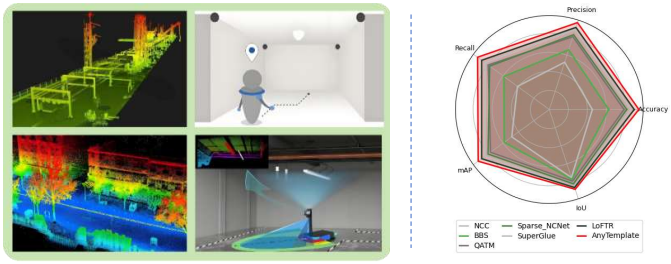
AnyTemplate: A Zero-shot Template Matching Approach for Arbitrary-Resolution and Scale-Agnostic Images

Dongjie Chen[1], Zhiwei Lv[1], Bowen Liu[2], Chengjie Fang[1], Jungang Xu[1*]
[1] University of Chinese Academy of Sciences, Beijing, China
[2] City University of Hong Kong, Hong Kong, China
chendongjie14@mails.ucas.ac.cn, *xujg@ucas.ac.cn

***Abstract*—Zero-shot Template Matching is challenging due to scale variations, aspect ratio distortions and background clutter. While foundation models like SAM3 and DINOv3 provide powerful features, cascading them often results in suboptimal performance due to geometric distortion and shallow feature interactions. To address this, we propose AnyTemplate, a novel framework unifying foundation models with deep feature interaction and geometry-aware refinement. Firstly, we introduce Aspect-Ratio-Preserving Normalization to maintain structural integrity and a learnable Cross-Attention Interaction Module to dynamically model semantic dependencies and suppress noise. Secondly, we propose a Multi-task Decoding Head capable of simultaneously predicting confidence and refining localization. Thirdly, we present DEAMPB dataset, a new multi-scale template matching benchmark. Extensive experiments show that AnyTemplate achieves SOTA performance on DEAMPB and standard benchmarks, significantly outperforming existing models in zero-shot scenarios.**

***Index Terms*—Zero-shot Template Matching, Visual Foundation Models, Cross-Attention, Deep Feature Interaction.**



(a) Applications of Template Matching (b) Metrics Comparison

Fig. 1. (a) shows the common application fields of Template Matching including point cloud registration, navigation, and visual localization; (b) shows the Accuracy/Precision/Recall/mAP/IoU metrics performance of the mainstream methods. By handling inconsistent resolutions and scales from various acquisition devices, AnyTemplate enables template matching with inputs of any scale or resolution.

I. Introduction

Template Matching is a fundamental and practical matching method in computer vision and is widely applied in many fields, such as video tracking, object detection, image stitching and 3D reconstruction. Traditional Template Matching methods often struggle with scale, orientation, and illumination variations. Unlike traditional methods that rely on handcrafted feature extraction, deep learning has revolutionized computer vision by enabling automatic feature extraction. Using deep features from deep neural networks, the focus of Template Matching has shifted from traditional methods to deep learning-based ones. Chen et al. proposed a template matching method called QATM [1], which can be used as a standalone template matching algorithm and can be easily embedded into a DNN model. Talmi et al. proposed a template matching measure named Deformable Diversity Similarity [2], which is robust to complex deformations, significant background clutter, and occlusions. Talmi et al. proved that their method outperforms the previous state of the art in detection accuracy while reducing computational complexity. Lai et al. proposed a fast and robust template matching method called A-MNS [3], which is based on Majority Neighbour Similarity (MNS) and the annulus projection transformation. Lai et al. demonstrate that A-MNS is able to estimate the rotation angle of the target object, overcome challenges such as background clutter, occlusion, arbitrary rotation transformation, and non-rigid deformation, while performing fast matching. Gao et al. proposed a template matching method based on a differentiable refinement of coarse-to-fine correspondence [4]. They used an edge-aware module to bridge the domain gap between the mask template and the grayscale image, thereby enabling robust matching.

We propose AnyTemplate, leveraging visual foundation models for generalization across arbitrary-resolution, scale-agnostic images. We utilize SAM3 [5] to generate region proposals and DINOv3 [6] to extract feature vectors from templates and candidates. Specifically, we introduce an Aspect-Ratio-Preserving Normalization strategy to maintain the structural integrity of arbitrary-shaped objects during feature embedding. Unlike previous methods based on shallow metrics, we introduce a learnable Cross-Attention Interaction Module to facilitate dynamic context exchange. Furthermore, Multi-task Decoding Head simultaneously predicts scores and refines

* is the corresponding author.

bounding boxes for high-precision localization for extreme aspect ratio objects. The main contributions are as follows:

- We propose AnyTemplate, a novel zero-shot framework that bridges the gap between generic foundation models and precise template matching via deep feature interaction.
- We introduce a geometry-aware pipeline featuring Cross-Attention Interaction Module and a Multi-task Decoding Head to tackle geometric distortion and localization coarse-ness.
- We establish a comprehensive benchmark for zero-shot template matching by introducing the DEAMPB dataset and validating our method across multiple domains.

II. Related Work

A. Template Matching

Existing Template Matching methodologies can be broadly categorized into traditional and deep learning-based methods. Traditional approaches typically measure similarity through pixel-level statistics or dense structural feature comparisons [7], such as Normalized Cross Correlation (NCC), Sum of Squared Difference (SSD) [8], Mutual Information (MI) [9], and Local Self-Similarity (LSS) [10]. While effective in constrained environments, the high computational cost of dense feature extraction often limits their early applications. In the deep learning era, template matching methods have evolved to employ Deep Neural Networks (DNNs) for feature extraction and similarity evaluation, as seen in QATM, BUPM [11], and Deep-DIM [12]. By automatically learning high-level semantic representations, these algorithms achieve superior performance compared to handcrafted features. Interestingly, the requirement for precise feature alignment in matching is also a core challenge in complex generation tasks; for instance, unified conditional frameworks for pose-guided person generation [13] leverage similar structural alignment principles to maintain consistency between target poses and source features.

B. DNN-based Image Classifier

The rapid evolution of Deep Neural Networks has empowered image classifiers with formidable feature representation capabilities. AlexNet [14] marked a breakthrough by winning the 2012 ImageNet challenge, significantly outperforming traditional methods with a top-5 error rate of 15.3%. Subsequent architectures like ResNet [15] introduced residual learning to facilitate the training of substantially deeper networks, while EfficientNet [16] utilized neural architecture search to optimize the balance between accuracy and computational efficiency. More recently, the Vision Transformer (ViT) [17] shifted the paradigm by applying self-attention mechanisms directly to image patch sequences, achieving state-of-the-art results with reduced training resources. Beyond classification, these powerful backbones have been extended to more complex generative scenarios. For example, motion-prior conditional diffusion models for long-term talking face generation [18] build upon these robust feature representations to ensure temporal stability and realistic motion synthesis across extended sequences. Furthermore, self-supervised encoders like DINOv2 [19] have bridged the gap between supervised and unsupervised feature learning, providing versatile backbones for a variety of downstream vision applications.

C. Image Similarity

Image similarity measurement is critical for various applications, including face recognition, object tracking, and person re-identification [20]. Traditional techniques often rely on intensity histograms, cosine similarity, or hashing algorithms (AHash, PHash, DHash) to compare images via binary codes or vector distances. Feature-based methods such as SIFT and ORB further quantify keypoints to calculate geometric distances between images. In contrast, Deep Learning-based approaches, particularly Siamese neural networks [21], utilize weight-sharing architectures to learn discriminative embeddings for direct comparison. Modern similarity research has shifted toward fine-grained feature interaction and controllable matching. This trend is evident in domains like fashion design, where fine-grained garment generation systems [22] require understanding local texture similarity and global style consistency. Similarly, customizable virtual dressing frameworks [23] emphasize the importance of maintaining high-fidelity similarity between the original garment and the generated result to ensure practical utility in controllable fashion design. These advancements demonstrate that image similarity is no longer determined solely by simple distance metrics, but rather by the sophisticated alignment of multi-scale, semantically rich features.

III. Methodology

Traditional Template Matching methods often struggle with variations in scale, orientation, and illumination. Fig. 2 depicts the overall pipeline of our method. We first employ **SAM3** [5] to generate a set of class-agnostic candidate patches from the input image. Instead of standard resizing, we apply an Aspect-Ratio-Preserving transformation to align the template and candidate regions to a unified resolution. These normalized inputs are processed by **DINOv3** [6] to extract robust visual features. To capture the deep semantic correlation between the template and candidates, we introduce a **Cross-Attention Interaction Module**. The output of this module is aggregated via **Global Pooling** into a unified interaction vector $\mathbf{f}_{\text{inter}} \in \mathbb{R}^D$. This vector serves as the input for our dual-branch prediction system: a **Similarity Scoring Branch** that evaluates the matching probability, and a **Bounding Box Refinement Branch** that fine-tunes the localization coordinates.

A. Zero-Shot Region Proposal Generation

The quality of region proposals largely determines the upper bound of template matching performance. Traditional exhaustive methods, such as Sliding Windows, impose heavy computational costs and often truncate objects because of fixed aspect ratios. Although heuristic methods like Selective Search reduce redundancy, they lack semantic understanding,

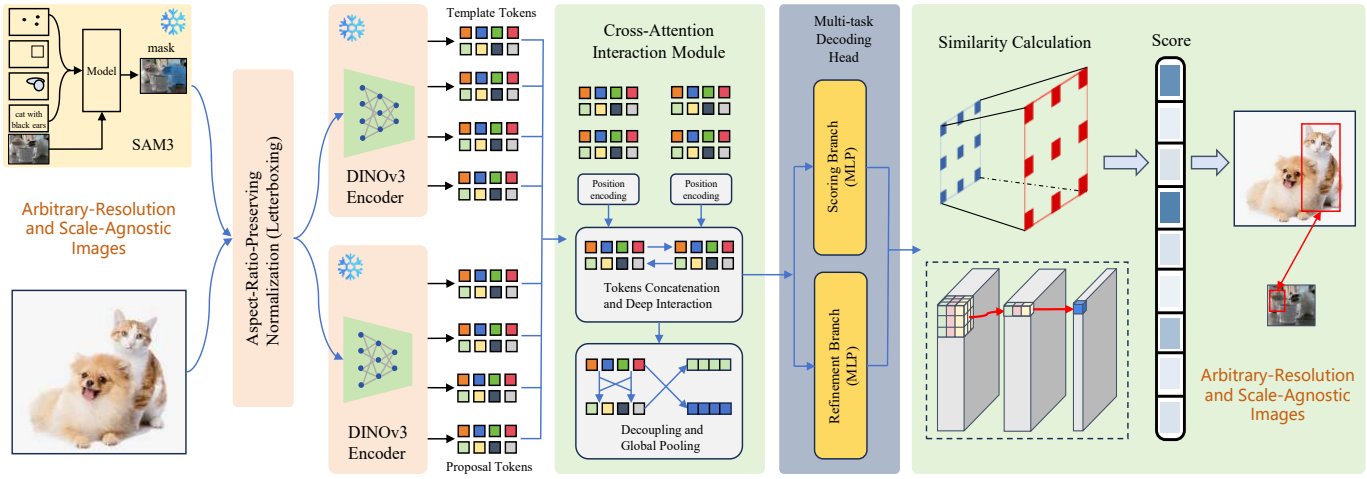


Fig. 2. The workflow of AnyTemplate. AnyTemplate takes any scale and resolution images as input, generates a valid mask via SAM3, and uses DINOv3 for feature extraction, calculates similarity, and yields matching results.

resulting in low recall in cluttered scenes. To address these limitations, we leverage the **Segment Anything Model (SAM3)** [5] as a generic, class-agnostic region proposal generator. SAM3 encodes strong semantic priors learned from billion-scale masks, enabling it to distinguish foreground objects from background noise without task-specific training.

Automatic Grid Prompting. Since the target object's location is unknown *a priori* (zero-shot setting), we employ an automatic mask generation pipeline. Specifically, we overlay a spatially uniform grid of points $\mathcal{G} = \{(x_i, y_i)\}_{i=1}^{N}$ on the input search image I_S. Each point serves as a geometric prompt to the SAM3 mask decoder.

Proposal Extraction. For each prompt, SAM3 predicts a binary segmentation mask $\mathbf{M} \in \{0,1\}^{H \times W}$ and a corresponding IoU confidence score. Unlike traditional detectors that output loose bounding boxes, SAM3's masks adhere tightly to object boundaries. We convert these valid masks into bounding box proposals $\mathcal{B} = \{b_k\}_{k=1}^{K}$ by computing the minimum bounding rectangle for each mask connected component:

$$b_k = (x_{\min}, y_{\min}, x_{\max} - x_{\min}, y_{\max} - y_{\min}) \quad (1)$$

where coordinates are derived from the non-zero indices of the k-th mask.

Quality Filtering. To mitigate redundancy, we apply a dual-filtering strategy: (1) *Stability Filter:* Masks that change significantly under small threshold perturbations are discarded; (2) *NMS:* We apply Non-Maximum Suppression (NMS) on the generated masks to remove near-duplicate proposals, ensuring a compact yet comprehensive set of candidate regions for the subsequent feature extraction stage.

We systematically compare the three methods' region proposal performance. For a 4000×3000 image, Table I shows their proposal counts. With 20×20 stride and a single 160×120 window, Sliding Window generates 25000 boxes. Selective Search considers object features such as color, texture, and size, but still produces approximately 7000 boxes.

TABLE I
THE NUMBER OF REGION PROPOSALS GENERATED BY THE THREE METHODS

Method	Number of Region Proposals
Sliding Window	25,000
Selective Search	7,000
Segment Anything	160

SAM3 generates only about 160 boxes, far fewer than the other methods, while accurately delineating object boundaries to ensure high-quality proposals.

B. Aspect-Ratio-Preserving Normalization

Standard approaches typically resize region proposals to a fixed square resolution to satisfy the input requirements of Vision Transformers like DINOv3. However, this naive resizing operation often introduces severe geometric distortion, particularly for objects with extreme aspect ratios (e.g., long poles or flat signage), which corrupts the semantic structure essential for matching.

To mitigate this issue, we introduce an **Aspect-Ratio-Preserving Normalization** strategy (also known as "Letterboxing"). As shown in Fig. 3, for a given template or region proposal I with original dimensions $H \times W$, we first compute a scaling factor $s = \frac{S}{\max(H,W)}$, where S is the target input size (448×448) of the encoder. The image is then resized to $(H', W') = (sH, sW)$, ensuring the longest dimension equals S while maintaining the original aspect ratio. Finally, we place the resized image onto the center of a square canvas $\mathbf{I}_{norm} \in \mathbb{R}^{S \times S \times 3}$ initialized with a constant padding value. This normalization ensures that the geometric properties of the objects remain intact during feature extraction, significantly enhancing the model's robustness to shape variations.

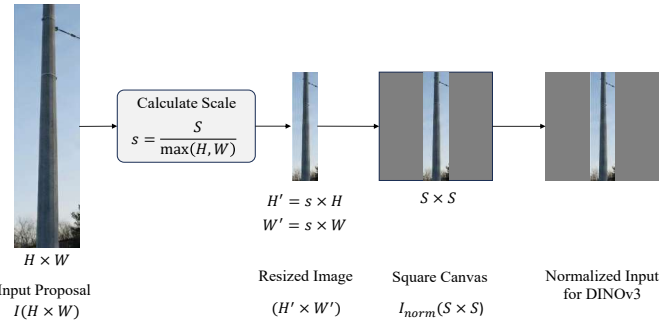


Fig. 3. Schematic of Aspect-Ratio-Preserving Normalization. Input images are resized to a fixed target dimension while preserving their original aspect ratio, with padding applied to fill the remaining canvas area.

C. Cross-Attention Interaction Module

To transcend the limitations of shallow similarity metrics (e.g., Cosine Similarity) that fail to capture complex semantic dependencies, we introduce a **Cross-Attention Interaction Module (CAIM)**. This module serves as a learnable bridge, enabling deep feature interaction between the template and the candidate region proposals. By leveraging the attention mechanism, the module explicitly models the pixel-level correspondence and suppresses background noise in the search region.

Formally, let $\mathbf{F}_T \in \mathbb{R}^{N_T \times C}$ and $\mathbf{F}_P \in \mathbb{R}^{N_P \times C}$ denote the flattened feature sequences of the template and a region proposal extracted by the frozen backbone, respectively, where N represents the number of patches and C is the feature dimension.

Positional Encoding. Since the standard Transformer architecture is permutation-invariant, we first supplement the feature sequences with learnable 2D positional encodings to preserve the spatial geometry of the object. The position-aware features $\mathbf{Z}_T$ and $\mathbf{Z}_P$ are obtained by element-wise addition of the features and their corresponding positional embeddings.

Multi-Head Cross-Attention. We treat the region proposal as the "Query" seeking semantic alignment, and the template as the "Key" and "Value" providing reference information. The queries $\mathbf{Q}$, keys $\mathbf{K}$, and values $\mathbf{V}$ are projected as follows:

$$\mathbf{Q} = \mathbf{Z}_P \mathbf{W}_Q, \quad \mathbf{K} = \mathbf{Z}_T \mathbf{W}_K, \quad \mathbf{V} = \mathbf{Z}_T \mathbf{W}_V \tag{2}$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{C \times d_{head}}$ are learnable projection matrices. We employ Multi-Head Attention (MHA) to capture distinct semantic aspects (e.g., texture, shape) in parallel subspaces. The attention weights are computed via scaled dot-product:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V} \tag{3}$$

where $\sqrt{d_k}$ is a scaling factor to prevent gradient vanishing.

Feature Aggregation. The output of the attention layer is processed by a Feed-Forward Network (FFN) with residual connections and Layer Normalization (LN), resulting in the interacted feature representation $\mathbf{F}_{inter}$. Finally, we apply Global Average Pooling (GAP) to condense the sequence into a unified descriptor vector, which is subsequently fed into the Multi-task Decoding Head for score prediction and box regression.

D. Similarity Calculation and Refinement

The similarity calculation phase computes cosine similarity between each region proposal and template vector, then uses NMS to select top-k results and their bounding boxes as matching outputs.

The Similarity Scoring Branch. To quantify the semantic alignment between the template and the region proposal, we employ a dedicated scoring branch that maps the interacted feature vector to a scalar similarity score. Unlike traditional approaches that rely on shallow metrics like cosine similarity, our scoring branch consists of a three-layer Multi-Layer Perceptron (MLP) with ReLU activations. This non-linear mapping allows the model to weigh the importance of different interactive features dynamically. The final layer uses a Sigmoid activation function to produce a normalized confidence score, representing the probability that the proposal matches the template. During inference, this score is used to rank all proposals, serving as the primary metric for object localization.

The Bounding Box Refinement Branch. Although the region proposals generated by the foundation model (SAM3) are generally high-quality, they are class-agnostic and may not perfectly align with the specific boundaries of the template object (e.g., due to loose fitting or occlusions). To address this, we introduce a regression-based refinement branch parallel to the scoring branch. This branch takes $\mathbf{f}_{\text{inter}}$ as input and predicts a set of four transformation offsets $\boldsymbol{\delta} = (\delta_x, \delta_y, \delta_w, \delta_h)$.

We adopt the standard parameterization for box regression. Given the initial proposal center coordinates and size (x_p, y_p, w_p, h_p), the refined bounding box (x', y', w', h') is computed as:

$$\begin{aligned} x' &= w_p \cdot \delta_x + x_p, \\ y' &= h_p \cdot \delta_y + y_p, \\ w' &= w_p \cdot \exp(\delta_w), \\ h' &= h_p \cdot \exp(\delta_h). \end{aligned}$$

This mechanism allows AnyTemplate to perform sub-pixel localization adjustments, significantly improving the **Intersection over Union (IoU)** with the ground truth, especially for objects with extreme aspect ratios or irregular shapes.

IV. EXPERIMENT

A. Dataset Details

As shown in Fig.4, we present an example from the DEAMPB dataset to evaluate our method. This dataset consists of 10,000 instances, each containing a template image, an input image, and a corresponding text file with detailed annotations. We also categorize the objects in DEAMPB according to the size categories from the COCO dataset [28] and our self-collected datasets for comprehensive analysis.

TABLE II

QUANTITATIVE COMPARISON WITH STATE-OF-THE-ART METHODS ON THE DEAMPB BENCHMARK. **BOLD** INDICATES THE BEST PERFORMANCE, AND UNDERLINED INDICATES THE SECOND BEST. OUR FULL MODEL ACHIEVES SOTA PERFORMANCE IN ZERO-SHOT SCENARIOS.

Category	Method	Accuracy (%) ↑	Precision (%) ↑	Recall (%) ↑	Robustness
Traditional	NCC (Normalized Cross-Correlation)	45.2	52.1	41.3	Low
	BBS (Best-buddies similarity) [24]	62.4	65.8	58.9	Mid
Deep Template	QATM [1]	78.5	80.2	75.1	High
	Sparse-NCNet [25]	81.3	82.5	79.4	High
Feature Matching	SuperGlue [26]	85.6	88.1	83.2	Very High
	LoFTR [27]	89.2	90.5	87.8	Very High
Ours	**AnyTemplate**	**94.1**	**95.8**	**93.2**	**Very High**

TABLE III

ZERO-SHOT GENERALIZATION PERFORMANCE ON THE **MS-COCO** VALIDATION SET.

Method	AP_{Small}	AP_{Medium}	AP_{Large}	mAP_{Total}
NCC	24.5	41.2	58.9	41.5
QATM [1]	55.3	68.7	76.4	66.8
SuperGlue [26]	62.1	79.5	85.2	75.6
LoFTR [27]	65.8	81.2	86.5	77.8
AnyTemplate (Ours)	**78.4**	**89.6**	**92.1**	**86.7**

Fig. 4. Some instances of the DEAMPB dataset. Each instance shows the template and test image are both arbitrary-resolution and scale-agnostic.

TABLE IV

ABLATION STUDY ON THE EFFECTIVENESS OF KEY COMPONENTS ON THE DEAMPB DATASET.

ID	Backbone	Interaction	Refine Head	mAP ↑	IoU ↑
#1	ResNet-50	Cosine	×	72.5	0.65
#2	ViT-B/16	Cosine	×	81.2	0.71
#3	DINOv3	Cosine	×	88.5	0.78
#4	DINOv3	Cross-Attention	×	91.3	0.82
#5	DINOv3	**Cross-Attention**	✓	**94.1**	**0.91**

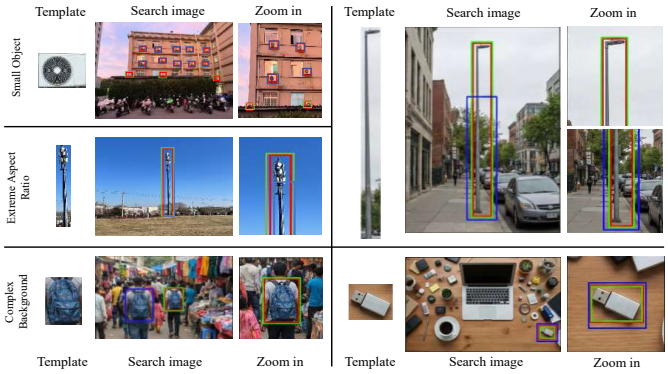


Fig. 5. The visual results of AnyTemplate demonstrate its outstanding performance in Template Matching tasks across diverse scenarios.

B. Experimental Results

Zero-Shot Generalization on DEAMPB. We conducted a comprehensive experimental comparison using standard metrics. As shown in Table II, traditional methods like NCC and Best-buddies similarity (BBS) struggle with complex background clutter and scale variations, yielding sub-optimal accuracy (45.2% and 62.4%, respectively).

While deep template matching methods like QATM leverage learnable similarity measures, they struggle with zero-shot generalization. Recent attention-based approaches like LoFTR and SuperGlue show strong robustness; however, our AnyTemplate outperforms LoFTR by 4.9% in accuracy and 5.3% in precision.

Zero-Shot Generalization on MS-COCO. We conducted a zero-shot evaluation on the **MS-COCO 2017 validation set**. Following standard COCO metrics, as shown in Fig. 5, we report Average Precision (AP) across three object scales: Small ($area < 32^2$), Medium ($32^2 < area < 96^2$), and Large ($area > 96^2$). As shown in Table III, AnyTemplate achieves a remarkable mAP of 86.7%, outperforming the state-of-the-art feature matcher LoFTR by **8.9%**. The performance gap is most pronounced in the **Small Object** (AP_{Small}) category, where our method leads by over **12%** (78.4% vs. 65.8%).

C. Ablation Study

To investigate the contribution of each component in our framework, we conducted a systematic ablation study on the DEAMPB dataset. The results are summarized in Table IV.

Impact of Backbone: Comparing Rows #1, #2, and #3, replacing the standard ResNet-50 with the self-supervised DINOv3 yields a substantial gain of **16.0% in mAP**. This confirms our hypothesis that robust foundation model is crucial for extracting scale-invariant features in zero-shot settings.

Effectiveness of Interaction: By replacing the shallow Cosine Similarity with our **Cross-Attention Interaction Module** (Row #3 vs. #4), the mAP improves from 88.5% to 91.3%. Dynamic context interaction enables the model to better distinguish targets than static metric learning.

Geometry and Precision: Multi-task Decoding Head (Row #5) boosts mAP to 94.1% by mitigating distortion for irregular objects. Incorporating the Cross-Attention Interaction Module (Row #5) yields the best results, significantly increasing IoU from 0.82 to 0.91. This confirms that the regression branch effectively refines SAM3's proposals, achieving precise pixel-level alignment with ground truth.

V. CONCLUSION

In this paper, we introduced AnyTemplate, a robust zero-shot framework that redefines template matching by bridging geometric integrity with deep feature interaction. Addressing the limitations of naive resizing and shallow similarity metrics, our approach integrates Aspect-Ratio-Preserving Normalization to prevent distortion and employs a learnable Cross-Attention Interaction Module to dynamically suppress background noise. Additionally, the novel Multi-task Decoding Head enables precise sub-pixel localization via simultaneous scoring and box refinement. AnyTemplate achieves state-of-the-art performance on the DEAMPB benchmark, outperforming traditional and deep matchers with superior generalization across arbitrary scales and resolutions.

REFERENCES

[1] Jiaxin Cheng, Yue Wu, Wael AbdAlmageed, and Premkumar Natarajan, "Qatm: Quality-aware template matching for deep learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11553–11562.

[2] Itamar Talmi, Roey Mechrez, and Lihi Zelnik-Manor, "Template matching with deformable diversity similarity," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 175–183.

[3] Jinxiang Lai, Liang Lei, Kaiyuan Deng, Runming Yan, Yang Ruan, and Zhou Jinyun, "Fast and robust template matching with majority neighbour similarity and annulus projection transformation," *Pattern Recognition*, vol. 98, pp. 107029, 2020.

[4] Zhirui Gao, Renjiao Yi, Zheng Qin, Yunfan Ye, Chenyang Zhu, and Kai Xu, "Learning accurate template matching with differentiable coarse-to-fine correspondence refinement," *Computational Visual Media*, vol. 10, no. 2, pp. 309–330, 2024.

[5] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al., "Sam 3: Segment anything with concepts," *arXiv preprint arXiv:2511.16719*, 2025.

[6] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al., "Dinov3," *arXiv preprint arXiv:2508.10104*, 2025.

[7] Lichun Mei, Caiyun Wang, Huaiye Wang, Yuanfu Zhao, Jun Zhang, and Xiaoxia Zhao, "Fast template matching in multi-modal image under pixel distribution mapping," *Infrared Physics & Technology*, vol. 127, pp. 104454, 2022.

[8] Yong Li, Jing Chen, Meng Ke, Linghao Li, Zegang Ding, and Yan Wang, "Small targets recognition in sar ship image based on improved ssd," in *2019 IEEE International Conference on Signal, Information and Data Processing (ICSIDP)*. IEEE, 2019, pp. 1–6.

[9] Xiaojie Li, Yao Hu, Tiantian Shen, Shaohui Zhang, Jie Cao, and Qun Hao, "A comparative study of several template matching algorithms oriented to visual navigation," in *Optoelectronic Imaging and Multimedia Technology VII*. SPIE, 2020, vol. 11550, pp. 66–74.

[10] Xin Xiong, Qing Xu, Guowang Jin, Hongmin Zhang, and Xin Gao, "Rank-based local self-similarity descriptor for optical-to-sar image matching," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 10, pp. 1742–1746, 2019.

[11] Jiaxin Cheng, Yue Wu, Wael Abd-Almageed, and Prem Natarajan, "Image-to-gps verification through a bottom-up pattern matching network," in *Asian Conference on Computer Vision*. Springer, 2018, pp. 546–561.

[12] Bo Gao and Michael W Spratling, "Robust template matching via hierarchical convolutional features from a shape biased cnn," in *The International Conference on Image, Vision and Intelligent Systems (ICIVIS 2021)*. Springer, 2022, pp. 333–344.

[13] Fei Shen and Jinhui Tang, "Imagpose: A unified conditional framework for pose-guided person generation," *Advances in neural information processing systems*, vol. 37, pp. 6246–6266, 2024.

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[16] Mingxing Tan and Quoc Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[18] Fei Shen, Cong Wang, Junyao Gao, Qin Guo, Jisheng Dang, Jinhui Tang, and Tat-Seng Chua, "Long-term talkingface generation via motion-prior conditional diffusion model," in *Forty-second International Conference on Machine Learning*.

[19] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al., "Dinov2: Learning robust visual features without supervision," *Transactions on Machine Learning Research Journal*, 2024.

[20] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 6, pp. 2872–2893, 2021.

[21] Vinod Nair and Geoffrey E Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[22] Fei Shen, Jian Yu, Cong Wang, Xin Jiang, Xiaoyu Du, and Jinhui Tang, "Imaggarment-1: Fine-grained garment generation for controllable fashion design," *arXiv preprint arXiv:2504.13176*, 2025.

[23] Fei Shen, Xin Jiang, Xin He, Hu Ye, Cong Wang, Xiaoyu Du, Zechao Li, and Jinhui Tang, "Imagdressing-v1: Customizable virtual dressing," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025, vol. 39, pp. 6795–6804.

[24] Tali Dekel, Shaul Oron, Michael Rubinstein, Shai Avidan, and William T Freeman, "Best-buddies similarity for robust template matching," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2021–2029.

[25] Ignacio Rocco, Relja Arandjelović, and Josef Sivic, "Efficient neighbourhood consensus networks via submanifold sparse convolutions," in *European conference on computer vision*. Springer, 2020, pp. 605–621.

[26] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich, "Superglue: Learning feature matching with graph neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4938–4947.

[27] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou, "Loftr: Detector-free local feature matching with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8922–8931.

[28] Swasti Jain, Sonali Dash, Rajesh Deorari, et al., "Object detection using coco dataset," in *2022 International Conference on Cyber Resilience (ICCR)*. IEEE, 2022, pp. 1–4.