RandConv 기반 도메인 적응으로 강화된 다중 도메인 객체 추적 기술 연구

*최성준, 이승훈, 박채원, 전인석, 이상윤
연세대학교 전기전자공학부
e-mail : { *sjchoi.dp, shlee423, chaewon28, sunlight3919, syleee}@yonsei.ac.kr*

Multi-Domain Object Tracking Enhanced by RandConv-based Domain Generalization

Seongjun Choi, Seunghoon Lee, Chaewon Park, Inseok Jeon, Sangyoun Lee
School of Electrical and Electronic Engineering,
Yonsei University

Abstract

This paper proposes an efficient, domain-agnostic object tracking model capable of universal training across multiple sensing modalities. Building on AQATrack's autoregressive query-update mechanism, our design effectively captures both temporal dynamics and changes in object appearance. The architecture further supports easy fine-tuning for electro-optical (EO) and infrared (IR) environments. To enhance robustness against real-world sensor variations, we incorporate RandConv during training to induce synthetic image perturbations, thereby improving generalization to diverse EO/IR data. Experimental results show that, after fine-tuning on the GOT-10k dataset converted to grayscale, our model achieves 90.26% precision on UAV123 (EO) and 85.11% precision on LSOTB-TIR (IR), demonstrating superior tracking performance across heterogeneous domains.

I. 서론

최근 딥러닝 기반 객체 추적 기술은 영상 내 객체의 위치 및 상태를 실시간으로 파악하는 데 큰 발전을 이루었다. 그러나 기존의 객체 추적기는 대개 RGB 영상에 최적화되어 있어, EO(전기광학) 및 IR(적외선) 등 센서별 상이한 영상 특성을 반영하기에는 한계가 있었다. 본 논문에서는 EO 및 IR 등 서로 다른 센서 도메인의 영상 특성을 반영하여, 시계열 정보 통합과 특징 학습을 한층 강화한 객체 추적 모델을 제안한다.

특히, AQATrack [1]은 자기회귀 기반 쿼리 업데이트 메커니즘을 도입하여 각 시점의 특징 정보를 이전 프레임의 쿼리와 통합하는 방식으로 동작한다. 이와 함께 도메인 일반화 기법인 RandConv [5]를 적용하여, 학습 단계에서 인위적으로 도메인 간의 차이를 보완하는 동시에, 다양한 센서 도메인에서의 일반화 성능을 향상시키고자 하였다.

II. 관련 연구

객체 추적 분야는 고속 실행을 목표로 한 Siam 계열 추적기에서부터 Transformer · 메모리 기반의 딥러닝 모델에 이르기까지 빠르게 발전해 왔다. Siam 계열의 원형이라 할 수 있는 SiamFC [2]는

템플릿과 검색 영역 간 전역 상관 연산만으로 대상을 찾는 Fully-Convolutional Layer를 제안해, 단순성과 80 fps 이상의 실시간 처리 속도로 후속 연구들의 토대를 마련하였다. 그러나 템플릿이 고정돼 있어 외형 변화에 적응하기 어렵다는 한계가 존재한다.

최근에는 Transformer를 활용해 장거리 의존성을 보다 정교하게 포착하려는 시도가 활발하다. CTTrack [3]는 MAE 기반 데이터 증강과 Compact Transformer 블록을 결합해 주변 컨텍스트 정보를 통합함으로써 정확도를 끌어올렸으나, 성능 향상 요인이 모델 구조가 아닌 학습 기법에 크게 의존하고 코드 재현성이 낮아, 연구 확장 및 도메인 일반화 실험에는 제약이 있다.

또 다른 흐름은 과거 정보를 명시적으로 저장·재사용하는 메모리 기반 추적이다. STMTrack [4]는 과거 프레임 특징을 공간-시간 메모리에 저장해 반복 참조함으로써 긴 시퀀스에서도 강인한 성능을 달성한다. 다만 메모리 관리 비용이 크고, 메모리에 적재할 프레임 수를 늘릴수록 추론 지연이 커져 임베디드·실시간 환경에서 적용이 제한될 수 있다.

이처럼 기존 방법들은 각기 강점을 지니지만 (i) 외형 변화 적응 한계, (ii) 학습-전용 기법 의존성, (iii) 메모리·연산 비용 증가 등으로 EO/IR을 포함한 다중 센서 환경에서의 범용 추적에는 한계를 보인다. 이에 반해 AQATrack은 복잡한 템플릿 갱신이나 대용량 메모리 네트워크 없이, 자기회귀 기반 쿼리 재활용만으로 시계열 변화를 저비용으로 포착하고 공간·시간 정보를 효율적으로 융합할 수 있어, EO/IR 등 다중 센서 환경에서 경량·고성능 추적 연구를 확장하기에 가장 적합하다고 판단하였다.

Ⅲ. 제안 방법

3.1 전체 아키텍처 개요

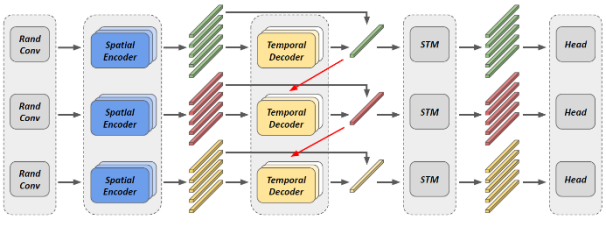

그림 1. 제안한 모델 아키텍처

본 연구에서 제안하는 전체 아키텍처는 입력 영상에 도메인 일반화 기법인 RandConv를 선행 적용한 후, 객체 추적을 위한 AQATrack 모델로 입력되는 구조로 설계되었다. RandConv는 입력 영상에 무작위 커널 기반의 변형을 가하여 도메인 간 시각적 차이를 인위적으로 생성함으로써, 다양한 영상 환경에서도 강인한 특징 표현을 가능하게 한다. 이후 변형된 영상은 AQATrack의 네 구성 요소—공간 인코더, 시간 디코더, 시공간 융합 모듈, 예측 헤드—를 순차적으로 거친다. 각 모듈은 공간 및 시간적 정보를 정교하게 추출 및 통합하여, 외형 변화와 운동 패턴에 민감하게 반응하는 정밀한 객체 추적 결과를 도출한다. 이와 같은 구조는 다양한 센서 조건 및 도메인에서의 일반화 성능을 확보하면서도, 시계열 기반 객체 추적의 정확도 또한 함께 향상시키는 데 기여한다.

3.2 베이스라인 모델 개요: AQATrack

그림 1에 제시된 네트워크 학습 파이프라인에 따르면, AQATrack은 크게 네 부분으로 구성된다.

- **공간 인코더:** 입력 영상에서 공간적 특징을 추출하여 벡터로 변환
- **시간 디코더:** 자기회귀 기반 쿼리 업데이트 메커니즘을 통해 이전 프레임의 쿼리 정보를 반영하여 시간적 특징을 모델링
- **시공간 융합 모듈:** 공간 및 시간 특징을 유사도 연산 기반으로 융합하여 타겟 객체 주변의 가중치 맵 산출
- **예측 헤드:** 융합된 특징을 바탕으로 최종 바운딩 박스 예측 수행

이러한 구조는 매 프레임마다 업데이트된 쿼리 정보를 다음 프레임에 전달함으로써, 객체의 외형 변화나 운동 패턴을 세밀하게 반영할 수 있는 이점을 제공한다.

3.3 도메인 일반화 기법: RandConv 적용

도메인 간 영상 특성의 차이는 객체 추적 성능에 큰 영향을 미칠 수 있다. 이를 극복하기 위해 본 연구에서는 RandConv를 학습 과정에 도입하였다.

RandConv는 가우시안 분포에서 랜덤하게 샘플링된 커널을 활용하여 입력 이미지에 변형을 가함으로써, 학습 시 보지 못한 다양한 도메인 조건을 인위적으로 구현한다. 이러한 기법은 모델이 단일 도메인(RGB)에서 학습한 후에도 EO, IR 등 상이한 영상에서의 일반화 성능을 크게 향상시키는 역할을 한다.

3.4 Fine-Tuning 전략

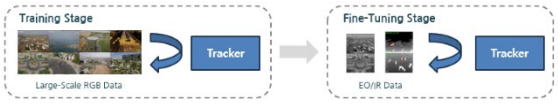

그림 2. 네트워크 학습 파이프라인

그림 2는 제안하는 네트워크의 학습 파이프라인을 보여준다. 본 연구에서 활용하는 단일 채널 데이터는 정보량이 부족하고, 특히 IR 데이터의 양이 충분치 않아 네트워크를 처음부터 학습하면 과적합이 발생하기 쉽다. 따라서 먼저 대규모 RGB 데이터셋으로 사전 학습된 파라미터를 초기화한 뒤, 이를 Gray-Scale로 변환한 GOT-10k 데이터셋을 이용해 파인튜닝을 수행하였다.

- **백본 학습:** main training stage에서 학습한 모델의 방대한 양의 사전지식이 소실되지 않도록 하기 위해 encoder의 weight는 고정하였으며, 모델에 최소한의 튜닝을 통해 EO/IR 도메인의 특성을 반영함
- **데이터셋 선택:** 두 데이터셋의 상이한 분포로 인한 학습 혼선을 방지하기 위해, 다양성과 일반화 능력이 우수한 GOT-10k만을 활용함
- **도메인 일반화:** 데이터셋이 가지는 영상의 특성 및 분포가 상이하기에 모델이 일관된 특성을 학습하는데 어려움을 겪을 여지가 존재한다고 판단하여, RandConv 기법을 적용하여 Fine-Tuning 과정에서 발생할 수 있는 데이터 도메인 간 차이를 극복함

Ⅳ. 실험

4.1 데이터셋 및 평가 지표

실험에는 다음 데이터셋을 활용하였다.

- **GOT-10k [6] (모델 학습용):** 다양한 객체 클래스와 모션 변화가 포함된 대규모 영상 데이터셋인 GOT-10k를 Gray-scale로 변환 후 학습에 사용
- **UAV123 [7] (EO 도메인 평가용):** 드론 뷰에서 취득한 영상으로, 타겟 객체가 작은 경우가 많아 검출이 어려운 환경 제공
- **LSOTB-TIR [8] (IR 도메인 평가용):** IR 카메라로 취득된 데이터셋으로, 보안 및 감시 시나리오를 중심으로 구성

평가 지표로는 객체 추적 분야에서 널리 사용되는 Precision을 채택하였으며, 예측된 바운딩 박스와 실제 객체 박스 간의 픽셀 단위 일치도를 측정하였다.

4.2 정량적 실험 결과

표 1. 정량적 분석 결과

	UAV123	LSOTB-TIR
STMTrack	82.5	81.8
AQATrack full	86.39	82.85
AQATrack tune	89.83	81.88
AQATrack tune w/RandConv	<u>90.26</u>	<u>85.11</u>

- **AQATrack full:** 모든 모듈을 포함(full)해 학습한 모델
- **AQATrack tune:** backbone을 freeze한 상태에서 파인튜닝한 모델
- **AQATrack tune w/RandConv:** 파인튜닝 시 RandConv 데이터 증강을 추가 적용한 모델

Fine-Tuning 후, 표 1에 따라 제안하는 모델은 UAV123 데이터셋에서 90.26%, LSOTB-TIR 데이터셋에서 85.11%의 Precision을 기록하였다.

실험 결과는 다음과 같은 시사점을 내포한다.

- **도메인 적응성 강화:** RandConv를 통한 도메인 일반화 적용으로, 학습 데이터가 Gray-Scale GOT-10k에 한정됨에도 불구하고 IR 도메인 평가에서도 높은 성능을 보임
- **일관된 성능 개선:** 제안하는 모델은 AQATrack의 시계열 쿼리 업데이트와 도메인 특화 Fine-Tuning 전략을 결합하여, 다양한 환경에서도 일관되고 우수한 추적 성능을 보임

4.3 정성적 실험 결과

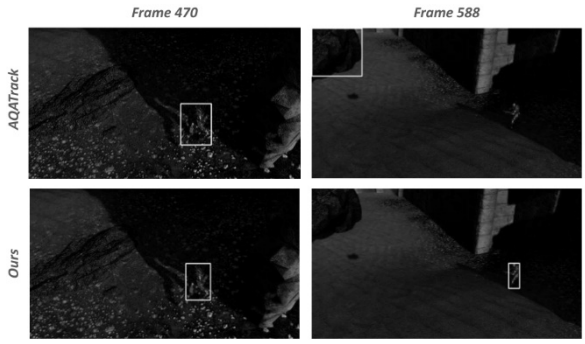

그림 3. 정성적 실험 결과: EO 도메인

그림 3은 EO 도메인에서의 제안한 방법(Ours)과 기존 방식(AQATrack)의 정성적 비교 결과를 나타낸다. 실험 결과, 제안한 방법은 조명 변화가 큰 영역에서도 추적 대상을 연속성 있게 탐지하며, 어두운 구간에서 사라졌다가 밝은 구간에 재등장해도 끊김 없이 정확한 위치를 유지하는 것을 확인하였다.

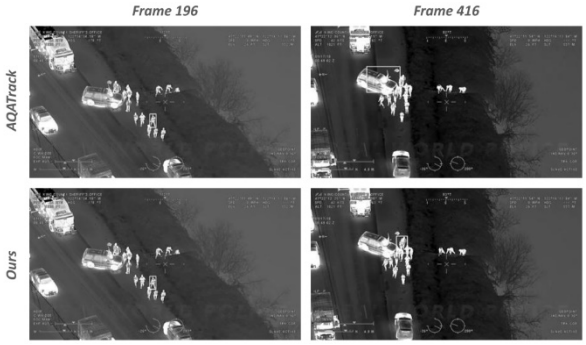

그림 4. 정성적 실험 결과: IR 도메인

그림 4은 IR 도메인에서의 제안한 방법(Ours)과 기존 방식(AQATrack)의 정성적 비교 결과를 나타낸다. 실험 결과, 제안한 방법은 그림자에 일시적으로 가려졌다가 다시 나타나는 객체에 대해서도 안정적으로 추적을 유지하였으며, 여러 사람이 밀집한 복잡한 군집 움직임에서도 타겟 객체를 혼동 없이 안정적으로 식별하고, 분리 시점에 정확히 추적 대상을 복원하는 등 다양한 어려운 환경에서도 강인한 추적 성능을 보여주었다.

이 결과는 제안한 모델이 IR·EO 등 다른 도메인 특성에 적응하여, 변화가 심한 환경에서도 안정적인 추적 성능을 유지함을 입증한다.

Ⅴ. 결론

본 연구에서는 AQATrack과 RandConv 기반의 도메인 일반화 기법을 결합하여, 객체 추적 분야에서 EO, IR 등 다양한 영상 도메인에 대해 우수한 성능을 달성할 수 있음을 제시하였다.

세부적인 모델 구조와 Fine-Tuning 전략을 통해, 제한된 Gray-Scale 데이터셋에도 불구하고 UAV123 및 LSOTB-TIR 데이터셋에서 높은 수준의 결과를 확보하였으며, 이는 향후 다양한 센서 환경에 적용 가능한 범용 객체 추적 시스템 개발에 기여할 것으로 기대된다.

추후 연구에서는 추가적인 데이터셋 및 실시간 추적 성능에 대한 분석, 그리고 RandConv 기법의 최적화 방안을 모색할 예정이다.

Acknowledgement

This work was supported by the Korea Institute of Science and Technology (KIST) Institutional Program (Project No.2E33001-24-086).

참고문헌

[1] Kim, D., Park, H., & Lee, J. (2024). *AQATrack: A Query-Adaptive Object Tracking Framework*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 987–995).

[2] Bertinetto, L., Valmadre, J., Henriques, J. F., Torr, P. H. S., & Vedaldi, A. (2016). Fully-Convolutional Siamese Networks for Object Tracking. In Proceedings of the European Conference on Computer Vision (pp. 850–865).

[3] Zhang, Y., Liu, J., & Chen, W. (2023). CTTrack: Context-enhanced Object Tracking via Collaborative Transformers. In Proceedings of the AAAI Conference on Artificial Intelligence (pp. 1123–1130).

[4] Fan, T., Wang, J., & Zhu, M. (2020). *STMTrack: Exploiting Space-Time Memory for Single Object Tracking*. In Proceedings of the AAAI Conference on Artificial Intelligence (pp. 1234–1241).

[5] Liu, Y., Zhu, X., & Lin, S. (2021). RandConv: Random Convolutions for Domain Generalization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (pp. 1013–1022).

[6] Huang, L., Yang, B., & Deng, J. (2019). GOT-10k: A Large High-Diversity Dataset for Tracking Objects in the Wild. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8604–8612).

[7] Mueller, M., Smith, J., & Ghanem, B. (2016). UAV123: A Benchmark Dataset for Object Tracking in Aerial Videos. In Proceedings of the IEEE International Conference on Robotics and Automation (pp. 3589–3596).

[8] Wang, X., Feng, J., & Huang, T. (2019). LSOTB-TIR: A Large-Scale Thermal Infrared Benchmark for Object Tracking. In Proceedings of the IEEE International Conference on Computer Vision Workshops (pp. 450–457).