl.ch (Zimin Xia); zhengxw@whu.edu.cn (Xianwei Zheng)
ORCID(s): 0009-0007-3297-0385 (Xu Pan); 0000-0002-4981-9514 (Zimin Xia); 0000-0001-9783-3030 (Xianwei Zheng)

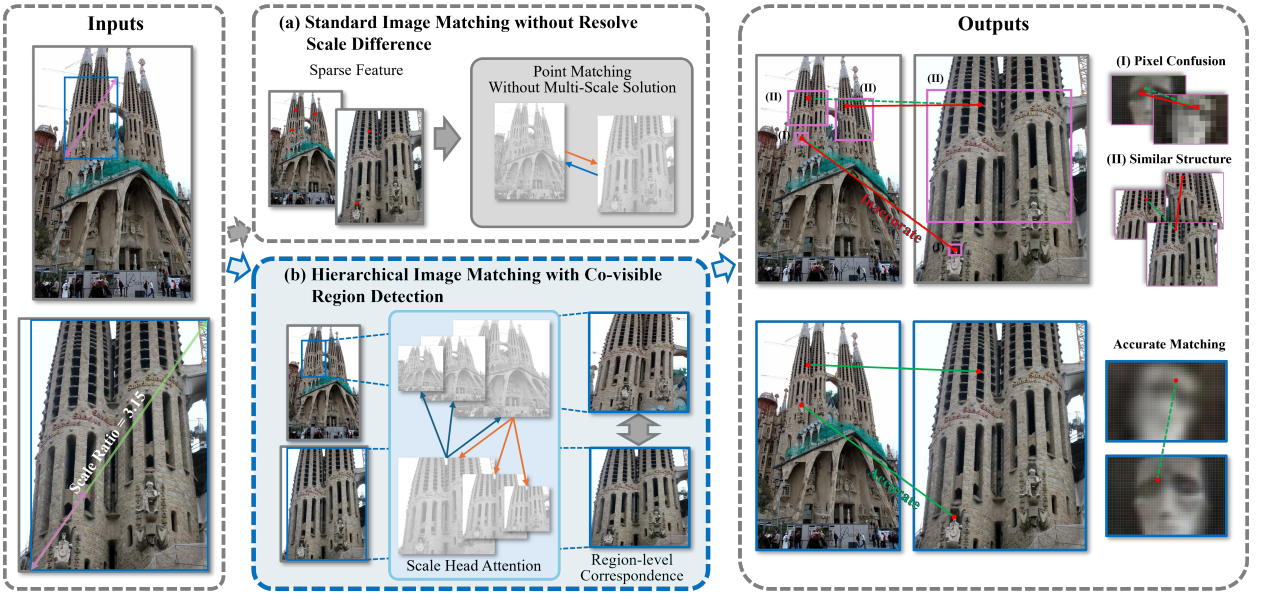


Figure 1: Comparison between direct point matching and hierarchical image matching under drastic scale variation. We define the scale ratio as the ratio of the diagonal lengths of the co-visible regions in image pairs. (a) Traditional methods directly establish point-level correspondences based on feature points. (b) Our method introduces intermediate region-level correspondence through co-visible region detection to form a hierarchical matching strategy.

Nevertheless, both traditional and learning-based methods rely mostly on point-level matching, which suffers from local ambiguity caused by scale differences and texture repetition. In complex scenes with significant variations in scale, visually similar but semantically different points may lead to erroneous matches. As shown in Figure 1(a), point-level matching fails to address the uncertainty caused by: (I) Variations in projection distances and angles. For instance, the number of pixels corresponding to the same points on the statue's head differs between the two images. (II) Repetitive patterns. The two spires are nearly identical in their shapes, and thus create confusion over which window corresponds to which. Due to both factors, point-level matching methods often result in wrong matches in complex scenes, as seen in the outputs in the top right of Figure 1.

To address such uncertainty, we propose a hierarchical matching strategy that first explicitly detects co-visible regions between input images to align their scale space, and then constrains point-level matching within these regions. As shown in Figure 1(b), we first detect the co-visible spire from the image pair to establish region-level correspondences and then rescale the detected windows. This alignment reduces the repetitive patterns and ensures that the 3D points, e.g., the ones on the statue's head, correspond to roughly the same amount of pixels across images, as shown by the outputs in the bottom right of Figure 1. Given the scale-aligned detected windows, precise point correspondences can be more easily matched. This hierarchical pipeline provides a coarse-to-fine matching strategy, where high-level region alignment facilitates robust fine-grained point matching under extreme scale variation.

In this work, we focus on a previously less-considered task, namely detecting co-visible regions between images. We propose a learning-based co-visible region detection method that significantly improves downstream point matching in a hierarchical framework, particularly when there are drastic scale differences between input images. Drawing inspiration from recent object detection research (Carion et al., 2020), we address co-visible region detection using a learnable query that gathers information from both input images. To mitigate scale differences between input images, we propose a Scale Head Attention mechanism that enables interaction between features across different scales. By addressing scale discrepancies in the input, our method demonstrates consistent improvements when used in combination with various downstream image matching methods. Our main contributions are:

- We propose an end-to-end co-visible region detection framework (SCoDe), which can be architecture-agnostic and can be seamlessly integrated into diverse matching backbones without altering their internal designs. SCoDe yields up to an 8.41% increase in point-matching precision on challenging large-scale benchmarks.

- A novel Scale Head Attention (SHA) projects input features to multiple scales and compares them across scales. Our ablation study confirms the effectiveness of the proposed SHA in enhancing co-visible region detection between images with significant scale differences for hierarchical image matching.
- Previous evaluations for co-visible region detection have not considered cases where the input image pair lacks co-visible regions. We introduce an additional similarity metric to assess the validity of the predicted co-visible regions.

2. Related Works

In this section, we first provide an overview of image matching and co-visible region detection from both communities, highlighting the challenges posed by scale variations. We then explain why the attention mechanisms used in existing methods fail to address these challenges.

2.1. Image Matching

Image matching aims to establish accurate correspondences between keypoints across image pairs. Traditional methods rely on keypoint detection and feature extraction using local descriptors such as SIFT (Lowe, 1999), SURF (Bay et al., 2006), and ORB (Rublee et al., 2011). These methods detect distinctive keypoints in both images and compute local descriptors based on the appearance of each keypoint. Correspondences are then established by matching keypoints with similar descriptors. However, these approaches often struggle in the presence of large variations in scale, viewpoint, or illumination, particularly in complex real-world scenes.

To filter out unreliable matches, various techniques such as neighborhood consensus (Cech et al., 2010; Yi et al., 2018; Zhang et al., 2019) are employed. Advanced methods like MAGSAC++ (Baráth et al., 2020) enhance sampling efficiency in high-outlier situations. Additionally, approaches such as Vector Field Consensus (VFC) (Torres et al., 2013) further refine outlier rejection by leveraging global consistency and probabilistic sampling. These traditional methods help mitigate false matches through robust estimation techniques such as epipolar geometry verification using RANSAC or its variants (Fischler and Bolles, 1981; Baráth et al., 2020; Li et al., 2021a). However, their effectiveness can be limited when the initial matches are heavily affected by large geometric distortion or occlusions, which reduce the inlier ratio and pose challenges for subsequent model fitting. Traditional methods rely on hand-crafted features that struggle to generalize well to challenging scenarios involving significant geometric changes or repetitive textures.

With the rise of deep learning, several learning-based methods (Jin et al., 2021; Xu et al., 2024; Zhang et al., 2025) have been developed to address the limitations of traditional techniques. For instance, SuperGlue (Sarlin et al., 2020) uses transformers (Dosovitskiy et al., 2021) to model geometric structures and apply graph neural networks (Wu et al., 2020) together with optimal transport algorithms (Peyré et al., 2019). By incorporating attention mechanisms (Vaswani et al., 2017), SuperGlue can capture long-range dependencies and improve robustness in scenarios with complex geometry. Another example, COTR (Jiang et al., 2021), combines convolutional neural networks (CNNs) with transformers to extract semantic features that enable more accurate image matching. D2-Net (Dusmanu et al., 2019) integrates feature detection and description into a single CNN, enhancing performance in challenging conditions. Moreover, Sparse-NCNet (Rocco et al., 2020) leverages 4D correlation volumes and neighborhood consensus networks to efficiently handle sparse matching tasks.

In photogrammetry, learned features such as SuperPoint have shown strong performance on large-scale aerial datasets (Song et al., 2024; Ioli et al., 2024), suggesting their potential to enhance traditional matching pipelines. However, adapting these methods to multi-view geometry remains challenging due to varying imaging conditions and geometric constraints. Both traditional and learning-based methods encounter difficulties in scenes exhibiting large-scale differences or complex backgrounds, where point-level correspondences are often ambiguous due to repetitive patterns or occlusion. This limitation motivates the need for region-level correspondences, which are more robust to these issues and can provide a higher-level understanding of the scene.

2.2. Co-visible Region Detection

Co-visible region detection aims to identify regions within two images that correspond to the same physical part of the scene. This task is essential when the images are captured from very different viewpoints and have significantly different scales. Traditional approaches to co-visible region detection often rely on keypoint-based matching, using hand-crafted or learned features (Sivic and Zisserman, 2003; Lowe, 2004; DeTone et al., 2018; Edstedt et al., 2024) to establish correspondences between the regions of interest. These keypoint-based methods are typically followed

by geometric constraints, such as epipolar geometry or stereo matching (Hartley and Sturm, 1997; Scharstein and Szeliski, 2002), to refine the detected regions and ensure geometric consistency. Methods like RANSAC (Fischler and Bolles, 1981) have been widely used to filter out mismatches and estimate the location and shape of co-visible regions. However, these bottom-up approaches are highly dependent on the accuracy of point-level correspondences, and they often fail when the initial keypoints are incorrectly matched. Furthermore, traditional approaches are sensitive to scale variations and frequently suffer from performance degradation in complex environments where occlusion or repetitive textures are prevalent.

Recent advancements in learning-based approaches have improved co-visible region detection by incorporating deep learning models that operate on a global scale. For example, MKPC (Song et al., 2023) uses a multi-scale keypoint-based model to estimate region-level correspondences based on the distribution range of matching points. These models improve robustness in cases of large viewpoint changes but remain sensitive to point-level matching errors, which limits their overall effectiveness in highly challenging scenarios.

Top-down methods like Normalized Surface Overlay (NSO) (Rau et al., 2020) use asymmetric feature spaces to predict overlapping regions in images, providing a rough estimate of co-visible areas. OETR (Chen et al., 2022) introduces attention mechanisms and multi-scale features to predict overlap regions in images with wide baselines, enhancing performance in scenarios with large viewpoint differences. Region-level correspondence methods like OETR (Chen et al., 2022) are also gaining attention in photogrammetry, particularly for wide-baseline aerial imagery. However, even with these advancements, point-level correspondence ambiguity remains a significant challenge, especially under large scale differences.

2.3. Challenges of Scale Variation in Attention Mechanisms

Previous methods have extensively used attention mechanisms due to their ability to compare features across different regions of an image. Attention has been effective in improving image matching by capturing long-range dependencies (Dosovitskiy et al., 2021). In particular, transformers have shown remarkable success in improving the robustness of point matching by allowing for feature interactions across the entire image. However, standard attention mechanisms have neglected the issue of scale variations.

Several strategies have been proposed to address this issue. Multi-scale attention mechanisms, such as those employed in PANet (Wang et al., 2021), combine information from different resolution levels to capture features at varying scales. CrossFormer (Wang et al., 2023b) takes this further by fusing both local and global attention to improve performance in cases of significant geometric changes. Pyramid Vision Transformers (PVT) (Mei et al., 2023) introduce keys and values fusion with a large stride to maintain both coarse and fine-grained details across different scales, improving the ability to detect features at various resolutions. The Shunted Transformer (Ren et al., 2022) enhances multi-scale modeling by downsampling feature keys and values, allowing the model to capture a broader range of scales efficiently. SMT (Lin et al., 2023) replaces multi-head attention with a lightweight multi-head hybrid convolution, allowing for more efficient multi-scale perception.

Despite these advances, a significant gap remains in handling extreme scale differences. Most existing attention mechanisms are either computationally expensive or fail to fully capture the necessary scale information to resolve the ambiguities introduced by large-scale variations. Our proposed SCoDe introduces a novel Scale Head Attention mechanism that addresses this issue by explicitly enabling feature correlation across multiple scale subspaces. By doing so, it ensures that even under large-scale discrepancies, reliable correspondences can be established between image pairs.

3. Methodology

This section begins by introducing the task of establishing point correspondences between image pairs. It then provides a detailed explanation of the proposed deep learning-based approach, SCoDe, for detecting co-visible regions between image pairs, including the descriptions of the core components and the loss function.

3.1. Problem Description

3.1.1. Point Matching

Given a pair of images, I_A and I_B, point matching aims to identify points in the two images that belong to the same 3D points in the scene. State-of-the-arts (Sarlin et al., 2020; Jiang et al., 2021) tackle this task with a deep neural

network, i.e.,

$$\mathcal{X} = \mathcal{F}(I_A, I_B) \tag{1}$$

In Equation 1, $\mathcal{X}$ is a set of N point correspondences, $\mathcal{X} := \{\{\mathbf{p}_A^1, \mathbf{p}_B^1\}, \{\mathbf{p}_A^2, \mathbf{p}_B^2\}\}, ..., \{\mathbf{p}_A^N, \mathbf{p}_B^N\}\}$. $\mathbf{p} := (u, v)$ is the point coordinate of the output correspondences in the input image I_A and I_B. Commonly, given an input image pair I_A and I_B, the model $\mathcal{F}$ operates on them directly, barely considering the potential large-scale difference between them (Xu and Zhang, 2020; Sun et al., 2021; Xie et al., 2024), and some methods (Yi et al., 2016; Dusmanu et al., 2019) implicitly handle the problem by leveraging scale-varied training data.

In practice, I_A and I_B can be captured at drastically different camera poses. The large baseline between the two cameras often causes the scale of I_A and I_B to be different. Such scale differences can lead to incorrect point correspondence, where a point in image I_B may be mismatched with different points in image I_A. Despite some methods (Xu and Zhang, 2020; Sarlin et al., 2020) aggregating image global and local information to reduce this confusion, our experiments will show that their estimation is not robust when the scale difference between I_A and I_B is large.

Instead, we address this issue by explicitly detecting the co-visible regions in I_A and I_B, and subsequently generating a new image pair I'_A and I'_B that contains only the areas visible in both views, thereby enabling a more reliable point matching.

3.1.2. Co-visible Region Detection

Given a pair of images, I_A and I_B, the objective of co-visible region detection is to find in I_A and I_B the window $I_A^{'}$ and $I_B^{'}$ that correspond to the same area in the 3D scene,

$$I_A^{'}, I_B^{'} = \mathcal{M}(I_A, I_B). \tag{2}$$

Generally, $I_A^{'}$ and $I_B^{'}$ are described using a rectangular bounding box (Chen et al., 2022), and we denote them using their anchor point coordinates p and the offsets b from the anchor point to the four sides.

We observe that the process of detecting co-visible regions bears similarities with salient object detection in a single image (Fan et al., 2021; Kong et al., 2022), as both aim to locate regions of interest at various scales and predict bounding boxes. Our proposed method draws inspiration from object detection and extends it to handle cross-view consistency.

Once the co-visible regions I'_A and I'_B are detected, we crop the corresponding patches from I_A and I_B, and normalize their relative scale by computing a scale factor from the bounding boxes and resampling one of the patches. To maximize the use of high-resolution information, we identify the region with the smaller spatial scale and resample it to match the other region based on the ratio of their shorter sides. The resulting scale-aligned regions are then fed into a feature matching network, which extracts and matches features within these regions. By restricting the search space to semantically consistent and geometrically plausible areas, this strategy effectively reduces mismatches caused by repetitive patterns and large scale differences, while also improving matching efficiency.

3.2. Proposed SCoDe

3.2.1. Overview

We propose SCoDe, an end-to-end **S**cale-aware **Co**-visible region **De**tection method that explicitly addresses the potential scale difference between input images. Similar to vision transformer-based object detection (Carion et al., 2020), SCoDe also leverages a learnable query for region detection. Unlike object detection methods that mostly focus on the single-view, SCoDe exchanges information from two input images for co-visible region detection.

As shown in Figure 2, a Siamese network, i.e., a CNN with two weight-sharing branches, extracts features f_A, f_B from input images I_A and I_B. Prior to feature extraction, both input images I_A and I_B are resized such that their longer side is specific pixels while maintaining the original aspect ratio. This rescaling standardizes the input dimensions and facilitates consistent processing in the Siamese CNN backbone. When I_A and I_B possess a scale difference, a large region in I_A can appear to be small in I_B. To match the corresponding regions, we design a novel Scale Head Attention (Section 3.2.2) to interact features from I_A with multi-scale features from I_B and vice versa,

$$\begin{aligned} f_A^{(i+1)} &= \mathcal{N}_{cross}(\mathcal{N}_{self}(f_A^{(i)}, f_A^{(i)}), \mathcal{N}_{self}(f_B^{(i)}, f_B^{(i)})), \\ f_B^{(i+1)} &= \mathcal{N}_{cross}(\mathcal{N}_{self}(f_B^{(i)}, f_B^{(i)}), \mathcal{N}_{self}(f_A^{(i)}, f_A^{(i)})). \end{aligned} \tag{3}$$

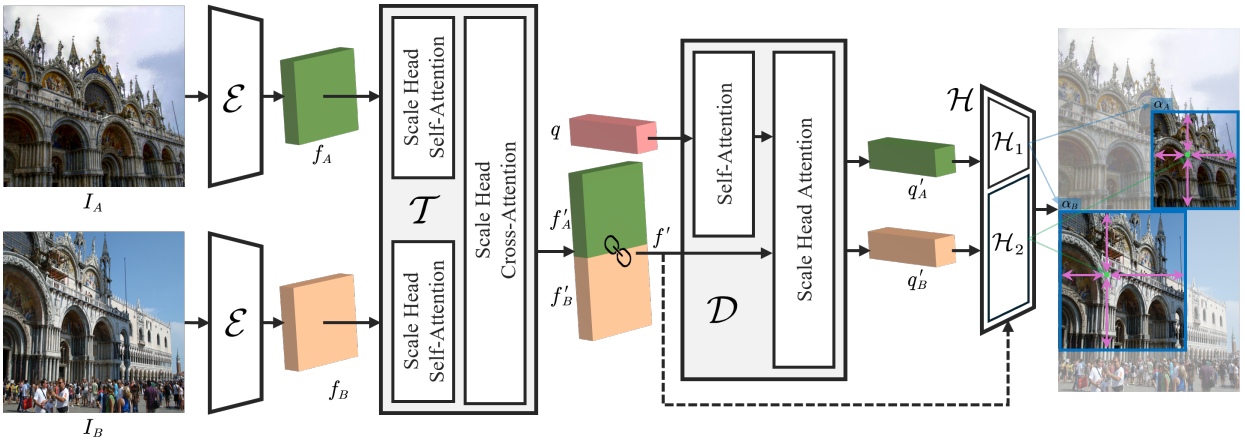


Figure 2: The SCoDe architecture. We pass the input image pairs through a shared CNN feature extractor $\mathcal{E}$ to generate feature maps f_A and f_B. The feature maps are fed into a transformer $\mathcal{T}$ with Scale Head Attention, and then we concatenate the outputs and feed them together with a learnable query tensor q into the decoder $\mathcal{D}$. Both transformer blocks $\mathcal{T}$ and $\mathcal{D}$ are based on SHA. The output query pairs together with the concatenated decoded features pass through the prediction header $\mathcal{H}$ to obtain anchor point p, border offset b, and confidence α representing the co-visible regions.

$\mathcal{N}$ represents our proposed SHA, and it includes Scale Head Self-Attention $\mathcal{N}_{self}$ and Scale Head Cross-Attention $\mathcal{N}_{cross}$. In Equation 3, $\mathcal{N}_{self}$ first highlights the salient features across multiple scales for each image, and then $\mathcal{N}_{cross}$ compares those features at different scales between views. We perform a total of M iterations, where at each iteration i ($i = 1, \dots, M$), the input features $f_A^{(i)}$ and $f_B^{(i)}$ are updated for the next iteration, $i+1$. After the final iteration M, the updated features $f_A^{(M)}$ and $f_B^{(M)}$ are denoted as $f_A^{'}$ and $f_B^{'}$. These features are then concatenated along the spatial dimension to form the final feature representation, i.e., $f^{'} = f_A^{'} \oplus f_B^{'}$, where $\oplus$ stands for concatenation along the vertical direction of the images.

Our learnable query decoder $\mathcal{D}$ (Section 3.2.3) then uses a query vector q to iteratively extract information from $f^{'}$. For both input images, I_A and I_B, $\mathcal{D}$ outputs feature vectors $q_A^{'}$ and $q_B^{'}$, that contains the information from both views. Then our prediction head $\mathcal{H}$ (Section 3.2.4) combines $q_A^{'}$ and $q_B^{'}$ with the concatenated scale-aware features $f^{'}$ for co-visible region detection,

$$q_A^{'}, q_B^{'} = \mathcal{D}(f^{'}, q), \quad I_A^{'}, I_B^{'} = \mathcal{H}(f^{'}, q_A^{'}, q_B^{'}). \tag{4}$$

Compared with OETR (Chen et al., 2022), SCoDe introduces key differences in both architecture and supervision. It incorporates a Scale Head Attention module for cross-scale interaction and a dual-head design to predict region geometry and confidence separately. SCoDe is also trained with a multi-task loss (Section 3.2.5) that jointly optimizes localization and confidence, while OETR focuses only on geometric regression. These changes improve robustness to scale differences and enhance confidence interpretability.

3.2.2. Scale Head Attention

Before introducing the Scale Head Attention (SHA) mechanism, we first review the standard attention mechanism (Dosovitskiy et al., 2021) and discuss its limitation when dealing with large scale variance. The standard attention mechanism is defined as,

$$\textbf{Atten}(f) = \text{softmax}\left(\frac{Q \cdot K^T}{\sqrt{d_K}}\right) \cdot V. \tag{5}$$

Here, Q, K, and V stand for query, key, and value mapped from the input features. In self-attention, the input feature is from a single image, and thus it does not exchange information across views. In cross-view attention, Q is mapped from one image while K and V are mapped from the other, allowing interacting information between views. Transformers (Dosovitskiy et al., 2021; Carion et al., 2020; Wang et al., 2023b), typically combine self-attention and cross-attention, allowing the model to better assess semantic matches between different views.

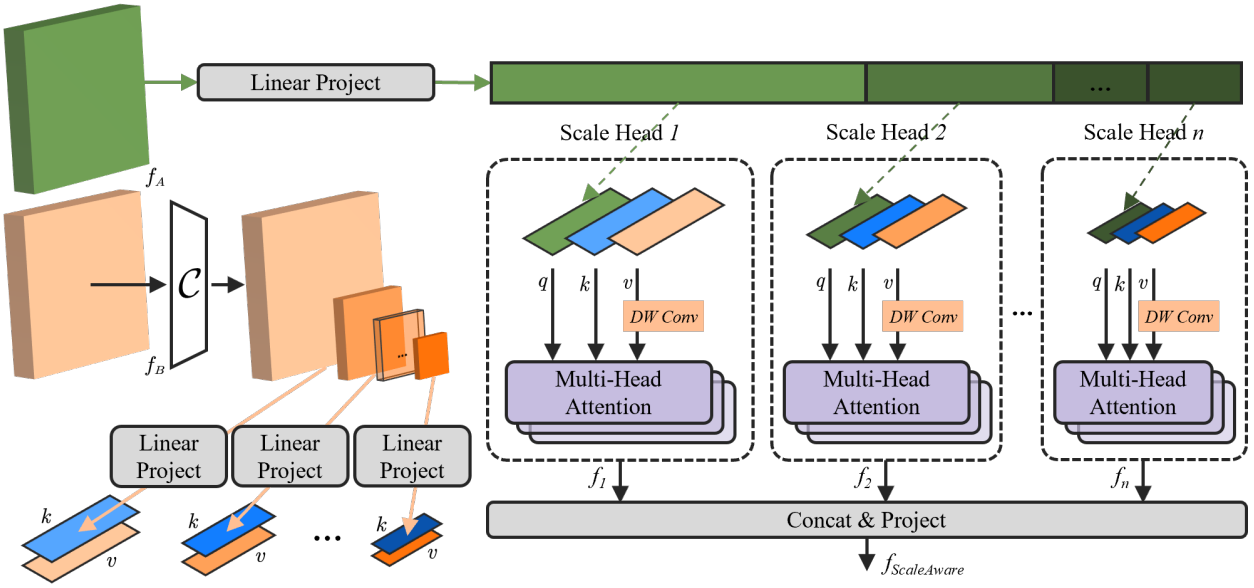


Figure 3: The schematic illustration of the proposed Scale Head Cross-Attention. Features are alternately projected as queries during the iterative process. Under the attention mechanism, features are projected into multiple subspaces of different scales through the convolution group $\mathcal{C}$. The novel design of the scale head is introduced.

However, standard attention mechanisms do not consider features of different scales. To address this limitation, we extend the standard self-attention and cross-attention into Scale Head Self-Attention and Scale Head Cross-Attention.

Scale Head Attention: Our Scale Head Attention $\mathcal{N}$ maps input 2D feature pairs into n scale subspace $s, s \in \{1, 2, \dots, n\}$. Here we explain Scale Head Cross-Attention $\mathcal{N}_{cross}$, where two features, f_A and f_B, are extracted from different images I_A and I_B. Similarly, Scale Head Self-Attention $\mathcal{N}_{self}$ is used when both features are from the same image. Scale $s = 1$ corresponds to the smallest scale, i.e. its feature has the highest spatial resolution, facilitating the detection of smaller objects. Similarly, $s = n$ represents the largest scale, suited for detecting larger objects. Each scale subspace s has corresponding mapping functions $P_q^s(\cdot)$, $P_k^s(\cdot)$, and $P_v^s(\cdot)$ to generate the query, key, and value for the attention at the current scale,

$$f_A^{(s)} = \textbf{Atten}_{MultiHead}(Q_A^{(s)}, K_B^{(s)}, V_B^{(s)}), \quad s \in \{1, 2, \dots, n\}. \tag{6}$$

As shown in Figure 3, we split flattened f_A into $Q_A^{(1)}, Q_A^{(2)}, \dots, Q_A^{(n)}$ and use multi-kernel convolution $\mathcal{C}$ to map f_B into keys and values of n corresponding scale subspaces, acquiring $K_B^{(1)}, K_B^{(2)}, \dots, K_B^{(n)}$ and $V_B^{(1)}, V_B^{(2)}, \dots, V_B^{(n)}$. Notably, a previous study (Wang et al., 2022) has shown that applying a depth-wise convolution (DW Conv) along the channel dimension of the value vector can effectively enhance local feature information. Therefore, we incorporate a DW Conv layer after $V_B^{(s)}$ to improve the query's sensitivity to spatial information. For an individual scale head s, as illustrated in Figure 3, the components $Q_A^{(s)}, K_B^{(s)}, V_B^{(s)}$ within the scale head contribute to the multi-head attention output

$$\begin{aligned} f_A^{(s)} &= \bigoplus_{j=1}^{m} \text{softmax}\left(\frac{Q_{Aj}^{(s)} \cdot K_{Bj}^{(s)T}}{\sqrt{d_K}}\right) \cdot \textbf{Conv}_{DW}(Q_{Bj}^{(s)}), \\ \mathcal{N}(f_A, f_B) &= \bigoplus_{s=1}^{n} f_A^{(s)}. \end{aligned} \tag{7}$$

In Equation 7, m denotes the number of attention heads within each scale head, and n is the total number of scale heads. The outputs from the individual scale heads, denoted as $f_A^{(s)}$, are concatenated to generate the final attention result. By swapping f_A and f_B during iterations, SHA enables multi-scale correlations across views.

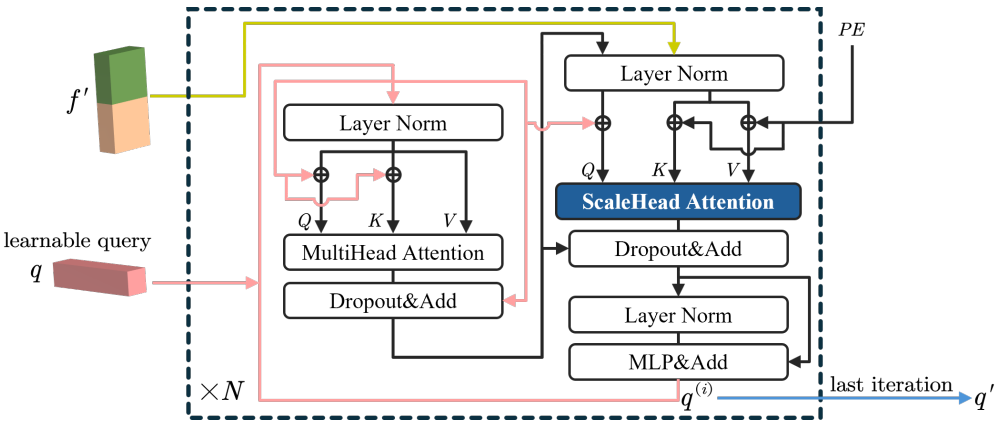


Figure 4: The architecture of the learnable query decoder. It decodes the learnable query by self-correlation and Scale Head Attention with the input scale-aware feature $f^{'}$.

Using the SHA mechanism, the encoder $\mathcal{T}$ then generates the scale-aware feature $f_A^{'}$ and $f_B^{'}$ for image I_A and I_B, respectively. These features are concatenated along the spatial dimension to form $f^{'}$, following Section 3.2. Subsequently, we design a decoder $\mathcal{D}$ to query the locations of co-visible regions from $f^{'}$.

3.2.3. Learnable Query Decoder

With the scale-aware feature $f^{'}$ extracted, the next step is to detect co-visible regions, which we frame as an object detection task. Inspired by COTR (Jiang et al., 2021), our decoder $\mathcal{D}$ employs a learnable query q to accurately locate co-visible regions within $f^{'}$, ensuring robust matching.

As illustrated in Figure 4, the decoder performs iterative querying N times. In each iteration i, the input query q undergoes standard self-attention (Equation 5), producing a self-correlated query $q^{(i)}$. To provide spatial cues, we add positional embeddings to both the queries and the encoded feature. The resulting $q^{(i)}$ is then used to query co-visible regions from the concatenated scale-aware feature $f^{'}$ using Scale Head Cross-Attention,

$$\begin{aligned} q^{(i)} &= \mathbf{Atten}(Q_q^{(i)}, K_q^{(i)}, V_q^{(i)}), \\ [q_A^{(i+1)}, q_B^{(i+1)}] &= \mathcal{N}_{cross}(q^{(i)}, f^{'}). \end{aligned} \tag{8}$$

In the encoder, Q, K, and V come from image features of different views. In the decoder, Q is from the learnable query, while K and V are from the encoded features f'. Here, q is projected to multi-scale queries, $Q^{(i)} = P_q(q^{(i)})$, and $f^{'}$ is mapped to the corresponding keys, $K = P_k(C(f^{'}))$, and values, $V = P_v(C(f^{'}))$. Concretely, we have,

$$\begin{aligned} [q_A^{(i+1)}, q_B^{(i+1)}] &= \bigoplus_{s=1}^{n} \mathbf{Atten}_{MultiHead}, [Q^{(s)}, K^{(s)}, V^{(s)}] \\ q_A^{'} &= q_A^{(N)}, q_B^{'} = q_B^{(N)}. \end{aligned} \tag{9}$$

After N iterations, $\mathcal{D}$ outputs the final query features $q_A^{'}, q_B^{'}$, which encapsulates the spatial information of co-visible regions. These features are then passed to the region prediction head $\mathcal{H}$, which predicts the specific locations and sizes of these co-visible regions within images I_A and I_B.

3.2.4. Region Prediction Head

Our prediction head $\mathcal{H}$ consists of two sub-modules: a co-visible region regressor $\mathcal{H}_1$ and a confidence predictor $\mathcal{H}_2$. It combines scale-aware features $f_A^{'}, f_B^{'}$ with query features $q_A^{'}, q_B^{'}$ to jointly predict the location, size, and confidence of co-visible regions:

$$\begin{aligned} \mathcal{H}(\cdot) &= \left(\mathcal{H}_1(\cdot), \mathcal{H}_2(\cdot)\right) \\ \mathbf{p}, \mathbf{b} &= \mathcal{H}_1(q_A^{'}, f_A^{'}, q_B^{'}, f_B^{'}) \\ \alpha &= \mathcal{H}_2(q_A^{'}, f_A^{'}, q_B^{'}, f_B^{'}). \end{aligned} \tag{10}$$

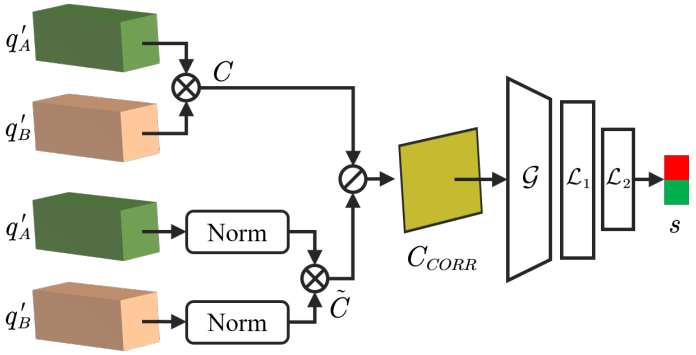


Figure 5: The prediction head of co-visible confidence. This module combines an unlearnable pseudo-cosine similarity structure with a learnable binary confidence regression network.

In Equation 10, $\mathbf{p}$ is the anchor point, $\mathbf{b}$ represents the bounding box offsets from the center, and α denotes the confidence score of the detected co-visible region bounding box. The co-visible region regression and confidence prediction will be detailed below.

Co-visible Region Regression $\mathcal{H}_1$: Bounding box offsets $\mathbf{b}$ are regressed using the query feature $q^{'}$ with a Multi-Layer Perceptron (MLP) to ensure both high accuracy and low computational cost.

The scale-aware feature $f^{'}$ captures global semantics of the image pair. Attention weights are computed by the dot product between $f^{'}$ and $q^{'}$:

$$\mathbf{w} = f^{'} \cdot q^{'}, \tag{11}$$

and converted into a probability distribution over spatial locations via a SoftMax function. Denoting $\mathbf{x}_i$ as candidate spatial coordinates (e.g., grid centers on the feature map), the semantic center $\mathbf{p}$ is calculated as the expected value of this distribution:

$$\mathbf{p} = \sum \mathbf{x}_i \cdot P(x = \mathbf{x}_i),\ P = \text{SoftMax}(\mathbf{w} \cdot f^{'}), \tag{12}$$

where $\mathbf{p}$ represents the predicted anchor point of the co-visible region.

Together, $\mathbf{p}$ and the regressed offsets $\mathbf{b}$ (as defined in Equation 10) determine the final bounding box specifying the spatial extent of the co-visible region in both I_A and I_B.

Co-visible Region Confidence $\mathcal{H}_2$: Even when the input images lack co-visible regions, the system may still output a bounding box. To address this, we propose a confidence score to assess the presence and reliability of the detected co-visible region. The query feature $q^{'}$ provides local semantic information to estimate the similarity confidence $\mathbf{s} := (y, n)$.

As illustrated in Figure 5, the pseudo-cosine similarity matrix C_{CORR} is computed between $q^{'}_A$ and $q^{'}_B$:

$$C_{CORR} = \frac{C}{\tilde{C}} = \frac{q^{'}_A \times q^{'}_B}{\|q^{'}_A\| \times \|q^{'}_B\|}, \tag{13}$$

where the initial correlation matrix C is normalized to obtain $\tilde{C}$, and C_{CORR} is the element-wise division of C by $\tilde{C}$.

The matrix C_{CORR} captures similarity information in a high-dimensional space, which is further processed by a convolution group $\mathcal{G}$ to integrate semantic and geometric details. Finally, linear layers $\mathcal{L}_1$ and $\mathcal{L}_2$ regress a binary vector $\mathbf{s}$, representing the confidence in the co-visible region's similarity.

3.2.5. Losses

Our object detection approach for region-level correspondences necessitates focusing on not only the accuracy of the bounding box locations and their sizes but also the precision of co-visible region matching. Additionally, it is crucial to evaluate the reliability of the identified co-visible regions. To ensure comprehensive detection and matching supervision, we introduce a loss group $\mathbf{L}$

$$\mathcal{L} = \Lambda \cdot \mathbf{L} = \Lambda \cdot [L_{loc}, L_{L1}, L_{iou}, L_{cycle}, L_{sim}] = \Lambda \cdot [|\mathbf{p} - \hat{\mathbf{p}}|, |\mathbf{b} - \hat{\mathbf{b}}|, \mathcal{L}_{iou}(\mathbf{r}, \hat{\mathbf{r}}), |\mathbf{c} - \tilde{\mathbf{c}}|, \mathcal{F}_{CE}(\mathbf{s} - \hat{\mathbf{s}})], \tag{14}$$

including detection losses like localization loss L_{loc}, Least Absolute Deviations L_{L1}, and generalized IoU loss L_{iou}, as well as matching losses like symmetric center consistency loss L_{cycle} and similarity cross-entropy loss L_{sim}.

In Equation 14, $\mathbf{p}$ and $\mathbf{b}$ denote the predicted center and size of the co-visible region, forming the bounding box $\mathbf{r} = (\mathbf{p}, \mathbf{b})$; $\hat{\cdot}$ indicates the corresponding ground truth. The binary vector $\mathbf{s}$ represents predicted region similarity confidence. The vector $\mathbf{c}$ is the predicted center coordinate of the co-visible region in one image, while $\tilde{\mathbf{c}}$ denotes the center derived by applying the same prediction process to the reversed input, i.e., by swapping the feature pair $(f_A^{'}, q_A^{'}) \rightarrow (f_B^{'}, q_A^{'})$. This enforces geometric consistency across the image pair via the symmetric center consistency loss L_{cycle}. The hyperparameters $\lambda_{loc}, \lambda_{L1}, \lambda_{iou}, \lambda_{cycle}, \lambda_{sim} \in \mathbb{R}$ in Λ balance the loss weights, ensuring stable supervision across the detection network for center estimation, bounding box regression, region correspondence, and confidence prediction.

4. Experiments

This section first introduces the used datasets, followed by the implementation details and evaluation metrics. Then we provide experimental results that demonstrate the superiority of SCoDe over previous state-of-the-art approaches for stand-alone co-visible region detection as well as for integrating it for point matching. Finally, we conduct ablation studies and assess the method's generalization ability across diverse scenarios.

4.1. Datasets

We use **MegaDepth** (Li and Snavely, 2018) dataset for both co-visible region detection and image matching, as it provides ground truth correspondences, along with sparse 3D reconstructions and depth maps generated by COLMAP (Schönberger and Frahm, 2016; Schönberger et al., 2016). MegaDepth consists of 1,000,000 images across 196 outdoor scenes, depicting various lighting conditions, scales, and viewpoints. These characteristics simulate the challenges of real-world image matching.

Following previous work (Chen et al., 2022), we select 10 scenes from MegaDepth for testing and randomly pick 3,000 image pairs with sufficient co-visible regions. All training and main evaluations are conducted at 640×640 to balance training cost, performance, and generalization. Ground truth co-visible regions are determined using ray-casting based on the provided camera parameters and depth maps. Negative samples are taken from unrelated scenes. Additional experiments on other resolutions are also conducted for comparison and general evaluation.

Additionally, we qualitatively assess generalization capability using **ScanNet** (Dai et al., 2017) and **GL3D** (Shen et al., 2018; Luo et al., 2018). ScanNet contains 2.5 million views from 1,513 indoor environment scans, primarily designed for indoor scene understanding. In contrast, GL3D features high-resolution images captured by drones with significant viewpoint and scale variations, making it a valuable resource for applications such as aerial photogrammetry and large-scale 3D reconstruction. Although both datasets lack ground truth co-visible regions, they present diverse and challenging scenarios that are suitable for qualitative analysis.

4.2. Implementation Details

Before feature extraction, input image pairs are resized such that their longer side is 640 pixels, while maintaining the original aspect ratio, as specified in Section 4.1. We use ResNet-50 (He et al., 2016) as our feature extractor $\mathcal{E}$. Scale Head Attention employs non-overlapping multi-kernel convolutions, where the receptive fields do not overlap, reducing both channel and spatial dimensions. During training and testing, we iteratively apply the encoder $\mathcal{T}$ four times ($M = 4$) and the decoder $\mathcal{D}$ two times ($N = 2$). The number of scale heads n is set to 3 to effectively capture information from objects of varying sizes. To facilitate reproducibility, Table 1 provides a concise summary of the main network architecture, including the key components and their configurations; the complete layer-wise details and parameter statistics are provided in Appendix A.

The hyperparameter group Λ is set to $[2.5, 2.5, 2, 2, 1]$, balancing model performance across different functions. For optimization, we use AdamW (Loshchilov and Hutter, 2017) with a batch size of 4 and an initial learning rate of 10^{-4}, which decays by a factor of 0.1 at the 25th epoch. The learning rate of the confidence regression module is initialized at 0.1 times the main learning rate to ensure synchronized convergence with the overall network. We trained SCoDe on MegaDepth using 128,000 triplet image sets, which include ground truth co-visible regions and negative samples. The network converged after 45 hours of training with 40 epochs on 4 NVIDIA RTX A6000 GPUs.

In all experiments, we retain the top $N = 2048$ keypoints per image after default non-maximum suppression. Descriptor matching is performed using either Nearest Neighbor (Gutin et al., 2002) search with Lowe's ratio test

Table 1
Summary of Key Architectural Components

Component	Module	I/O Shape	Params	Note
Extractor	ResNet-50	(B,3,H,W) → (B,1024,H/16,W/16)	Pretrained	
	Projection	(B,1024,...) → (B,256,H/32,W/32)	1×1 conv, s=2	
Transformer	Reshape + Query	(B,256,...) → (B,N,256), (B,2,256)	–	Flatten + Embed
	Encoder $\mathcal{T}$	(B,N,256) → (B,N,256)	4×2 blocks	SHA + MLP
	Decoder $\mathcal{D}$	(Query + Feature) → (B,2,256)	2 layers	Self/Cross-attn
SHA Block	Paths 0/1/2	(B,256,...) → (B,128/256,...)	k=1/3/5, s=1/3/5	Multi-scale
TLBR Head $\mathcal{H}_2$	Linear ×2	(B,2,256) → (B,2,4)	256→256→4	Box regression
Heatmap Head	Conv ×2	(B,256,...) → (B,1,...)	3×3, 1×1	
Conf. Head $\mathcal{H}_1$	CNN $\mathcal{G}$	(B,1,...) → (B,32,...)	3×3 + pool ×2	Feature extract
	MLP $\mathcal{L}_{1,2}$	(B,32×H×W) → (B,2)	256→2, Softmax	Classification

(threshold = 0.7) (Lowe, 2004) or via learned matchers such as SuperGlue (Sarlin et al., 2020). For classical feature pipelines like SIFT (Lowe, 1999), descriptors are matched directly without learned refinement. In deep pipelines, matches are established within the detected co-visible regions, ensuring spatial consistency.

4.3. Evaluation Metrics

We adopt a variety of metrics to evaluate our method on both co-visible region detection and image matching tasks.

Co-visible Region Detection. We compute Intersection-over-Union (IoU) between predicted and ground-truth boxes. To evaluate robustness, we report the average IoU over thresholds (e.g., 0.5, 0.75, 0.9). We also report Overlap IoU (OIoU), which measures IoU within overlapping visible areas. Note that the so-called mean IoU in our context refers to averaging IoU over test samples rather than semantic classes.

To assess confidence reliability, we report Recall@Threshold, measuring the fraction of ground-truth regions with predicted confidence above a threshold. This metric applies only to SCoDe, which explicitly learns region-level confidence via supervised training.

Image Matching. We follow common practice (Sun et al., 2021; Sarlin et al., 2020) and evaluate pose estimation using rotation and translation errors. Metrics include Area Under the Curve (AUC), Accuracy (Acc), and mean Average Accuracy (mAA) computed from the pose error distribution. Matching Score (MS) denotes the mean confidence of all inlier correspondences. Precision (P) is computed as the percentage of correspondences with epipolar error below a threshold after pose recovery.

4.4. Evaluation on Co-visible Region Detection

We compare our proposed co-visible region detection approach with the state-of-the-art method OETR (Chen et al., 2022). Because this task has been largely overlooked in the literature, OETR is the only comparable method. To ensure a fair comparison, we re-trained OETR under the same conditions to eliminate implementation discrepancies and verify the OETR's pre-trained weights. OETR provides two sets of pre-trained weights, *best* and *cyclecenter*, for evaluation. We compare our method to the *cyclecenter* weights, as they perform best on benchmark testing datasets.

The test results on the MegaDepth are shown in Table 2. Overall, SCoDe achieves consistent and substantial improvements over OETR in all metrics. Compared to the baseline *cyclecenter*, the proposed SCoDe improves IoU@0.75 by 12.4%, OIoU@0.9 by 22.46%, and mIoU by 4.69%. Notable gains are also observed across other thresholds. This demonstrates the significant improvement of SCoDe for co-visible region detection. On the other hand, SCoDe achieves an exceptionally high recall rate, exceeding 99% across all thresholds while estimating the confidence of co-visible regions. These improvements can be attributed to the explicit focus of SCoDe on the representation of co-visible regions across multiple scale levels. By leveraging the Scale Head Attention mechanism, SCoDe enables better feature interaction across scales, resulting in more robust and precise co-visible region detection and matching.

As our approach shows significant improvements in co-visible region detection, the next section will explore how these advancements contribute to more accurate image matching.

Table 2
Quantitative evaluation results on MegaDepth. OETR* denotes our reproduced results obtained by re-training under the same conditions on the same dataset. We report recall at multiple thresholds for intersection-over-union (IoU) and overlap intersection-over-union (OIoU), as well as mean intersection-over-union (mIoU) and confidence recall for co-visible region prediction. The boldface text indicates the best method under each metric. '-' means the method does not predict the confidence.

Methods	Confidence Recall			IoU Recall			OIoU Recall			mIoU
	@0.5	@0.75	@0.9	@0.5	@0.75	@0.9	@0.5	@0.75	@0.9	
OETR* (reproduced)	-	-	-	87.40	60.71	36.51	98.03	89.17	70.90	76.71
OETR (best)	-	-	-	88.94	57.53	34.43	96.13	77.80	52.87	76.33
OETR (cyclecenter)	-	-	-	90.71	59.03	35.92	95.93	76.51	52.36	77.29
SCoDe	**99.74**	**99.83**	**99.89**	**94.69**	**71.43**	**42.64**	**98.74**	**91.74**	**74.82**	**81.98**

Table 3
Evaluation on MegaDepth for all-range scale differences. We report the area under the curve (AUC), accuracy (Acc), and mean average accuracy (mAA) at 5, 10, and 20 degrees, as well as precision (P) and matching scores (MS). Based on the matching pipeline, we compare the baseline ('-' means without co-visible region detection), OETR guided, and SCoDe matching paradigms for pose estimation and matching performance. The boldface text indicates the best method under each group of matching methods.

Methods		AUC			Acc				mAA			P	MS
		@5	@10	@20	@5	@10	@15	@20	@5	@10	@20		
LoFTR	-	41.79	56.05	66.87	63.52	74.25	77.72	79.89	63.52	68.89	73.85	60.37	50.81
	+OETR	42.69	57.24	68.58	65.09	**76.14**	**80.10**	82.08	65.09	70.61	75.85	**71.02**	52.57
	+SCoDe	**43.30**	**58.03**	**68.97**	**66.66**	76.12	79.93	**82.71**	**66.66**	**71.39**	**76.35**	70.89	**52.96**

4.5. Evaluation on Image Matching

Detecting co-visible regions before point matching provides critical prior information, which is essential for improving the accuracy and reliability of image matching. In our evaluation, we compared SCoDe with the baseline method, OETR (Chen et al., 2022), using the LoFTR (Sun et al., 2021) point matching backbone. As illustrated in Table 3, SCoDe outperforms the baseline in all metrics.

SCoDe is specifically designed to address scale variations. To verify its effectiveness in large scale difference scenarios, we conducted an evaluation on the MegaDepth test set, focusing on image pairs with a scale ratio greater than 2. Table 4 shows the comparison between the performance of combining the same feature extraction and matching methods with OETR and with SCoDe. SCoDe demonstrated improvements in AUC, accuracy, and mAA within a 20-degree range by 10% to 20%, especially with traditional matching methods like Nearest Neighbour (Gutin et al., 2002). Significant improvements were also observed compared to the learning-based SuperPoint+SuperGlue matching backbone.

As shown in Table 2, SCoDe estimates co-visible regions more accurately than OETR, and Table 4 highlights its superior matching performance under large scale differences. This demonstrates that precise co-visible region estimation leads to more reliable pose estimation and image matching. The precision of 87.55% attained by SCoDe highlights its effectiveness, bringing an 8.41% improvement to the SuperPoint+SuperGlue matching pipeline and consistently outperforming both traditional and learning-based matching methods, especially in challenging scenarios with large scale variations.

To further verify the effectiveness of our region-guided matching strategy beyond learning-based methods, we conduct additional experiments using classical SIFT (Lowe, 1999) features. Following the protocol in Table 3, we randomly select a subset of MegaDepth test pairs with diverse scale variation. SIFT keypoints and descriptors are extracted, and matching is performed via Nearest Neighbour (Gutin et al., 2002) search without any learned refinement. Our co-visible region detection is then applied to constrain matching within geometrically meaningful areas. The results are reported in Table 5.

Table 4
Evaluation on MegaDepth for larger scale differences. Each row corresponds to a complete matching pipeline. Specifically, we consider combinations of feature extractors and matchers, including SuperPoint (SP) (DeTone et al., 2018), DISK (Tyszkiewicz et al., 2020), D2-Net (D2) (Dusmanu et al., 2019), ContextDesc (CON) (Luo et al., 2019), ASLFeat (ASL) (Luo et al., 2020), R2D2 (Revaud et al., 2019), and LoFTR (Sun et al., 2021). These are paired with either Nearest Neighbor (NN) (Gutin et al., 2002) or SuperGlue (SG) (Sarlin et al., 2020) for matching, except LoFTR, which is an end-to-end dense matching framework.

Methods		AUC			Acc				mAA			P	MS
		@5	@10	@20	@5	@10	@15	@20	@5	@10	@20		
SP+NN	-	2.26	3.81	5.91	3.91	6.15	7.64	9.90	3.91	5.03	6.90	11.54	3.13
	+OETR	10.16	16.54	23.55	18.52	26.24	30.77	33.49	18.52	22.38	27.26	31.29	7.92
	+SCoDe	**13.75**	**22.57**	**32.33**	**24.49**	**35.81**	**42.27**	**46.34**	**24.49**	**30.15**	**37.23**	**41.30**	**10.21**
DISK+NN	-	3.30	4.95	6.70	5.31	7.12	8.29	9.19	5.31	6.22	7.48	16.89	0.52
	+OETR	21.04	33.35	46.19	37.36	51.80	59.41	64.15	37.36	44.58	53.18	60.88	12.11
	+SCoDe	**23.10**	**35.81**	**48.81**	**40.15**	**54.67**	**61.97**	**67.02**	**40.15**	**47.41**	**55.95**	**63.72**	**14.05**
D2+NN	-	0.36	0.61	1.35	0.53	1.05	1.76	2.80	0.53	0.79	1.53	6.12	2.68
	+OETR	4.17	6.73	10.18	7.26	10.90	13.38	16.00	7.26	9.08	11.89	23.10	2.87
	+SCoDe	**5.70**	**9.54**	**14.53**	**10.59**	**15.32**	**19.38**	**22.79**	**10.59**	**12.95**	**17.02**	**29.86**	**3.23**
CON+NN	-	20.43	32.56	46.10	36.19	51.86	60.08	64.77	36.19	44.03	53.23	61.69	7.68
	+OETR	23.32	37.10	51.29	41.75	57.81	65.78	70.62	41.75	49.78	58.99	67.45	11.21
	+SCoDe	**25.63**	**39.72**	**53.90**	**44.58**	**60.77**	**68.56**	**72.94**	**44.58**	**52.67**	**61.71**	**68.80**	**12.32**
ASL+NN	-	11.44	19.21	28.79	20.67	31.46	38.67	43.73	20.67	26.07	33.64	38.68	18.40
	+OETR	21.95	35.94	50.88	40.25	57.20	66.22	71.17	40.25	48.73	58.71	60.89	34.55
	+SCoDe	**23.83**	**38.28**	**53.41**	**42.87**	**60.39**	**68.97**	**73.52**	**42.87**	**51.63**	**61.44**	**64.55**	**38.17**
LoFTR	-	25.37	37.82	49.61	42.86	55.44	61.45	66.17	42.86	49.15	56.48	48.21	40.76
	+OETR	31.70	47.14	60.55	53.71	68.73	74.24	77.75	53.71	61.22	68.61	71.22	46.70
	+SCoDe	**34.11**	**50.14**	**64.35**	**57.55**	**72.63**	**78.72**	**82.16**	**57.55**	**65.09**	**72.76**	**76.54**	**48.42**
R2D2+NN	-	12.86	22.46	33.22	25.34	36.81	44.52	49.13	25.34	31.08	38.95	39.60	85.15
	+OETR	24.96	39.90	55.20	44.52	61.85	71.19	75.42	44.52	53.19	63.24	65.09	88.39
	+SCoDe	**27.20**	**42.96**	**58.25**	**48.60**	**66.18**	**74.26**	**77.99**	**48.60**	**57.39**	**66.76**	**68.84**	**88.64**
DISK+SG	-	15.69	25.91	37.27	28.88	42.00	48.79	53.40	28.88	35.44	43.27	35.45	6.98
	+OETR	21.04	33.35	46.19	37.36	51.80	59.41	64.15	37.36	44.58	53.18	60.88	12.11
	+SCoDe	**23.10**	**35.81**	**48.81**	**40.15**	**54.67**	**61.97**	**67.02**	**40.15**	**47.41**	**55.95**	**63.72**	**14.05**
SP+SG	-	25.11	39.23	54.03	44.23	60.33	69.44	74.86	44.23	52.28	62.21	79.14	9.88
	+OETR	29.32	45.56	61.37	51.02	69.02	77.99	82.07	51.02	60.02	70.03	86.17	19.00
	+SCoDe	**31.18**	**48.27**	**63.77**	**55.06**	**72.38**	**79.85**	**83.87**	**55.06**	**63.72**	**72.79**	**87.55**	**21.51**

Although SIFT is inherently designed to handle scale variations, our method still achieves consistent improvements across all evaluation metrics, while maintaining a stable matching score (MS). This suggests that the effectiveness of SCoDe arises not from altering the keypoint detector or descriptor itself, but from constraining the matching process to geometrically and semantically meaningful regions. By limiting correspondence search to areas that are both co-visible and scale-aligned, our method suppresses false matches commonly caused by repetitive structures or scale ambiguities—issues that traditional hand-crafted features like SIFT struggle with. This targeted filtering reduces descriptor ambiguity and increases the distinctiveness of potential matches, even without any learning-based refinement. These results indicate that the gains introduced by SCoDe stem from a fundamental rethinking of the matching strategy, making them broadly applicable and not confined to specific neural architectures.

Given the strong performance of SuperPoint and SuperGlue in previous evaluations, we selected them alongside SCoDe for our visualizations. Figure 6 presents qualitative results from the MegaDepth test scenes. The first row shows the matching results on the original image pair without addressing scale differences, while the second and third rows show results with co-visible regions detected by OETR and our proposed SCoDe respectively.

Table 5
Comparison of SIFT matching performance under different region constraints on scale-diverse image pairs from the MegaDepth test set. Our method SCoDe consistently improves SIFT matching performance compared to the baseline and OETR across all metrics, enhancing its robustness through co-visible region filtering.

Methods		AUC			Acc				mAA			P	MS
		@5	@10	@20	@5	@10	@15	@20	@5	@10	@20		
SIFT+NN	-	6.55	11.24	16.96	12.16	18.16	22.94	25.87	12.16	15.16	19.78	13.20	**85.72**
	+OETR	8.11	13.58	20.25	14.91	22.23	26.87	30.92	14.91	18.57	23.73	16.60	85.09
	+SCoDe	**8.79**	**15.03**	**22.18**	**16.30**	**24.52**	**29.23**	**33.10**	**16.30**	**20.41**	**25.79**	**17.60**	85.26

The co-visible region detection of SCoDe significantly improves correspondence reliability under large scale ratios. For instance, in Figure 6(a), OETR fails to detect co-visible regions in the church image, leading to false matches. In contrast, SCoDe accurately identifies these regions, even with minimal visible area. This success is due to that its scale heads capture rich scale subspace correlations. In Figure 6(b), where more statues become visible at smaller scales, OETR's detection is biased, but SCoDe reliably matches the regions, even under semantic asymmetry. In Figure 6(c), SCoDe offers precise co-visible boundaries, avoiding overly broad or narrow regions that lead to geometric errors.

4.6. Resolution-Aware Matching Performance

To evaluate the scalability and robustness of our co-visible region prior under varying input resolutions, we conduct experiments using two representative feature matching pipelines: SuperPoint (DeTone et al., 2018) combined with

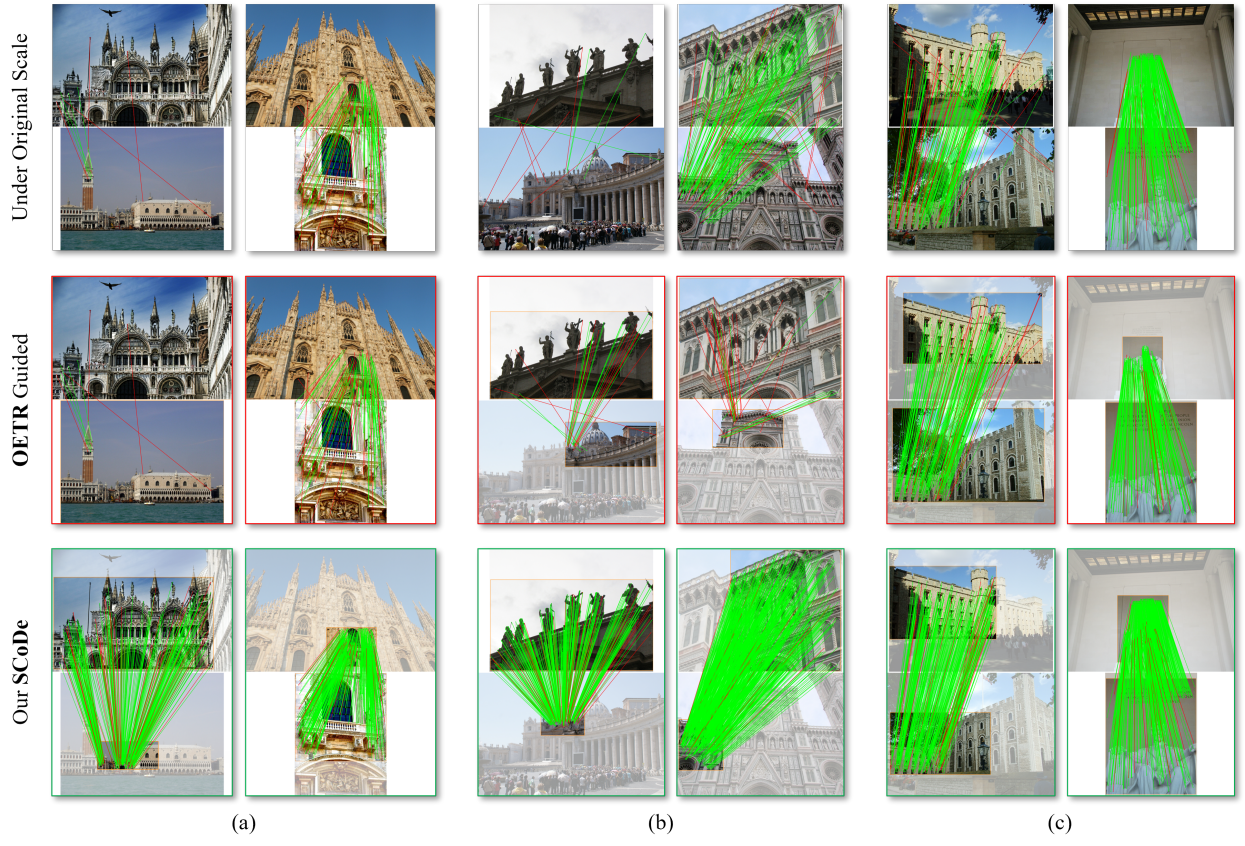


Figure 6: Qualitative examples in test scenes. We compare with the state-of-the-art method on testing scenarios. The highlighted windows in the figure represent the detected co-visible regions.

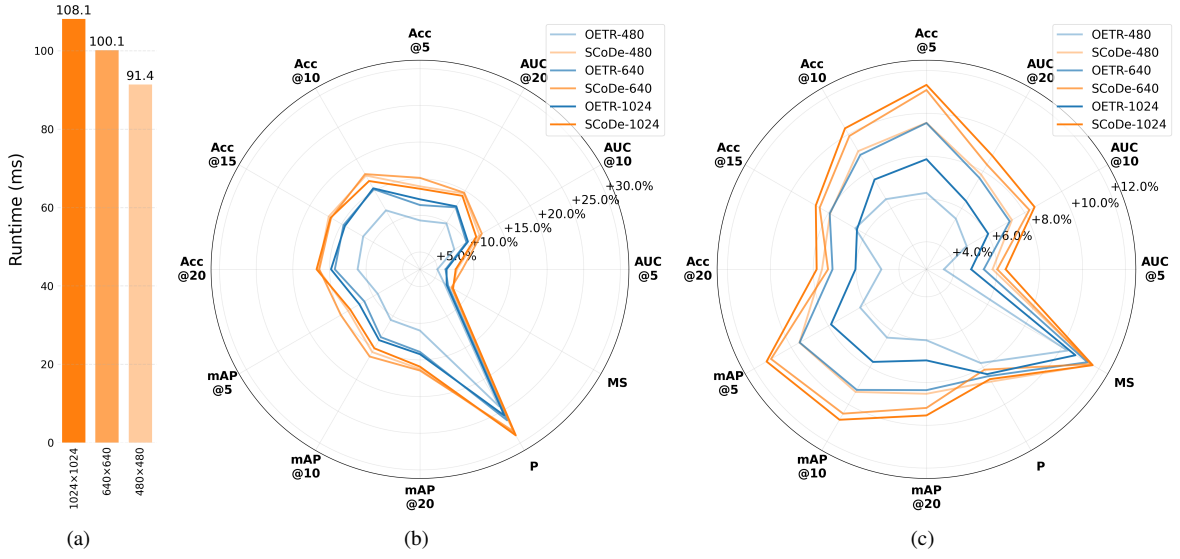


Figure 7: Performance and Runtime of SCoDe Across Input Resolutions. Performance gain is measured as percentage improvement over the baseline (0%), representing the respective matcher without SCoDe. (a) Average runtime for co-visible region detection at 480×480, 640×640, and 1024×1024, averaged over scenes and matchers. (b) Relative performance gain using LoFTR. (c) Performance gain using SuperPoint + SuperGlue. Orange bars: SCoDe (ours); Blue bars: OETR.

SuperGlue (Sarlin et al., 2020), which represents a two-stage architecture, and LoFTR (Sun et al., 2021), a fully end-to-end approach. We test these matchers at three input resolutions: 480×480 and 640×640, and 1024×1024, to reflect realistic image scales in practical applications. Performance is assessed using standard metrics such as matching precision, inlier ratio, and pose accuracy. To provide an intuitive comparison, we visualize results using radar charts, where improvements are expressed as percentage gains over a baseline pipeline centered at the origin. Orange and blue lines indicate our method and the OETR prior, respectively, with lighter and darker shades denoting different resolutions.

Results in Figure 7(b) and (c) demonstrate that our method consistently outperforms both the original pipeline and OETR across all evaluation metrics and resolutions. In particular, while OETR shows sensitivity to resolution degradation, our approach remains robust, maintaining consistent performance gains at lower and higher resolutions. This resolution-invariant behavior highlights the generality and stability of our region prior. The improvements are especially pronounced when integrated into the SuperPoint+SuperGlue pipeline, which lacks built-in global spatial modeling. This suggests that our method effectively compensates for architectural limitations by providing spatial context through semantically and geometrically meaningful region constraints.

In addition to accuracy, we evaluate computational scalability by measuring the average runtime of co-visible region extraction. For both pipelines, we compute the per-scene extraction time and then aggregate across all scenes and matchers to minimize variance. As shown in Figure 7(a), the runtime increases only marginally from 640 to 1024 resolution, confirming the efficiency of our method. This property is particularly relevant to scalable photogrammetric applications, where high-resolution images are prevalent and runtime efficiency is critical. While our experiments are conducted up to 1024×1024, in practice, co-visible region detection can be performed on downsampled inputs (e.g., from 4K–16K aerial images) to efficiently localize overlapping areas, followed by fine-level feature matching at full resolution within the detected regions. Unlike traditional learning-based detectors that often degrade under resolution variation or increased noise, SCoDe maintains robust performance by explicitly aligning the scale space and constraining correspondence to co-visible regions. This reduces spatial and semantic ambiguity, enhances resilience to repetitive patterns, and ensures high-quality matches. These characteristics make SCoDe highly suitable for scalable 3D reconstruction and large-scale mapping workflows.

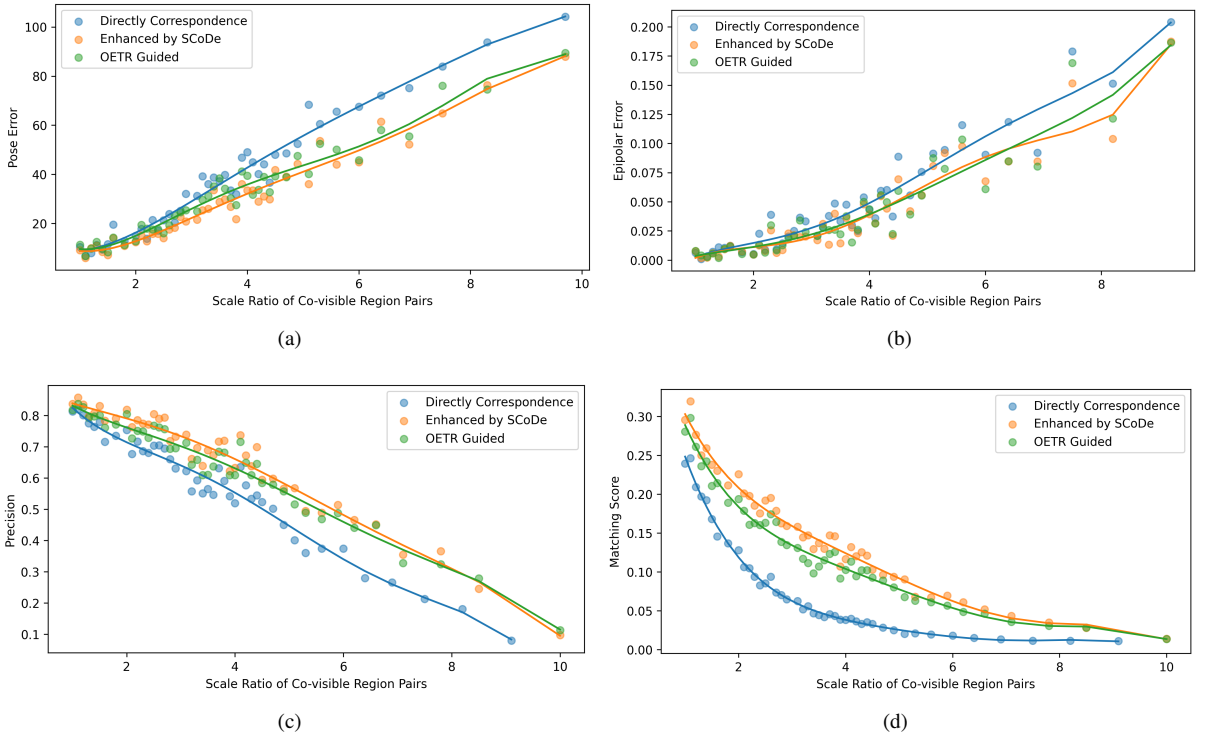


Figure 8: Visualizations of the relational trend of pose estimation and point matching effect metrics at different scales. Figures (a)-(d) show the relational trend of pose error, epipolar error, accuracy, and matching confidence at different scales, respectively. The point data represents the mean value of the metric in the center interval of the scale, and the line fits the relational trend of the metric and the scale. It should be noted that the scale ratio is the average of different test scenarios.

4.7. Evaluation under Various Scale Differences

Following the qualitative evaluation on selected samples, we now study how the performance of model changes as the scale difference varies. To ensure the generality of the metrics, we grouped image pairs by scale ratio, keeping the number of pairs consistent across groups. Figure 8 depicts the relationship between the centers of each group and mean metrics within that scale interval.

In the first row, Figure 8(a) and (b) show the trends of pose error and epipolar error as the scale ratio increases. We observe that both errors rise with larger scale differences. However, the proposed SCoDe consistently achieves lower errors across all scales, especially at higher ratios, significantly outperforming OETR. Similarly, in the second row, Figure 8(c) and (d) illustrate the decline in accuracy and confidence with increasing scale ratios. Even across a wide range of scale differences, our proposed method establishes more reliable image matching, demonstrating its superior performance.

4.8. Rotation Robustness Evaluation

To evaluate the robustness of region-constrained matching under in-plane camera rotations, we conduct a controlled synthetic experiment. We randomly sample 36 image pairs from the MegaDepth validation set and rotate one image in each pair from 0° to 180° in 2° increments, resulting in 91 rotation levels per pair.

We benchmark four representative local feature extractors: DISK (Tyszkiewicz et al., 2020), R2D2 (Revaud et al., 2019), D2-Net (Dusmanu et al., 2019), and SuperPoint (DeTone et al., 2018), as well as three rotation-invariant matchers: SIFT (Lowe, 2004), ALIKED (Zhao et al., 2023), and SE2-LoFTR (Bökman and Kahl, 2022). Each method is evaluated with and without SCoDe's region guidance. For all settings, keypoints and descriptors are extracted independently, followed by brute-force matching (Bradski, 2000) and geometric verification via RANSAC (Fischler and Bolles, 1981). We compute the ratio of correct matches at each rotation level and average results across all pairs.

As shown in Figure 9, SCoDe enhances matching robustness for rotation-sensitive methods by improving match quality within their effective rotation range (typically below 40°). The benefit is especially notable for methods without explicit orientation modeling (e.g., DISK, D2-Net), while rotation-invariant methods such as SIFT and SE2-LoFTR receive limited but non-negative gains, indicating strong compatibility. Overall, SCoDe serves as a rotation-robustness enhancer by reinforcing local geometric consistency within the rotation tolerance of the underlying detector. Under substantial in-plane rotations, co-visible region detection remains feasible but may exhibit reduced stability due to the absence of explicit rotation modeling. In such scenarios, simple augmentation strategies (e.g., discrete rotation sampling) can further mitigate performance fluctuations, suggesting a direction for extending the enhancer's applicability.

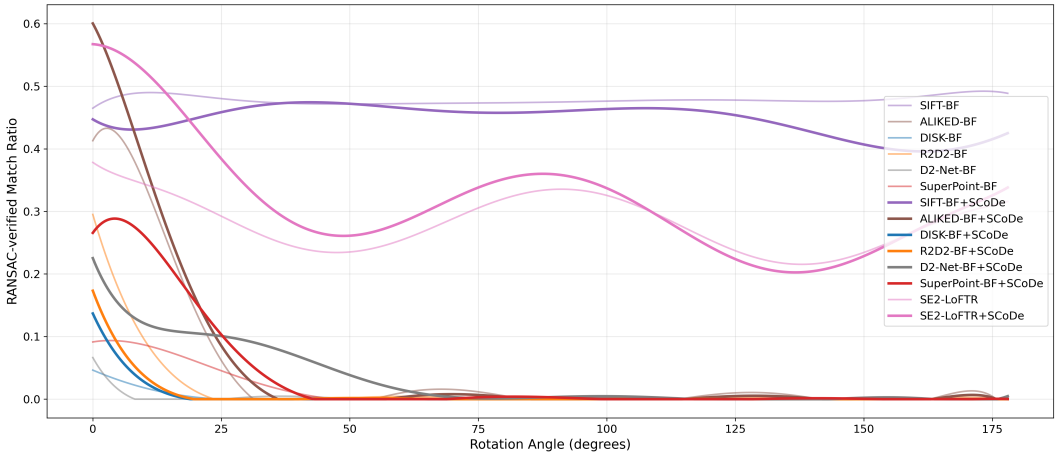


Figure 9: Rotation invariance with and without region constraints. Average correct match ratio under increasing in-plane rotation (0°–180°). Dark curves: with SCoDe region constraints. Light curves: without constraints. SCoDe improves robustness especially for methods lacking inherent rotation handling (e.g., ALIKED (Zhao et al., 2023), D2-Net (Dusmanu et al., 2019)), and does not significantly alter the performance of rotation-invariant matchers like SIFT (Lowe, 2004) and SE2-LoFTR (Bökman and Kahl, 2022).

4.9. Ablation Study

We validate the effectiveness of our proposed modules by comparing the full model with variants that have specific modules disabled. The quantitative results are presented in Table 6.

1) Scale Head Attention: Replacing all Scale Head Attention in the transformer with standard multi-head attention leads to significant drops in all the metrics except OIoU. The mIoU, reflecting the accuracy of co-visible region detection, decreases by nearly 5%, and the IoU regression rate across thresholds drops by over 7%. The scale head architecture introduces multiple scale subspaces into the attention mechanism, establishing semantic associations among them. The relatively high OIoU indicates that after removing the scale head, the model detects co-visible regions more roughly. The sharp decline in other metrics highlights the critical role of scale heads in accurately detecting co-visible regions.

2) Depth-wise Convolution: We also tried removing the depth-wise convolution in each MLP and SHA. Since this module enhances the surrounding information of each feature pixel along the depth dimension, its removal weakens spatial location awareness. The broad decline in evaluation metrics confirms the importance of depth-wise convolution for detecting co-visible regions.

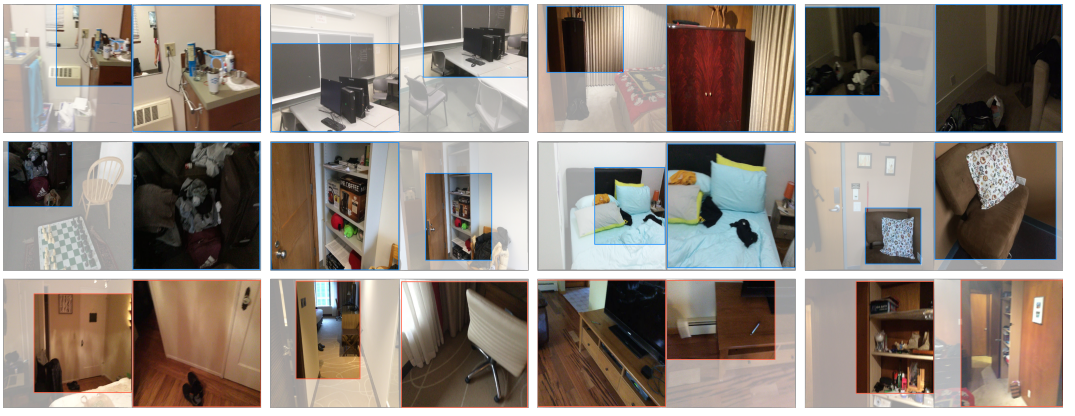
3) Confidence Prediction Module: Adding a confidence prediction module to the detection head improves the model's overall perception of co-visible regions. As Table 6 shows, introducing the additional information from confidence prediction to constrain the model's learning process improves its generalization ability. The confidence prediction module not only provides validity-checking information but also helps the detector more accurately focus on co-visible regions across varying scale differences.

4) Extra Multi-Scale Aggregation: Conventional methods (Chen et al., 2022; Wang et al., 2023b) for addressing multi-scale issues often involve augmenting multi-scale aggregation on features extracted by the backbone network.

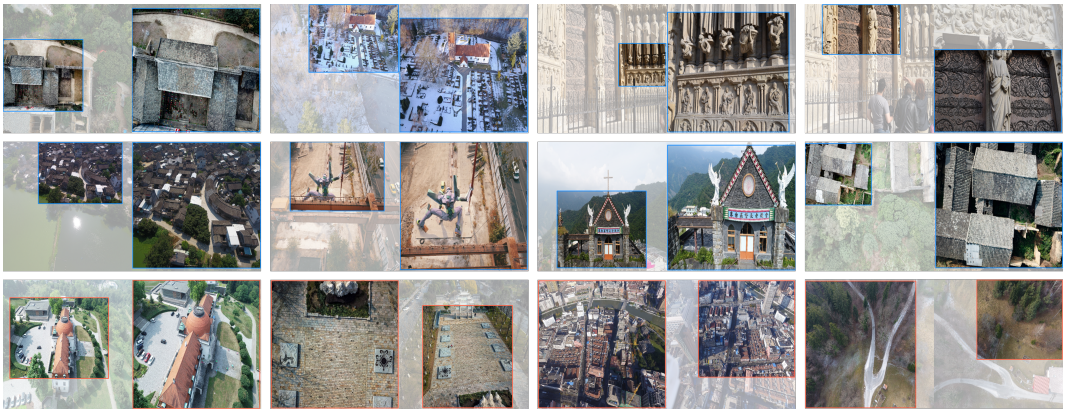
Table 6
Quantitative ablation results on the MegaDepth dataset for all crucial components of SCoDe.

Methods	mIoU	Mean of Recall		
		Conf	IoU	OIoU
w/o ScaleHead	77.29	98.69	62.42	88.61
w/o DWConv	78.41	98.74	64.03	87.19
w/o ConfModule	79.86	-	66.53	**89.18**
Extra MSA	79.30	99.10	65.51	87.40
SCoDe	**82.02**	**99.62**	**69.82**	87.66

We conducted similar experiments but found that adding extra multi-scale aggregation to the features yielded reduced performance across all indicators, as shown in Table 6. This indicates that multi-scale aggregation disrupts spatial location information, affecting the precise detection of co-visible regions.



(a) Co-visible region detection results of SCoDe on indoor scenes from ScanNet



(b) Co-visible region detection results of SCoDe on outdoor scenes from GL3D

Figure 10: Qualitative examples in different scenes. We select ScanNet and GL3D for qualitative experiments in indoor and outdoor scenes, respectively. The visualization shows the large scale co-visible area prediction in both indoor and outdoor scenes. The blue rectangular boxes in the image represent accurate detection results, while the red ones indicate inaccurate detection.

4.10. Generalizability Test

We further report co-visible region detection results under more diverse perspectives, including indoor, outdoor, and bird's-eye view. As shown in Figure 10(a), indoor scenes with large scale variations pose challenges due to repetitive repetitive textures, structural similarity, and perspective changes. SCoDe improves generalization in such environments by accurately localizing co-visible regions and suppressing ambiguous repetitions. In outdoor scenes (Figure 10(b)), which feature a wide range of scale and perspective differences, including aerial and handheld shots, SCoDe accurately identifies co-visible regions with similar semantics. This robustness is particularly valuable in drone-based photogrammetry applications, where large-scale viewpoint variation and altitude shifts are common. By integrating multi-scale features via Scale Head Attention (SHA), the model effectively addresses complex geometric variations and supports large-scale 3D reconstruction from aerial data.

The qualitative results in Figure 10 confirm the robustness of SCoDe in both indoor and outdoor scenarios, demonstrating its capability to generalize across diverse scenes with varying perspectives and scale differences. This makes it a reliable solution for real-world applications where visual conditions are often unpredictable and highly variable, particularly in geospatial domains such as drone photogrammetry, remote sensing, and urban reconstruction.

The last row of Figure 10(a) and (b) presents several failure cases, with red boxes indicating inaccurate areas. These failures are primarily caused by extreme scale differences, large viewpoint changes, or significant rotations, which lead to insufficient feature overlap or perceptual difficulty. Future research could address these challenges by incorporating stronger geometric constraints, such as epipolar geometry, or by leveraging context-aware features to improve robustness in extreme conditions.

5. Conclusion

We propose Scale-aware Co-visible Region Detector (SCoDe), a novel method that improves image matching by detecting and aligning co-visible regions before point-level correspondence. This strategy confines the matching process to spatially consistent regions that are likely to be truly visible in both views. At the core of SCoDe is the Scale Head Attention mechanism, which models scale-space dependencies across spatial hierarchies. By explicitly encoding scale information into the attention process, it enables robust region localization under large scale variations and in scenes with complex structure. By focusing the matching process on well-aligned, co-visible regions, SCoDe improves the reliability of feature associations and reduces incorrect matches caused by scale inconsistency. Extensive experiments demonstrate that SCoDe consistently outperforms existing methods in both region detection and image matching tasks. In particular, it achieves up to 8.41% improvement in matching precision across different pipelines, including SuperPoint+SuperGlue. These results show that incorporating scale-aware regional constraints provides a robust and scalable solution for image matching, with strong potential for deployment in large-scale 3D reconstruction and mapping systems.

References

Baráth, D., Noskova, J., Ivashechkin, M., Matas, J., 2020. MAGSAC++, a fast, reliable and accurate robust estimator, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1301–1309. doi:10.1109/CVPR42600.2020.00138.

Bay, H., Tuytelaars, T., Van Gool, L., 2006. Surf: Speeded up robust features, in: Proceedings of the European Conference on Computer Vision, Springer. pp. 404–417.

Bökman, G., Kahl, F., 2022. A case for using rotation invariant features in state of the art feature matchers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5110–5119.

Bradski, G., 2000. The opencv library. Dr. Dobb's Journal of Software Tools .

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., 2020. End-to-end object detection with transformers, in: Proceedings of the European Conference on Computer Vision, Springer. pp. 213–229.

Cech, J., Matas, J., Perdoch, M., 2010. Efficient sequential correspondence selection by cosegmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 32, 1568–1581.

Chen, M., Voinov, A., Ames, D.P., Kettner, A.J., Goodall, J.L., Jakeman, A.J., Barton, M.C., Harpham, Q., Cuddy, S.M., DeLuca, C., Yue, S., Wang, J., Zhang, F., Wen, Y., Lü, G., 2020. Position paper: Open web-distributed integrated geographic modelling and simulation to enable broader participation and applications. Earth-Science Reviews 207, 103223. doi:10.1016/j.earscirev.2020.103223.

Chen, Y., Huang, D., Xu, S., Liu, J., Liu, Y., 2022. Guide local feature matching by overlap estimation, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 365–373.

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/CVF Conference on Computer Vision and Pattern Recognition.

DeTone, D., Malisiewicz, T., Rabinovich, A., 2018. Superpoint: Self-supervised interest point detection and description, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 224–236.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An image is worth 16x16 words: Transformers for image recognition at scale, in: Proceedings of International Conference on Learning Representations. URL: https://openreview.net/forum?id=YicbFdNTTy.

Dusmanu, M., Rocco, I., Pajdla, T., Pollefeys, M., Sivic, J., Torii, A., Sattler, T., 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.

Edstedt, J., Sun, Q., Bökman, G., Wadenbäck, M., Felsberg, M., 2024. Roma: Robust dense feature matching, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19790–19800.

Fan, D.P., Li, T., Lin, Z., Ji, G.P., Zhang, D., Cheng, M.M., Fu, H., Shen, J., 2021. Re-thinking co-salient object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 44, 4339–4354.

Fischler, M.A., Bolles, R.C., 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM 24, 381–395.

Gutin, G., Yeo, A., Zverovich, A., 2002. Traveling salesman should not be greedy: domination analysis of greedy-type heuristics for the tsp. Discrete Applied Mathematics 117, 81–86.

Harris, C., Stephens, M., et al., 1988. A combined corner and edge detector, in: Alvey vision conference, Citeseer. pp. 10–5244.

Hartley, R.I., Sturm, P., 1997. Triangulation. Computer vision and image understanding 68, 146–157.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 770–778.

Ioli, F., Dematteis, N., Giordan, D., Nex, F., Pinto, L., 2024. Deep learning low-cost photogrammetry for 4d short-term glacier dynamics monitoring. PFG–Journal of Photogrammetry, Remote Sensing and Geoinformation Science , 1–22.

Jiang, W., Trulls, E., Hosang, J., Tagliasacchi, A., Yi, K.M., 2021. Cotr: Correspondence transformer for matching across images, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6207–6217.

Jin, Y., Mishkin, D., Mishchuk, A., Matas, J., Fua, P., Yi, K.M., Trulls, E., 2021. Image matching across wide baselines: From paper to practice. International Journal of Computer Vision 129, 517–547.

Kong, Y., Zheng, Y., Yao, C., Liu, Y., Wang, H., 2022. Scale adaptive fusion network for rgb-d salient object detection, in: Proceedings of the Asian Conference on Computer Vision, pp. 3620–3636.

Li, H., Zheng, X., Dong, M., Xia, G.S., Xiong, H., 2021a. Locally nonlinear affine verification for multisensor image matching. IEEE Transactions on Geoscience and Remote Sensing 60, 1–16.

Li, L., Han, L., Ding, M., Cao, H., Hu, H., 2021b. A deep learning semantic template matching framework for remote sensing image registration. ISPRS Journal of Photogrammetry and Remote Sensing 181, 205–217. doi:10.1016/j.isprsjprs.2021.09.012.

Li, Z., Snavely, N., 2018. Megadepth: Learning single-view depth prediction from internet photos, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Lin, W., Wu, Z., Chen, J., Huang, J., Jin, L., 2023. Scale-aware modulation meet transformer, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6015–6026.

Lindenberger, P., Sarlin, P.E., Pollefeys, M., 2023. Lightglue: Local feature matching at light speed, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 17627–17638.

Loshchilov, I., Hutter, F., 2017. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 .

Lowe, D.G., 1999. Object recognition from local scale-invariant features, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, Ieee. pp. 1150–1157.

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

Luo, Z., Shen, T., Zhou, L., Zhang, J., Yao, Y., Li, S., Fang, T., Quan, L., 2019. Contextdesc: Local descriptor augmentation with cross-modality context, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2527–2536.

Luo, Z., Shen, T., Zhou, L., Zhu, S., Zhang, R., Yao, Y., Fang, T., Quan, L., 2018. Geodesc: Learning local descriptors by integrating geometry constraints, in: Proceedings of the European Conference on Computer Vision.

Luo, Z., Zhou, L., Bai, X., Chen, H., Zhang, J., Yao, Y., Li, S., Fang, T., Quan, L., 2020. Aslfeat: Learning local features of accurate shape and localization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6589–6598.

Mei, Y., Fan, Y., Zhang, Y., Yu, J., Zhou, Y., Liu, D., Fu, Y., Huang, T.S., Shi, H., 2023. Pyramid attention network for image restoration. International Journal of Computer Vision 131, 3207–3225.

Peyré, G., Cuturi, M., et al., 2019. Computational optimal transport: With applications to data science. Foundations and Trends® in Machine Learning 11, 355–607.

Rau, A., Garcia-Hernando, G., Stoyanov, D., Brostow, G.J., Turmukhambetov, D., 2020. Predicting visual overlap of images through interpretable non-metric box embeddings, in: Proceedings of the European Conference on Computer Vision, Springer. pp. 629–646.

Ren, S., Zhou, D., He, S., Feng, J., Wang, X., 2022. Shunted self-attention via multi-scale token aggregation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10853–10862.

Revaud, J., De Souza, C., Humenberger, M., Weinzaepfel, P., 2019. R2d2: Reliable and repeatable detector and descriptor. Advances in Neural Information Processing Systems 32.

Rocco, I., Arandjelović, R., Sivic, J., 2020. Efficient neighbourhood consensus networks via submanifold sparse convolutions, in: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (Eds.), Proceedings of the European Conference on Computer Vision, Springer International Publishing, Cham. pp. 605–621. doi:10.1007/978-3-030-58545-7_35.

Rosten, E., Drummond, T., 2006. Machine learning for high-speed corner detection, in: Proceedings of the European Conference on Computer Vision, Springer. pp. 430–443.

Rublee, E., Rabaud, V., Konolige, K., Bradski, G., 2011. Orb: An efficient alternative to sift or surf, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, Ieee. pp. 2564–2571.

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.

Scharstein, D., Szeliski, R., 2002. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. International Journal of Computer Vision 47, 7–42.

Schönberger, J.L., Frahm, J.M., 2016. Structure-from-motion revisited, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Schönberger, J.L., Zheng, E., Pollefeys, M., Frahm, J.M., 2016. Pixelwise view selection for unstructured multi-view stereo, in: Proceedings of the European Conference on Computer Vision.

Shen, S., 2013. Accurate multiple view 3D reconstruction using patch-based stereo for large-scale scenes. IEEE Transactions on Image Processing 22, 1901–1914. doi:10.1109/TIP.2013.2237921.

Shen, T., Luo, Z., Zhou, L., Zhang, R., Zhu, S., Fang, T., Quan, L., 2018. Matchable image retrieval by learning from surface reconstruction, in: Proceedings of the Asian Conference on Computer Vision.

Sivic, Zisserman, 2003. Video google: a text retrieval approach to object matching in videos, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1470–1477. doi:10.1109/ICCV.2003.1238663.

Song, H., Kashiwaba, Y., Wu, S., Wang, C., 2023. Efficient and accurate co-visible region localization with matching key-points crop (mkpc): A two-stage pipeline for enhancing image matching performance. arXiv preprint arXiv:2303.13794 .

Song, S., Morelli, L., Wu, X., Qin, R., Albanwan, H., Remondino, F., 2024. Evaluating learning-based tie point matching for geometric processing of off-track satellite stereo. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 48, 393–400.

Sun, J., Ji, L., Zhu, J., 2023. Shared coupling-bridge scheme for weakly supervised local feature learning. IEEE Transactions on Multimedia .

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.

Tang, S., Zhang, Y., Li, Y., Yuan, Z., Wang, Y., Zhang, X., Li, X., Zhang, Y., Guo, R., Wang, W., 2019. Fast and automatic reconstruction of semantically rich 3D indoor maps from low-quality RGB-d sequences. Sensors 19. doi:10.3390/s19030533.

Torres, J.C., López, L., Romo, C., Arroyo, G., Cano, P., Lamolda, F., Villafranca, M.M., 2013. Using a cultural heritage information system for the documentation of the restoration process, in: 2013 Digital Heritage International Congress (DigitalHeritage), pp. 249–256. doi:10.1109/DigitalHeritage.2013.6744761.

Tu, D., Cui, H., Shen, S., 2023. Panovlm: Low-cost and accurate panoramic vision and lidar fused mapping. ISPRS Journal of Photogrammetry and Remote Sensing 206, 149–167. URL: https://www.sciencedirect.com/science/article/pii/S0924271623003179, doi:https://doi.org/10.1016/j.isprsjprs.2023.11.012.

Tyszkiewicz, M., Fua, P., Trulls, E., 2020. Disk: Learning local features with policy gradient. Advances in Neural Information Processing Systems 33, 14254–14265.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Advances in Neural Information Processing Systems 30.

Wang, C., Xu, L., Xu, R., Xu, S., Meng, W., Wang, R., Zhang, X., 2023a. Triple robustness augmentation local features for multi-source image registration. ISPRS Journal of Photogrammetry and Remote Sensing 199, 1–14. doi:10.1016/j.isprsjprs.2023.03.023.

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.

Wang, W., Chen, W., Qiu, Q., Chen, L., Wu, B., Lin, B., He, X., Liu, W., 2023b. Crossformer++: A versatile vision transformer hinging on cross-scale attention. IEEE Transactions on Pattern Analysis and Machine Intelligence .

Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L., 2021. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 568–578.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., Philip, S.Y., 2020. A comprehensive survey on graph neural networks. IEEE Transactions on Neural Networks and Learning Systems 32, 4–24.

Xie, T., Dai, K., Wang, K., Li, R., Zhao, L., 2024. Deepmatcher: a deep transformer-based network for robust and accurate local feature matching. Expert Systems with Applications 237, 121361.

Xu, H., Zhang, J., 2020. Aanet: Adaptive aggregation network for efficient stereo matching, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1959–1968.

Xu, S., Chen, S., Xu, R., Wang, C., Lu, P., Guo, L., 2024. Local feature matching using deep learning: A survey. Information Fusion 107, 102344.

Ye, Y., Zhu, B., Tang, T., Yang, C., Xu, Q., Zhang, G., 2022. A robust multimodal remote sensing image registration method and system using steerable filters with first- and second-order gradients. ISPRS Journal of Photogrammetry and Remote Sensing 188, 331–350. URL: https://www.sciencedirect.com/science/article/pii/S0924271622001083, doi:https://doi.org/10.1016/j.isprsjprs.2022.04.011.

Yi, K.M., Trulls, E., Lepetit, V., Fua, P., 2016. Lift: Learned invariant feature transform, in: Proceedings of the European Conference on Computer Vision, Springer. pp. 467–483.

Yi, K.M., Trulls, E., Ono, Y., Lepetit, V., Salzmann, M., Fua, P., 2018. Learning to find good correspondences, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2666–2674.

Zhang, J., Sun, D., Luo, Z., Yao, A., Zhou, L., Shen, T., Chen, Y., Quan, L., Liao, H., 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.

Zhang, J., Xia, Z., Dong, M., Shen, S., Yue, L., Zheng, X., 2025. Comatcher: Multi-view collaborative feature matching, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 21970–21980.

Zhang, X., Zhou, Y., Qiao, P., Lv, X., Li, J., Du, T., Cai, Y., 2023. Image registration algorithm for remote sensing images based on pixel location information. Remote Sensing 15, 436. doi:10.3390/rs15020436.

Zhao, X., Wu, X., Chen, W., Chen, P.C., Xu, Q., Li, Z., 2023. Aliked: A lighter keypoint and descriptor extraction network via deformable transformation. IEEE Transactions on Instrumentation and Measurement 72, 1–16.

Appendix A. Detailed Model Architecture

Table A.1: Main Architecture Components

Module	Submodule	Input	Output	Description
Backbone (layer0-3)	ResNet-50	(B,3,H,W)	(B,1024,H/16,W/16)	Feature extraction
Input Projection	CNN	(B,1024,H/16,W/16)	(B,256,H/32,W/32)	Dim. reduction
Transformer	Encoder+Decoder	(B,256,H/32,W/32)	(B,2,256)	Cross-attn process
TLBR Regression	MLP	(B,2,256)	(B,2,4)	Box regression
Heatmap	CNN	(B,256,H/32,W/32)	(B,1,H/32,W/32)	Heatmap generation
Confidence Prediction	CNN+MLP	(B,1,H/32,W/32)	(B,2)	Overlap classification

Table A.2: Transformer ($\mathcal{T}$ & $\mathcal{D}$) Flow

Step	Operation	Input Shape	Output Shape	Description
1	Feature Reshape	(B, 256, H/16, W/32)	(B, HW/512, 256)	Flatten spatial dimensions
2	Learnable Query	-	(B, 2, 256)	Learnable query embeddings
3	Encoder $\mathcal{T}$	(B, HW/512, 256)	(B, HW/512, 256)	Multi-stage feature processing
4	Decoder $\mathcal{D}$	(B, 2, 256) + (B, HW/512, 256)	(B, 2, 256)	Cross-attention decoding

Table A.3: Encoder $\mathcal{T}$ Detailed Architecture

Stage	Block	Input	Output	Conv Params	MLP
0	0	(B,256,H/32,W/32)	same	1×1→256, 3×3→128, 5×5→128	256→1024→256
0	1	(B,256,H/32,W/32)	same	1×1→256, 3×3→128, 5×5→128	256→1024→256
1	0	(B,256,H/32,W/32)	same	1×1→256, 3×3→128, 5×5→128	256→1024→256
1	1	(B,256,H/32,W/32)	same	1×1→256, 3×3→128, 5×5→128	256→1024→256
2	0	(B,256,H/32,W/32)	same	1×1→256, 3×3→128, 5×5→128	256→1024→256
2	1	(B,256,H/32,W/32)	same	1×1→256, 3×3→128, 5×5→128	256→1024→256
3	0	(B,256,H/32,W/32)	same	1×1→256, 3×3→128, 5×5→128	256→1024→256
3	1	(B,256,H/32,W/32)	same	1×1→256, 3×3→128, 5×5→128	256→1024→256

Table A.4: SHA Multi-Scale Processing $\mathcal{C}$

Path	Conv (stride)	Input	Output	Function
Path 0	1×1, 256 (1)	(B, 256, H/32, W/32)	(B, 256, H/32, W/32)	Original scale
Path 1	3×3, 128 (3)	(B, 256, H/32, W/32)	(B, 128, H/96, W/96)	Coarse features
Path 2	5×5, 128 (5)	(B, 256, H/32, W/32)	(B, 128, H/160, W/160)	Global context

Table A.5: Decoder $\mathcal{D}$ Detailed Architecture

Layer	Component	Input	Output	Attention	Params
0	Self-Attn	(B,2,256)	(B,2,256)	Linear	-
0	Cross-Attn	(B,2,256)+(B,HW/512,256)	(B,2,256)	SHA	1×1→256, 3×3→128, 5×5→128
0	MLP	(B,2,256)	(B,2,256)	-	256→1024→256
1	Self-Attn	(B,2,256)	(B,2,256)	Linear	-
1	Cross-Attn	(B,2,256)+(B,HW/512,256)	(B,2,256)	SHA	1×1→256, 3×3→128, 5×5→128
1	MLP	(B,2,256)	(B,2,256)	-	256→1024→256

Table A.6: Head Modules $\mathcal{H}$ Specifications

Head	Layer	Input	Output	Params	Act	#Params
$\mathcal{H}_2$ TLBR	Linear 1	(B,2,256)	(B,2,256)	256→256	ReLU	65,536
$\mathcal{H}_2$ TLBR	Linear 2	(B,2,256)	(B,2,4)	256→4	None	1,028
$\mathcal{H}_2$ Heatmap	Conv2d 1	(B,256,H/32,W/32)	same	3×3, out=256	ReLU	590,080
$\mathcal{H}_2$ Heatmap	Conv2d 2	(B,256,H/32,W/32)	(B,1,H/32,W/32)	1×1, out=1	None	257
$\mathcal{H}_1$ CNN $\mathcal{G}$	Conv2d 1	(B,1,H/32,W/32)	(B,16,H/32,W/32)	3×3, out=16	ReLU	160
$\mathcal{H}_1$ CNN $\mathcal{G}$	MaxPool 1	(B,16,H/32,W/32)	(B,16,H/64,W/64)	2×2, s=2	None	0
$\mathcal{H}_1$ CNN $\mathcal{G}$	Conv2d 2	(B,16,H/64,W/64)	(B,32,H/64,W/64)	3×3, out=32	ReLU	4,640
$\mathcal{H}_1$ CNN $\mathcal{G}$	MaxPool 2	(B,32,H/64,W/64)	(B,32,H/128,W/128)	2×2, s=2	None	0
$\mathcal{H}_1$ MLP	Flatten	(B,32,H/128,W/128)	(B,32×HW/16384)	-	None	0
$\mathcal{H}_1$ MLP	Linear $\mathcal{L}_1$	(B,32×HW/16384)	(B,256)	→256	ReLU	Var.
$\mathcal{H}_1$ MLP	Linear $\mathcal{L}_2$	(B,256)	(B,2)	256→2	Softmax	514

Table A.7: Detailed Convolution Layer Parameters

Module	Layer	Kernel Size	Stride	Padding	Groups	Input Ch.	Output Ch.
Input Projection	Conv1x1	1×1	1	0	1	1024	256
Input Projection 2	Conv1x1	1×1	2	0	1	256	256
$\mathcal{H}_2$ Heatmap Conv	Conv1	3×3	1	1	1	256	256
$\mathcal{H}_2$ Heatmap Conv	Conv2	1×1	1	0	1	256	1
$\mathcal{H}_1$ CNN	Conv1	3×3	1	1	1	1	16
$\mathcal{H}_1$ CNN	Conv2	3×3	1	1	1	16	32
SHAttention MSA	Reduction0	1×1	1	0	1	256	256
SHAttention MSA	Reduction1	3×3	3	0	1	256	128
SHAttention MSA	Reduction2	5×5	5	0	1	256	128
DWConv (SHA)	DWConv	3×3	1	1	128	128	128
DWConv (MLP)	DWConv	3×3	1	1	1024	1024	1024