

# MammoTracker: Mask-Guided Lesion Tracking in Temporal Mammograms

Xuan Liu<sup>1,3</sup>, Yinhao Ren<sup>2,3</sup>, Marc D. Ryser<sup>4</sup>, Lars J. Grimm<sup>3</sup>, and Joseph Y. Lo<sup>1,2,3</sup>

<sup>1</sup> Department of Electrical and Computer Engineering, Duke University, Durham, NC

<sup>2</sup> Department of Biomedical Engineering, Duke University, Durham, NC

<sup>3</sup> Department of Radiology, Duke University School of Medicine, Durham, NC

<sup>4</sup> Departments Population Health Science and Mathematics, Duke University School of Medicine, Durham, NC

{xuan.liu115,joseph.lo}@duke.edu

**Abstract.** Accurate lesion tracking in temporal mammograms is essential for monitoring breast cancer progression and facilitating early diagnosis. However, automated lesion correspondence across exams remains a challenge in computer-aided diagnosis (CAD) systems, limiting their effectiveness. We propose MammoTracker, a mask-guided lesion tracking framework that automates lesion localization across consecutive exams. Our approach follows a coarse-to-fine strategy incorporating three key modules: global search, local search, and score refinement. To support large-scale training and evaluation, we introduce a new dataset with curated prior-exam annotations for 730 mass and calcification cases from the public EMBED mammogram dataset, yielding over 20000 lesion pairs, making it the largest known resource for temporal lesion tracking in mammograms. Experimental results demonstrate that MammoTracker achieves 0.455 average overlap and 0.509 accuracy, surpassing baseline models by 8%, highlighting its potential to enhance CAD-based lesion progression analysis. Our dataset will be available at <https://gitlab.oit.duke.edu/railabs/LoGroup/mammotracker>.

**Keywords:** Object Tracking · Mask-guided Mechanism · Mammogram · Computer-Aided Diagnosis.

## 1 Introduction

Temporal analysis of mammograms across multiple consecutive exams provides valuable insights for breast cancer screening [1, 2]. An emerging area of computer-aided diagnosis (CAD) research focuses on monitoring lesion changes over time. This is achieved by establishing correspondences between current and prior mammograms, enabling the assessment of disease progression [3–5]. However, manually tracking lesion locations across different exams is labor-intensive. Therefore, an automated CAD framework is needed to streamline the tracking process, reducing radiologists’ workload while improving efficiency and consistency. This automated lesion correspondence serves as a crucial pre-processing step for

downstream temporal CAD models, enhancing both detection and classification performance.

Lesion tracking, a key component of temporal lesion monitoring, remains underdeveloped [6]. Existing mammogram tracking approaches primarily rely on image registration techniques to align lesion locations [4, 7–9]. However, breast tissue, being soft and deformable, undergoes slight variations across imaging instances, making rigid-body registration methods ineffective. Inspired by object tracking in natural image analysis [11–14], Siamese-based tracking models have shown success in video processing. Nevertheless, adapting such deep learning-based tracking methods for lesion tracking in global mammograms presents unique challenges. Compared to natural images, mammograms have significantly higher spatial resolution (dimensions of 2K–4K), and lesions exhibit variations in size and appearance over time. Conventional down-sampling approaches lead to substantial information loss, particularly in calcification cases, thereby negatively impacting tracking performance.

In this work, we introduce a new temporal lesion tracking dataset based on the public EMBED dataset [1], providing precise lesion annotations for over 700 patients, with exams spanning up to 8 years. To address lesion tracking challenges, we propose MammoTracker, a mask-guided framework mimicking radiologists’ reading behavior. It consists of (1) a global search step using registration-based approach, (2) a local search step with a mask-guided anchor-free tracking model, and (3) a score refinement step through a mask-guided similarity learning model.

We summarize our contributions as follows:

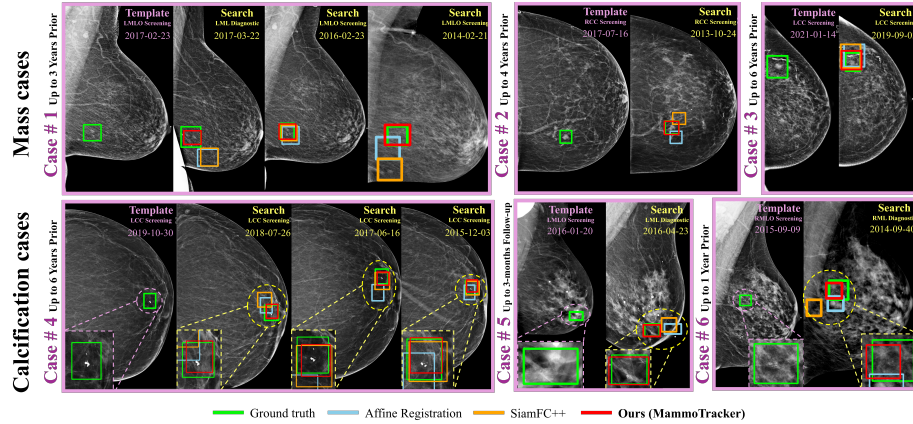
1. We propose MammoTracker, a novel lesion tracking framework to precisely identify lesion locations in temporal mammograms. As shown Fig. 1, MammoTracker outperforms both registration methods and deep learning-based Siamese trackers, with quantitative evaluation metrics confirming its superiority.
2. We release the largest temporal mammogram dataset with lesion annotations for 518 mass and 212 calcification cases. With over 20000 exhaustive lesion pairs spanning up to 8 years, this dataset could be valuable for facilitating future research in temporal lesion analysis and CAD.

## 2 Related Work

**Registration-based Approach.** Temporal lesion tracking leverages spatial consistency, as breast structure remains stable over time. This enables global registration methods (rigid, affine, Demons) for lesion alignment [4, 10], effective for large or stable lesions but less sensitive to local deformations [9]. In this work, registration serves as the global search stage in our cascade tracking framework.

**Anchor-free Tracking Model.** Siamese-based tracking is widely used in visual object tracking [11–14]. Inspired by anchor-free object detection [19], anchor-free tracking removes predefined anchor boxes, improving efficiency and achiev-





**Fig. 1.** Representative experimental results comparing MammoTracker with two baseline trackers on both screening and diagnostic images, demonstrating superior performance in scale adaptation, aspect ratio consistency, and lesion localization precision for both mass and calcification cases.

ing state-of-the-art (SOTA) performance [15, 20]. In this work, we integrate an anchor-free tracking model as the local search component in our framework.

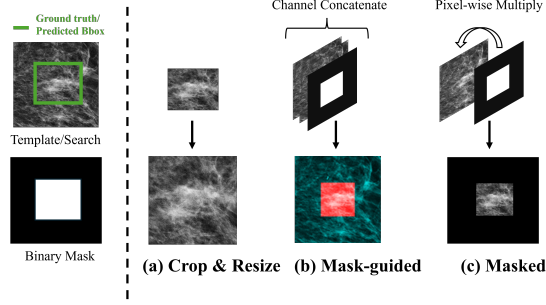
**Mask-guided Mechanism.** Mask-guided mechanisms, as illustrated in Fig. 2 (b), extract robust, background-invariant features, as demonstrated in person re-identification. Chunfeng et al. [16] use RGB-Mask pairs to remove background noise and preserve shape information, while Honglong et al. [17] apply mask-guided attention for improved tracking. Inspired by this, we incorporate mask guidance into our anchor-free tracking and similarity learning models, enhancing lesion-aware feature learning.

### 3 Method

As illustrated in Fig. 3, the proposed MammoTracker framework consists of three main components: (1) Global Search: An affine registration-based approach aligns mammograms at the breast level, narrowing the search area for finer tracking (2) Local Search: A mask-guided anchor-free tracking model accurately localizes lesions within the refined region, improving lesion-background separation. (3) Score Refinement: A mask-guided similarity learning module refines confidence scores for predicted bounding boxes, ensuring more reliable lesion tracking.

#### 3.1 Global Search: Registration Alignment

In the global search, we use affine registration [18] to align images by solving  $\tau_{\text{Aff}} = \arg \min \|\tau_{\text{Aff}}(I_t) - I_s\|_1$ , where  $I_t$  and  $I_s$  denotes the template and search



**Fig. 2.** Three different inputs. (a) Cropped & resized input; (b) Mask-guided input; (c) Masked input.

images, respectively. To improve computational efficiency, both images are down sampled by a factor of 8.

Most lesion sizes in our dataset range from 5 mm and 60 mm, with some reaching 120 mm. Based on this distribution, we define a local template size of 80 mm ( $\sim 1143$  pixels at 0.07 mm spacing). Lesions exceeding this size retrain the registration output as the final tracking result without further refinement.

For comprehensive lesion coverage during local search, the search region size is set to 110 mm ( $\sim 1571$  pixels at 0.07 mm spacing), capturing 97% of lesions based on the center of the registration coordinates. Template patches are resized to  $512 \times 512$  pixels, and search patches to  $1024 \times 1024$  pixels, which are then used for training and inference in the cascade tracking model.

### 3.2 Local Search: Mask-guided Anchor-free Tracking Model

As shown in Fig. 2 (b), template binary masks are generated using ground-truth bounding boxes and concatenated with corresponding template patches to form the input for tracking.

For the anchor-free tracking model, we use pre-trained MobileNetV2 [22] as the backbone with 3-channel input. Following [21], the center  $7 \times 7$  template feature regions are cropped for similarity matching and a depth-wise cross-correlation layer is applied.

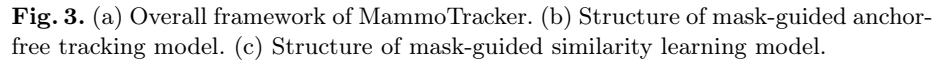
To suppress low-quality predicted bounding boxes, we integrate a center-ness mechanism [13, 19]. The center-ness score is computed as:

$$\text{centerness} = \sqrt{\left( \frac{\min(l, r)}{\max(l, r)} \times \frac{\min(t, b)}{\max(t, b)} \right)} \quad (1)$$

where  $l, r, t, b$  are distances from the predicted bounding box center to its boundaries. The final score is  $cls = \text{centerness} \times \text{classification}$ .

During training, focal-loss is used for classification and center-ness losses, while EIoU [23] is applied for regression. The total loss function is defined as:

$$\text{Loss} = \lambda_1 L_{\text{classification}} + \lambda_2 L_{\text{centerness}} + \lambda_3 L_{\text{reg}} \quad (2)$$


$$L_{\text{reg}} = L_{\text{IoU}} = L_{\text{IoU}} + L_{\text{dis}} + L_{\text{asp}} = 1 - \text{IoU} + \frac{\rho^2(b, b^{gt})}{(w^c)^2 + (h^c)^2} + \frac{\rho^2(w, w^{gt})}{(w^c)^2} + \frac{\rho^2(h, h^{gt})}{(h^c)^2} \quad (3)$$

### 3.3 Score Refinement: Mask-guided Similarity Learning Model

The model generates binary masks from predicted boxes, concatenating them with search patches, as shown in Fig. 2 (b). To optimize efficiency, all patches are resized to 512x512 pixels. IoU-based distance is used as 0 if  $\text{IoU} > 0.5$ , 1 if  $\text{IoU} < 0.3$ , and ignored in training but used in inference otherwise.

Since most predicted search bounding boxes correspond to negatives, we mitigate class imbalance problem by sampling negatives at twice the rate of positives. Additionally, subtraction-based feature distance computation is used, outperforming concatenation and cosine similarity. The model is trained with binary cross-entropy loss, with the final similarity score defined as *Similarity\_Score* =  $1 - distance$ . The final output of MammoTracker combines predicted bounding

boxes from the local tracking model with their corresponding similarity scores, ensuring improved lesion tracking accuracy.

## 4 Experiments and Results

### 4.1 Dataset and Experiment Setup

**Dataset.** EMBED is a large-scale publicly available mammogram dataset containing both screening and diagnostic images with a maximum follow-up period of 8 years [1]. However, lesion annotations are primarily available for the latest study dates, with limited prior annotations. Therefore, we manually annotate lesion locations at each prior time point for 730 cases within the 20% open subset, using the provided region of interest (ROI) annotations as references. The curated dataset comprises approximately 70% screening and 30% diagnostic images, totaling 20426 lesion pairs for training and tests in this study. Table 1 summarizes the detailed mass/calcification and train/test split. All images are rescaled to a reference pixel spacing of 0.07 mm x 0.07 mm.

**Table 1.** Comparison of training and testing datasets with different lesion types, where exhaustive lesion pair is set as the collection of all unique lesion-to-lesion combinations derived from each case.

|                  | Train       |             |                               | Test |      |      |
|------------------|-------------|-------------|-------------------------------|------|------|------|
|                  | Case        | View        | Pair                          | Case | View | Pair |
| Mass             | 352         | 625         | 8062                          | 166  | 317  | 4688 |
| Calcification    | 156         | 329         | 5690                          | 56   | 120  | 1986 |
| Train/Test Total | 518         | 954         | 13752                         | 212  | 437  | 6674 |
| <b>Total</b>     | <b>Case</b> | <b>View</b> | <b>Exhaustive Lesion Pair</b> |      |      |      |
|                  | 730         | 1391        | 20426                         |      |      |      |

**Evaluation metrics.** Following natural image evaluation practices [11], we assess our framework using five metrics. Average overlap (AO) measures the mean IoU across all lesion pairs, while accuracy represents the mean IoU for successful tracking. Robustness evaluates the tracking failure ratio, and average center point L2 distance computes the Euclidean distance (in mm) between ground-truth and predicted bounding box centers. The success plots show the proportion of successfully tracked pairs across IoU thresholds (0 to 1), with the area under the curve (AUC) serving as a comprehensive ranking metric.

**Implementation Details.** The proposed framework is implemented using TensorFlow 2.2 and trained on four NVIDIA 2080 Ti GPUs. Each component is trained independently. The Adam optimizer is used with a learning rate of 0.00005 for both the tracking and similarity learning models. The tracking model is trained for 50 epochs with a batch size of 16, while the similarity learning

model is trained for 5 epochs with a batch size of 8. The affine registration method follows the settings described in [18].

**Table 2.** MammoTracker comparisons on the test dataset.

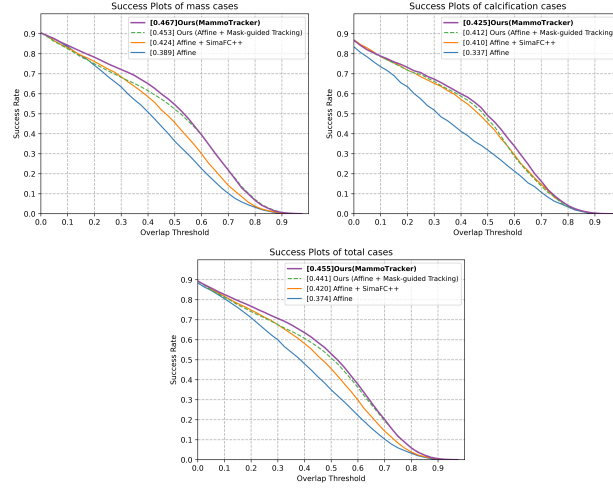
| Lesion Type | Method               | AO $\uparrow$ | Accuracy $\uparrow$ | Robustness $\downarrow$ | L2 distance $\downarrow$ (mm) |
|-------------|----------------------|---------------|---------------------|-------------------------|-------------------------------|
| Mass        | Affine [18]          | 0.389         | 0.430               | 0.095                   | 12.694                        |
|             | SiamFC++ [13]        | 0.424         | 0.469               | <b>0.094</b>            | 11.966                        |
|             | Mask-guided Tracking | 0.453         | 0.501               | 0.097                   | 11.535                        |
|             | <b>MammoTracker</b>  | <b>0.467</b>  | <b>0.516</b>        | 0.095                   | <b>11.057</b>                 |
| Calc        | Affine [18]          | 0.338         | 0.404               | 0.165                   | 13.761                        |
|             | SiamFC++ [13]        | 0.410         | 0.471               | <b>0.130</b>            | 11.794                        |
|             | Mask-guided Tracking | 0.412         | 0.475               | 0.133                   | 11.716                        |
|             | <b>MammoTracker</b>  | <b>0.425</b>  | <b>0.490</b>        | 0.133                   | <b>11.239</b>                 |
| Total       | Affine [18]          | 0.374         | 0.423               | 0.116                   | 13.011                        |
|             | SiamFC++ [13]        | 0.420         | 0.469               | <b>0.105</b>            | 11.915                        |
|             | Mask-guided Tracking | 0.441         | 0.494               | 0.108                   | 11.588                        |
|             | <b>MammoTracker</b>  | <b>0.455</b>  | <b>0.509</b>        | 0.107                   | <b>11.111</b>                 |

## 4.2 Model Comparison

We compare our proposed MammoTracker framework against affine registration [18] and SiamFC++ [13], representing registration-based and anchor-free tracking, respectively. Table 2 presents quantitative results. First, our mask-guided anchor-free tracking model outperforms the SiamFC++ baseline, where template patches are cropped and resized to 512x512 pixels, as illustrated in Fig. 2 (a). AO improves from 0.420 to 0.441 ( $\uparrow 5.0\%$ ) and accuracy increases from 0.469 to 0.49 ( $\uparrow 4.5\%$ ), primarily due to gains in mass lesion tracking, where AO and accuracy increase by 6.8% and L2 distance decreases by 3.6%.

Next, we evaluate the full MammoTracker framework, which includes cascade mask-guided similarity learning model. As shown in Table 2, it achieves notable improvements across AO and accuracy. Specifically, for mass lesions, AO increases from 0.424 to 0.467 ( $\uparrow 10.1\%$ ), and accuracy from 0.469 to 0.516 ( $\uparrow 10.0\%$ ). For calcifications, AO rises from 0.410 to 0.425 ( $\uparrow 3.7\%$ ), while accuracy improves from 0.471 to 0.490 ( $\uparrow 4.0\%$ ). Additionally, L2 distance is reduced by over 5% for both lesion types. Fig. 4 further illustrates that MammoTracker consistently achieves higher success rates across all IoU thresholds compared to SiamFC++ for both mass and calcification lesions.

However, SiamFC++ shows slightly better robustness, indicating potential areas for improvement in failure recovery for challenging cases.



**Fig. 4.** Success plots show a comparison of our tracker with others in different lesion types. From top left to top right and then bottom center: (a) Mass Cases; (b) Calcification Cases; (c) Both mass and calcification cases.

### 4.3 Ablation Study

We conducted ablation studies to evaluate the effectiveness of our proposed methods in 2 key aspects within the mask-guided anchor-free tracking model.

**Center-ness.** We evaluate the impact of incorporating center-ness in our anchor-free tracking model. As shown in Table 3, we compared two approaches: (1) using only the classification score and (2) using the product of the classification and center-ness score. Results show that center-ness significantly improves AO, accuracy and L2 distance by suppressing low-quality bounding boxes, leading to more reliable tracking performance.

**Mask-guided vs Masked Template Input.** We further examine the effect of different template input types by training the tracking model using three variations: (1) crop & resize; (2) mask-guided and (3) masked template, as shown in Fig. 2. As summarized in Table 3, the mask-guided template consistently outperforms the other methods across AO, accuracy and L2 distance metrics. This superiority can be attributed to three factors, aligning with findings from [16]: (1) rectangular masks effectively separate lesions from the background, enhancing discriminative feature learning; (2) the mask preserves stable bounding box shape information over time, preventing abrupt aspect ratio shifts; and (3) unlike fully masked templates, mask-guided templates retain weak background features, providing contextual cues for improved lesion localization.

**Table 3.** Ablation study on each module in anchor-free tracking model for both mass and calcification cases.

| Centerness | Template Inputs |             |        | AO $\uparrow$ | Accuracy $\uparrow$ | Robustness $\downarrow$ | L2 distance $\downarrow$<br>(mm) |
|------------|-----------------|-------------|--------|---------------|---------------------|-------------------------|----------------------------------|
|            | Crop & Resize   | Mask-guided | Masked |               |                     |                         |                                  |
|            | ✓               |             |        | 0.413         | 0.461               | <b>0.104</b>            | 12.024                           |
|            |                 | ✓           |        | 0.431         | 0.484               | 0.110                   | 11.715                           |
|            |                 |             | ✓      | 0.407         | 0.464               | 0.123                   | 11.776                           |
| ✓          | ✓               |             |        | 0.420         | 0.469               | 0.105                   | 11.915                           |
| ✓          |                 |             | ✓      | 0.416         | 0.479               | 0.132                   | 11.778                           |
| ✓          |                 | ✓           |        | <b>0.441</b>  | <b>0.494</b>        | 0.108                   | <b>11.588</b>                    |

## 5 Conclusion

In this study, we introduce MammoTracker, a mask-guided lesion tracking framework for temporal mammograms that enables accurate lesion localization across multiple time points. It follows a coarse-to-fine strategy that replicates radiologists’ approach to reading sequential images and identifying corresponding lesions. We also release a large-scale dataset with over 20000 tracking pairs, based on the EMBED dataset. MammoTracker outperforms baseline models in tracking accuracy, average overlap and L2 distance. As part of our ongoing work, we are expanding our experiments to additional datasets, including OPTIMAM and a private Duke dataset, to facilitate comprehensive analysis of lesion growth and shape changes over time. In future work, we will extend our framework to downstream CAD tasks, such as lesion detection and classification, to further enhance breast cancer diagnosis.

**Acknowledgments.** This work was supported by NIH/NCI R01 CA271237.

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References

1. Jeong, J.J., Vey, B.L., Bhimireddy, A., Kim, T., Santos, T., Correa, R., Dutt, R., Mosunjac, M., Oprea-Ilie, G., Smith, G. and Woo, M., 2023. The EMory BrEast imaging Dataset (EMBED): A racially diverse, granular dataset of 3.4 million screening and diagnostic mammographic images. *Radiology: Artificial Intelligence*, 5(1), p.e220047.
2. Halling-Brown, M.D., Warren, L.M., Ward, D., Lewis, E., Mackenzie, A., Wallis, M.G., Wilkinson, L.S., Given-Wilson, R.M., McAviney, R. and Young, K.C., 2020. Optimam mammography image database: a large-scale resource of mammography images and clinical data. *Radiology: Artificial Intelligence*, 3(1), p.e200103.
3. Li, H., Robinson, K., Lan, L., Baughan, N., Chan, C.W., Embury, M., Whitman, G.J., El-Zein, R., Bedrosian, I. and Giger, M.L., 2023. Temporal Machine Learning Analysis of prior mammograms for breast Cancer risk prediction. *Cancers*, 15(7), p.2141.

4. Loizidou, K., Skouroumouni, G., Nikolaou, C. and Pitris, C., 2020. An automated breast micro-calcification detection and classification technique using temporal subtraction of mammograms. *IEEE Access*, 8, pp.52785-52795.
5. Kooi, T. and Karssemeijer, N., 2017. Classifying symmetrical differences and temporal change for the detection of malignant masses in mammography using deep neural networks. *Journal of Medical Imaging*, 4(4), pp.044501-044501.
6. Cai, J., Tang, Y., Yan, K., Harrison, A.P., Xiao, J., Lin, G. and Lu, L., 2021. Deep lesion tracker: monitoring lesions in 4D longitudinal imaging studies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 15159-15169).
7. Zheng, Y., Yang, C. and Merkulov, A., 2018, May. Breast cancer screening using convolutional neural network and follow-up digital mammography. In *Computational Imaging III* (Vol. 10669, p. 1066905). SPIE.
8. Sharma, M.K., Jas, M., Karale, V., Sadhu, A. and Mukhopadhyay, S., 2019. Mammogram segmentation using multi-atlas deformable registration. *Computers in biology and medicine*, 110, pp.244-253.
9. Timp, S., van Engeland, S. and Karssemeijer, N., 2005. A regional registration method to find corresponding mass lesions in temporal mammogram pairs. *Medical physics*, 32(8), pp.2629-2638.
10. Díez, Y., Oliver, A., Llado, X., Freixenet, J., Martí, J., Vilanova, J.C. and Martí, R., 2011. Revisiting intensity-based image registration applied to mammography. *IEEE Transactions on Information Technology in Biomedicine*, 15(5), pp.716-725.
11. Bertinetto, L., Valmadre, J., Henriques, J.F., Vedaldi, A. and Torr, P.H., 2016. Fully-convolutional siamese networks for object tracking. In *Computer vision—ECCV 2016 workshops: Amsterdam, the Netherlands, October 8-10 and 15-16, 2016, proceedings, part II 14* (pp. 850-865). Springer International Publishing.
12. Li, B., Wu, W., Wang, Q., Zhang, F., Xing, J. and Yan, J., 2019. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4282-4291).
13. Xu, Y., Wang, Z., Li, Z., Yuan, Y. and Yu, G., 2020, April. SiamFC++: Towards robust and accurate visual tracking with target estimation guidelines. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 34, No. 07, pp. 12549-12556).
14. Chen, Z., Zhong, B., Li, G., Zhang, S. and Ji, R., 2020. Siamese box adaptive network for visual tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6668-6677).
15. Han, G., Su, J., Liu, Y., Zhao, Y. and Kwong, S., 2021. Multi-stage visual tracking with siamese anchor-free proposal network. *IEEE Transactions on Multimedia*, 25, pp.430-442.
16. Song, C., Huang, Y., Ouyang, W. and Wang, L., 2018. Mask-guided contrastive attention model for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1179-1188).
17. Cai, H., Wang, Z. and Cheng, J., 2019. Multi-scale body-part mask guided attention for person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 0-0).
18. Lowekamp, B.C., Chen, D.T., Ibáñez, L. and Blezek, D., 2013. The design of SimpleITK. *Frontiers in neuroinformatics*, 7, p.45.
19. Tian, Z., Shen, C., Chen, H. and He, T., 2019. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 9627-9636).
20. Zhang, J., Huang, B., Ye, Z., Kuang, L.D. and Ning, X., 2021. Siamese anchor-free object tracking with multiscale spatial attentions. *Scientific reports*, 11(1), p.22908.



21. Valmadre, J., Bertinetto, L., Henriques, J., Vedaldi, A. and Torr, P.H., 2017. End-to-end representation learning for correlation filter based tracking. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2805-2813).
22. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. and Chen, L.C., 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4510-4520).
23. Zhang, Y.F., Ren, W., Zhang, Z., Jia, Z., Wang, L. and Tan, T., 2022. Focal and efficient IOU loss for accurate bounding box regression. *Neurocomputing*, 506, pp.146-157.