Learning Random Noise Salient Feature Fusion Siamese Network for Low-Resolution Object Tracking (Student Abstract)

Md Maklachur Rahman, Tracy Hammond*

Department of Computer Science and Engineering, Texas A&M University, College Station, TX, USA
{maklachur, hammond}@tamu.edu

Abstract

Despite Siamese trackers' substantial potential, they offer sub-optimal tracking performance in low-resolution (LR) contexts. We introduce a Random Noise Salient Feature Fusion Learning Network to address this issue. This method integrates random noise-infused feature maps into a similarity-learning matching model. This integration acts as an effective regularization technique, enhancing the network's generalization capabilities in LR environments. Additionally, by integrating attention mechanisms, we enhance the discriminative ability of the network, assigning more weights to important features. This directs the network's focus toward the most salient regions of the feature map, ensuring improved accuracy without a significant increase in parameter overhead, and maintaining a high operating speed. To validate the effectiveness of our method, we performed qualitative and quantitative comparisons with state-of-the-art (SOTA) trackers.

Introduction

Visual object tracking (VOT) is an important task in computer vision, aiming to track objects from the initial video frame. Despite deep learning advancements, creating a suitable tracker remains challenging, especially in LR scenarios.

Our work focuses on addressing LR challenges, which are particularly relevant in real-life applications such as low-cost UAVs, remote sensing, and intelligent surveillance. LR imagery is often compromised by sparse target details, making it challenging for traditional tracking algorithms to perform effectively.

Recently, many trackers (Chen et al. 2021), and (Gao et al. 2022) have been proposed, but they are not suitable for real-time applications. Siamese tracker (Bertinetto et al. 2016) was introduced to advance the tracking performance, which gained popularity and many follow-up works emerged. Still, they struggle with LR challenges, impacting real-life applications. To overcome this limitation, we propose a novel tracking framework called Random Noise Salient Feature Fusion Siamese Network shown in Figure 1. We combined randomly selected noise-infused image features with traditional features and integrated attention mechanisms to learn salient target information.

*Corresponding author

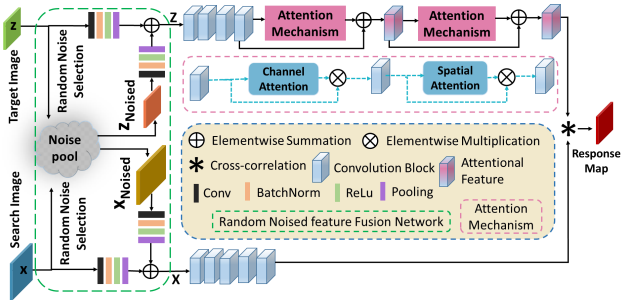

Figure 1: The proposed tracker pipeline: (z, x) pairs represent target and search images, while (z_{noised}, x_{noised}) are corresponding noised image pairs. Fused features and the response map are the input-output, respectively.

The proposed integrated framework effectively addresses the LR tracking challenges, as demonstrated through both qualitative results in Figure 2 and quantitative comparisons in Table 1 when compared to SOTA trackers.

Methodology

Baseline Siamese Tracking Pipeline: We begin by employing the SiamFC framework (Bertinetto et al. 2016) as our baseline, which consists of two identical Fully-Convolutional Neural Networks (FCN) branches: the template branch, responsible for the target image, and the search branch, handling subsequent video frames. These branches learn through parameter sharing. The core operation involves cross-correlation ($corr(.)$) to compute a similarity score between their feature maps, predicting the target's location as the position with the highest similarity score in the response map, expressed as:

$$f_\theta(z, x) = corr(\Psi_\theta(z), \Psi_\theta(x)), \quad (1)$$

where $f_\theta(z, x)$ represents the response map between target image(z) and search image(x) feature maps.

Random Noise Feature Fusion Learning: Usually, SiamFC inputs a pair of images (target, search) from the same image sequence, which offers minimal variation and can lead to overfitting. To mitigate this, we introduce a random noised input feature fusion learning strategy by altering

the baseline SiamFC architecture. Mathematically:

$$Z = \psi_\theta(z) \oplus \psi_\theta(z_{noised}), \tag{2}$$

$$X = \psi_\theta(x) \oplus \psi_\theta(x_{noised}), \tag{3}$$

$$f_\theta(Z, X) = corr(\overline{S_{at}}(\Psi_\theta(Z_s)), \Psi_\theta(X)), \tag{4}$$

where Z and X represent the fused features of the original inputs z and x, Z_s stands for the salient feature map of Z, $\oplus$ denotes element-wise summation as the feature fusion operation, and $\overline{S_{at}}$ denotes the salient attentional feature map.

During training, besides the conventional image pair (z, x), a noised-infused image pair is fed to the network to learn the feature variation. This technique aims to regularize the network and improve its tracking capabilities in diverse scenarios leveraging various noise types like gaussian, salt and pepper, poisson, speckle, gaussBlur, and sharpen. The noise fusion approach stabilizes the network, emphasizing the significance of finer details in LR images and enhancing tracking performance in such conditions.

Salient Feature Attention Network: Inspired by the attention mechanism benefits in computer vision tasks, we introduced a salient feature-based attention mechanism leveraging CBAM (Woo et al. 2018). This mechanism learns high-level features from the later layers of FCN, which contain more semantic and abstract image feature details. Our attention mechanism can be summarized as:

$$\overline{S_{at}} = S_{lf} \oplus S_{at} + S_{hf} \oplus S_{at}, \tag{5}$$

where $\overline{S_{at}}$ represents the salient feature attention mechanism, S_{lf} and S_{hf} denote the feature maps for low and high convolutional block features extracted from the salient feature map, respectively. S_{at} represents the block-wise CBAM attention network.

The attention mechanism especially plays a pivotal role in LR scenarios. By concentrating on high-level, salient features, our model improves its discriminative ability by learning to prioritize key target details: allocating more weights to important target details and less to non-target information. This approach reduces the ambiguity between the target and background, a challenge often exacerbated by LR. We experimentally finalize the placement of the attention mechanism.

Overall, noise infusion and salient attention feature fusion in the Siamese pipeline enhance tracker capabilities and provide a reliable way to track objects in LR.

Experiments

We trained our network using logistic loss and stochastic gradient descent (SGD) with momentum (0.9), exponential learning rate decay (from 10^{-2} to 10^{-5}), and weight decay (5×10^{-4}). We used 127×127 target images and 255×255 search images for training, applying noise infusion only during this phase. We fine-tuned the noise infusion process by experimenting with different layers and proportions for optimal performance.

To evaluate our method, we compared it with SOTA tracking methods on LR challenges, using popular OTB100 benchmarks with curated challenging attributes. We assessed tracker performance using overlap scores (success and precision scores) and tracking speed.

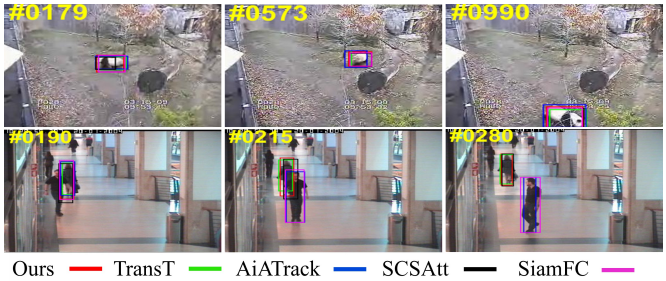

Figure 2: SOTA trackers weaknesses and qualitative comparison with our tracker on panda and walking2 sequences.

Tracker	Ours	TransT	AiATrack	SCSAtt	SiamFC
Success	**0.685**	0.678	0.663	0.639	0.573
Precision	**0.937**	0.891	0.918	0.872	0.815
Speed (FPS)	68	50	38	61	**86**

Table 1: Performance comparison with SOTA methods.

Our results, as shown in Figure 2, demonstrate that our tracker (highlighted in red) accurately tracks the target over frames. Table 1 shows a performance comparison with TransT (Chen et al. 2021), AiATrack (Gao et al. 2022), SCSAtt (Rahman, Fiaz, and Jung 2020), and baseline SiamFC (Bertinetto et al. 2016) trackers. Our method achieved notable success (AUC) with a score of 0.685 and a precision score of 0.937, outperforming SOTA methods. Additionally, our method operates at over real-time speed, achieving 68 FPS. These results highlight the efficacy of our approach in challenging LR tracking scenarios.

Conclusion

The proposed method boosts the baseline Siamese network strength by strategic noise-infusion and integrating attention mechanisms to enhance object-tracking performance. By addressing the limitations inherent in Siamese tracking frameworks, particularly in LR scenarios, our method demonstrated its efficacy and dominance in success and precision scores compared to SOTA trackers. Furthermore, our approach operates in real-time, making it applicable to real-life applications.

References

Bertinetto, L.; Valmadre, J.; Henriques, J. F.; Vedaldi, A.; and Torr, P. H. 2016. Fully-convolutional siamese networks for object tracking. In *ECCV*, 850–865.

Chen, X.; Yan, B.; Zhu, J.; Wang, D.; Yang, X.; and Lu, H. 2021. Transformer tracking. In *CVPR*, 8126–8135.

Gao, S.; Zhou, C.; Ma, C.; Wang, X.; and Yuan, J. 2022. Aiatrack: Attention in attention for transformer visual tracking. In *ECCV*, 146–164.

Rahman, M. M.; Fiaz, M.; and Jung, S. K. 2020. Efficient Visual Tracking With Stacked Channel-Spatial Attention Learning. *IEEE Access*, 8: 100857–100869.

Woo, S.; Park, J.; Lee, J.-Y.; and So Kweon, I. 2018. Cbam: Convolutional block attention module. In *ECCV*, 3–19.