UDC 004.93: 004.8

HYBRID TRANSFORMER-GRAPH NEURAL NETWORK FEATURE MATCHING BASED METHODOLOGY FOR ROBUST TEMPLATE OBJECT LOCALIZATION

Bach N.V.

Belarusian State University of Informatics and Radioelectronics, Minsk, Republic of Belarus

Tsviatkou V.Yu. – Doctor of Technical Sciences, Professor

Annotation. This paper proposes a novel framework that combines transformer-based and graph neural network-based feature matching techniques for accurate template object localization. The proposed pipeline consists of four main components: a hybrid features matching module, a non-linear geometric transformation module, and a bounding box refinement module. By integrating the strengths of both global contextual understanding from transformers and structural relationship modeling from graph neural networks, the method achieves improved robustness and precision in detecting and localizing objects under challenging conditions.

Keywords. Graph neural network, feature matching, template object localization, hybrid neural network.

Introduction. Template object localization is fundamental in applications ranging from industrial inspection and autonomous navigation to augmented reality. Traditional keypoints methods for template object localization such as scale-invariant feature transform (SIFT) or oriented FAST and rotated BRIEF (ORB) often struggle with large viewpoint changes, non-rigid deformations, and varying illumination. To overcome these challenges, it is proposed a hybrid strategy that intelligently combines the complementary strengths of efficient local feature matching (ELoFTR) with transformer [1, 2] and SuperPoint–SuperGlue with graph neural network (GNN) [3, 4] for creating a more robust and versatile feature matching system that can adapt to different image characteristics and application requirements and improve the accuracy and efficiency of finding correspondences between features in images. This approach integrates the powerful representational capabilities of transformers with the structural awareness of GNNs to create robust and adaptable matching models. Transformer owing to using self-attention and global view mechanisms is effective and flexible at capturing long-range dependencies and contextual information within data sequences or feature sets, understanding of relationships between features and adapting to various matching tasks by adjusting their architecture and training objectives. GNN due to using structure awareness mechanism and local neighborhood aggregation operates on graph-structured data, effectively capturing structural relationships and dependencies between features based on their connections and aggregates information from neighboring nodes, allowing them learn local feature representations and identify clusters.

This project presents a novel pipeline that integrates transformer-based and graph-based feature matching to accurately localize objects in a source image given a template. To achieve robust and precise object localization, our framework intelligently combines the hybrid features matching module, non-linear geometric transformation module and bounding box processing and analysis module.

The hybrid features matching module can integrate multi feature matching modules, but in project we use two submodels: transformer-based matcher (ELoFTR) and GNN-based matcher (SuperPoint–SuperGlue) (Figure 1). Integration of the ELoFTR and SuperPoint–SuperGlue fearture matching module to generate good keypoints is based on computing confidence score that estimates how likely the module outputs are correct.

Input Stage. The template image (visual prompt) annotated with an initial bounding box in the XYXY format, defining the object's spatial extent. The template and the source images are a resolution frame (460×512) or (1920×1080) with bit depth varying between 8 and 24 bits. The template and the source images to be recognized may exhibit multi-dimensional differences, such as varying spectral bands, platforms, sources, resolutions, or perspectives.

Hybrid Feature Matching Module. At the core of our pipeline, we combine a transformer-based matcher (ELoFTR) with a graph-based matcher (SuperPoint + SuperGlue) to harvest both dense, globally consistent correspondences and sparse, locally refined keypoints. ELoFTR first divides the template and source images into fixed-size patches, embeds each via a linear projection, and injects learned positional encodings. A coarse cross-attention stage then establishes rough matches, which a fine-grained refinement head sharpens to sub-pixel accuracy—each correspondence tagged with a confidence score. In parallel, SuperPoint detects salient 2D keypoints and computes 256-dimensional descriptors; SuperGlue builds a bipartite graph over template and source points, performs iterative cross-graph attention, and solves an optimal-transport problem to yield robust match pairs.

Non-linear geometric transformation module. In this module, we first prune and concatenate the keypoints correspondences generated by ELoFTR and SuperGlue. The retained matches are then used to robustly estimate a projective homography *H* through RANSAC algorithm, which automatically rejects

outliers. Finally, the four corners of the template's object bounding box are warped through *H* to yield a provisional object region in the source image, ready for downstream contour-based refinement.

Bounding-box refinement module. In the final stage, we refine the provisional bounding box by leveraging local edge and contour information to achieve precise object delineation. First, we extract a crop of the source image defined by the warped corners of the provisional box. This region is converted to grayscale and subjected to adaptive Gaussian thresholding, which dynamically segments foreground and background based on local intensity variations. From the resulting binary mask, we detect all connected contours and evaluate each by computing its area overlap with the provisional box. The contour exhibiting the maximum overlap is assumed to correspond to the target object's silhouette. Finally, we fit a minimum-area enclosing rectangle around this contour, and convert its corner coordinates into the standard XYXY format to produce the final, refined bounding box.

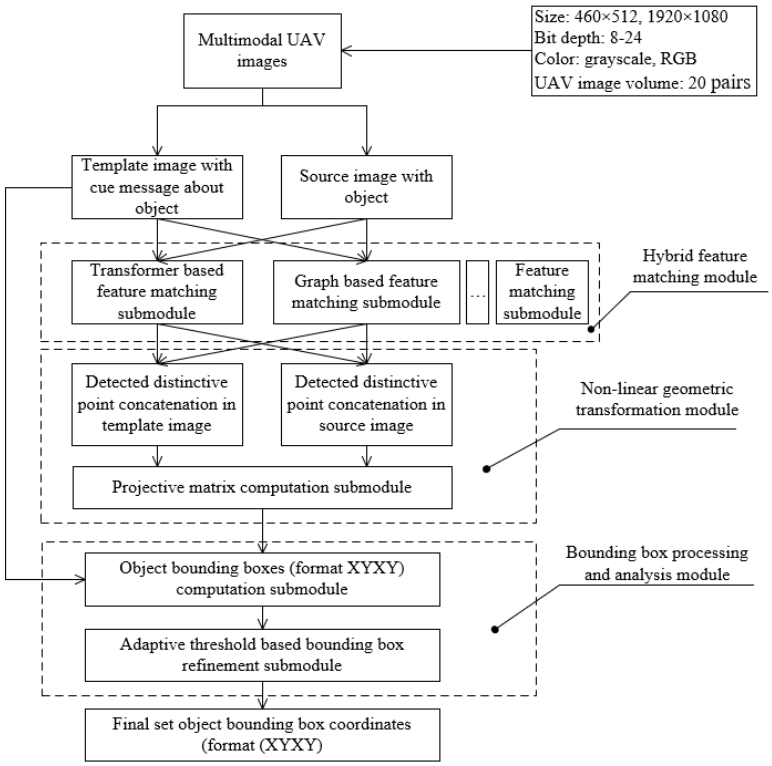


Figure 1 – Integrated modular feature matching structure

Estimation of the hybrid features matching module performance. Table 1 summarizes the average IoU-based score values (1) providing by using the ELoFTR and SuperPoint+SuperGlue matcher submodules and hybrid feature matching module on our test sets of 20 image pairs.

$$scorevalue = \begin{cases} 1 + (IoU - 0.75) \cdot 4, 0.5 < IoU \leq 1.0 \\ 0, 0.3 \leq IoU \leq 0.5 \\ -0.5, IoU < 0.3 \end{cases} \quad (1)$$

Table 1 – Estimation of template object detection score values for submodules and hybrid module

Type of submodule	Score value
ELoFTR submodule	12.39
SuperPoint+SuperGlue submodule	5.6
Hybrid Feature Matching Module	15.6

It follows from Table 1 provides the score value of template object detection is increased from 3.21 to 10 score value with using the hybrid feature matching module for the given multimodal UAV images.

Figure 2 and Figure 3 present comparative results of different feature matching approaches under various image conditions. It can be observed that the hybrid feature matching module consistently achieves the most accurate object localization, with bounding boxes closely aligned to the target objects. In contrast, the SuperPoint+SuperGlue method shows lower precision due to limited global context, while ELoFTR provides better global matching but lacks fine localization accuracy. Overall, the results demonstrate that the hybrid approach effectively combines the strengths of both methods, leading to improved robustness and detection performance across different scenarios.

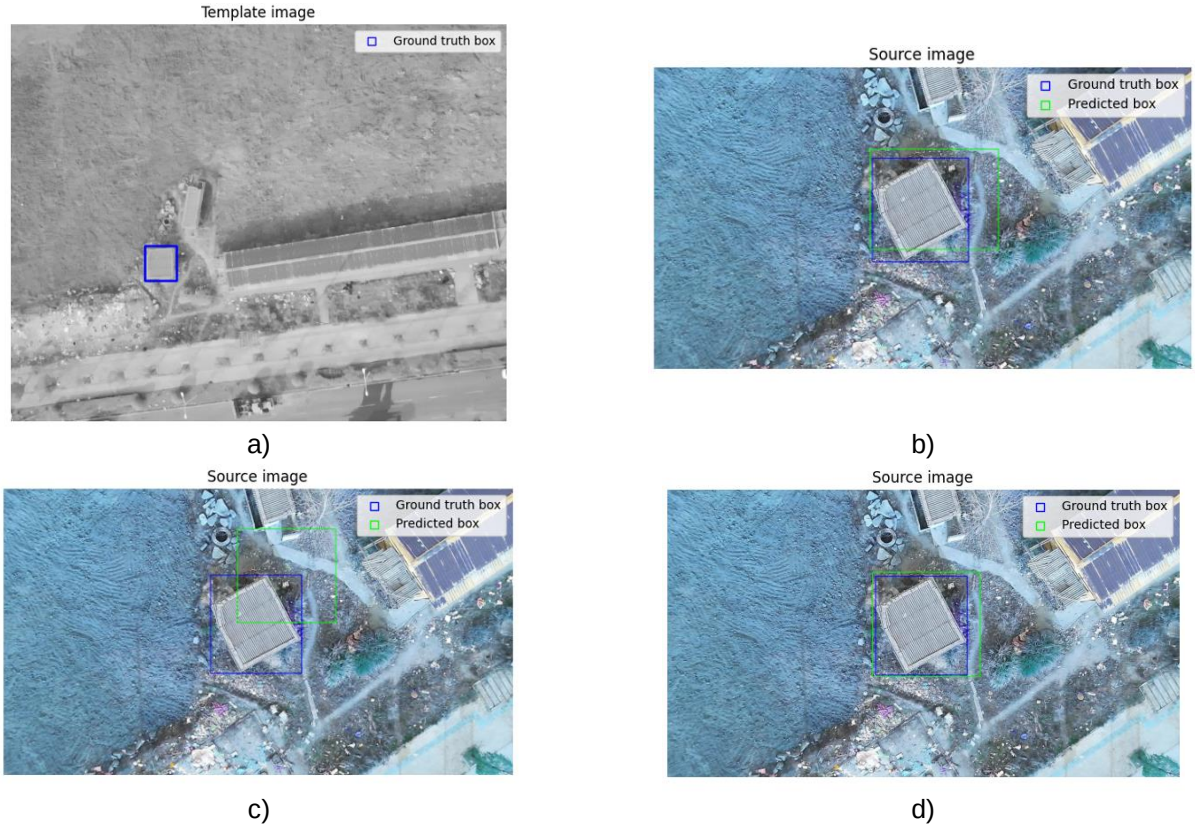


Figure 2 – Grayscale template (a) and color source images with detected object bounding box using hybrid feature matching module (b), only SuperPoint+SuperGlue submodule (c), only ELoFTR submodule (d)

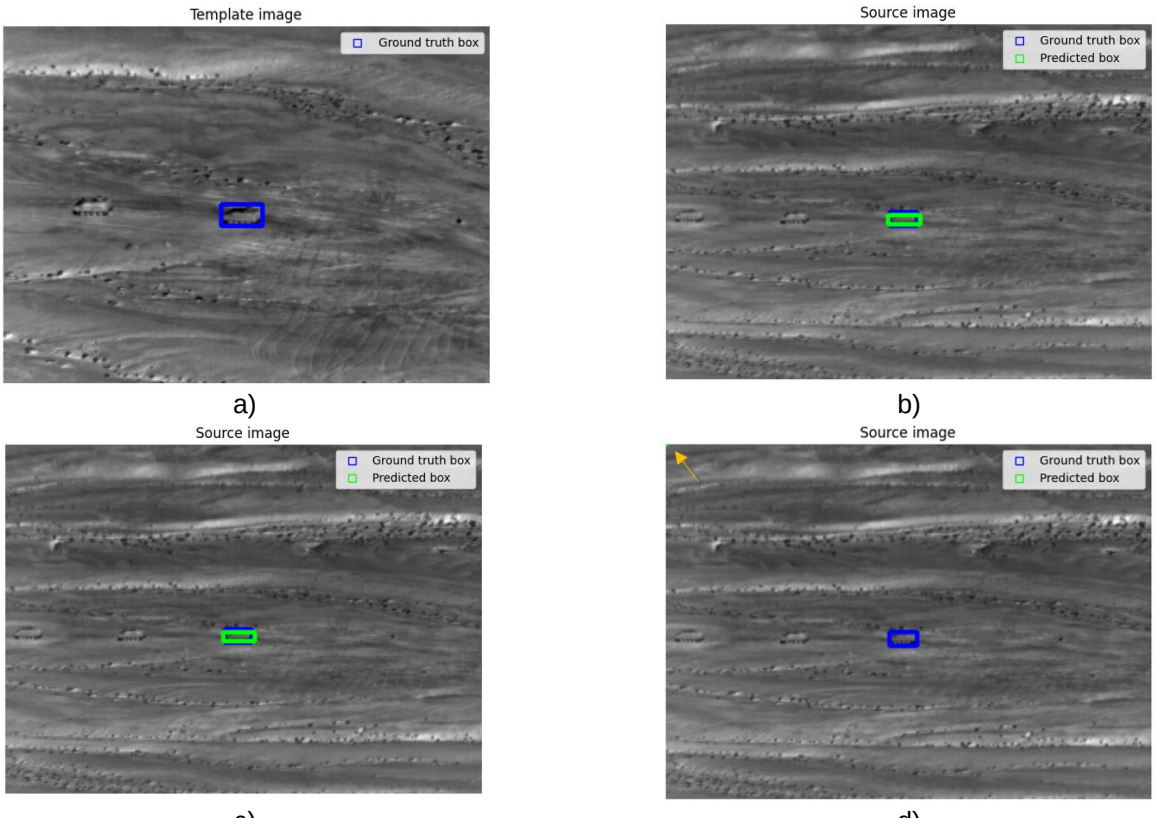
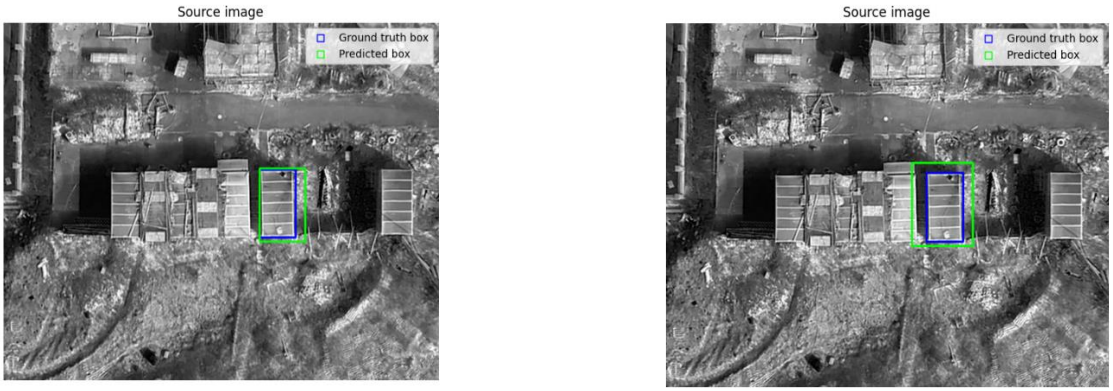


Figure 3 – Grayscale template (a) and source images with detected object bounding box using hybrid feature matching module (b), only SuperPoint+SuperGlue submodule (c), only ELoFTR submodule (d)

Visualization of bounding-box refinement module efficiency. The Figure 4 contains detected object image with (a) and without (b) using bounding box processing and analysis module.



a) b)

Figure 4 – Detected object bounding box with (a) and without (b) bounding box processing and analysis module

It follows from Figure 4 that bounding box processing and analysis module improves template object detection accuracy for the given multimodal UAV images.

Conclusion. In this paper, we proposed a hybrid object localization framework that integrates transformer-based and graph neural network-based feature matching with robust geometric transformation and contour-based bounding box refinement. By combining global contextual modeling and structural relationship learning, the proposed approach achieves high localization accuracy and robustness under challenging conditions such as scale variation and multimodal differences. Experimental results demonstrate that the hybrid feature matching module significantly outperforms individual methods, confirming the effectiveness of the overall pipeline.

References:

1. He, X. Matchanything: Universal cross-modality image matching with large-scale pre-training / H. Yu, S. Peng, D. Tan, Z. Shen, H. Bao, X. Zhou // arXiv preprint arXiv, 2025. – P. 2501.07556.

2. Wang, Y. Efficient LoFTR: Semi-dense local feature matching with sparse-like speed / X. He, S. Peng, D. Tan, and X. Zhou // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2024. – P. 21666-21675.

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

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