Evaluating the Robustness of TransTrack in Multi-Object Tracking (MOT)*

*Note: Sub-titles are not captured in Xplore and should not be used

1st Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

2nd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

***Abstract*—This paper first studies on Multi-Object Tracking (MOT) based on TransTrack, then evaluates its robustness under challenging conditions such as lighting changes, snow, and motion blur. By augmenting the MOT17 dataset and applying transfer learning, we assess how these factors impact tracking accuracy. Results highlight key strengths and limitations of TransTrack in real-world scenarios, offering insights for future improvements.**

***Index Terms*—TransTrack, multi-object tracking (MOT), robustness, environmental simulation, transfer learning**

I. Introduction

In the fields of autonomous driving and intelligent surveillance, Multi-Object Tracking (MOT) is a critical technology designed to simultaneously detect and track multiple dynamic targets, such as pedestrians, vehicles, and other objects. The effectiveness of MOT directly impacts the decision-making capabilities and safety of intelligent systems. With the rapid advancement of deep learning, algorithms based on Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have become the dominant approaches in MOT research. [1] These methods learn target features and dynamic behaviors, achieving high accuracy and real-time performance even in complex environments.

Despite the promising performance of modern MOT algorithms, such as TransTrack based on transformer, in ideal settings, significant challenges remain in real-world applications. Adverse weather conditions (e.g., rain, snow, fog), varying lighting conditions (e.g., contrasts between day and night) [2], and motion blur can severely affect tracking accuracy [3]. Research indicates that traditional tracking algorithms often struggle to handle target occlusions, mutual interference, and complex background changes, highlighting the need for enhanced robustness in MOT systems under challenging environmental conditions.

To address these challenges, this project will investigate the TransTrack algorithm, aiming to assess the impact of different environmental factors on tracking performance. We will utilize OpenCV to generate simulated environmental scenarios, including effects of rain, snow, fog, and motion blur, to diversify the training data. [4] This data augmentation strategy will aid in improving the model's adaptability in complex environments.

Additionally, the project will incorporate transfer learning and fine-tuning techniques from this course, training on specific environmental conditions based on a pre-trained model to further enhance the performance of the MOT system. This project aims to experimentally validate and improve the robustness of MOT technology in practical applications.

II. Background and related work

Talking about the Multiple object tracking, it detects and tracks multiple targets in videos, such as pedestrians, vehicles, and animals, while assigning unique IDs to each target, is crucial for subsequent tasks like trajectory prediction and precise retrieval. Each target is assigned a different ID to ensure clear tracking.

In multi-object tracking, transformer-based architectures have been applied across various areas of computer vision, including MOT. The paper TransTrack represents one of the first attempts to utilize transformers in multi-object tracking.

In this field, TransTrack, which utilizes the query-key mechanism from single object tracking, is a model that warrants in-depth study and exploration. The basic Transformer is a neural network architecture based on the self-attention mechanism, originally applied to natural language processing tasks. The self-attention mechanism allows the model to weigh each element in a sequence when processing sequential data, capturing relationships between different elements. TransTrack builds on the Transformer's encoder-decoder architecture, combining both object detection and object tracking.

Compared to traditional multi-stage tracking-by-detection methods, the TransTrack pipeline greatly simplifies multi-object tracking. Traditional methods separate detection and re-identification, leading to increased complexity and missed detections. Simply transferring the query-key mechanism from single-object tracking to multi-object tracking misses new objects, causing detection gaps.

What makes TransTrack unique is its core architecture, which employs two parallel decoders, each handling a different

task. [5] One decoder, the Detection Decoder, is responsible for detecting new objects. Its input consists of object queries learned by the Transformer, which are designed to detect newly appearing objects in the current frame. The Detection Decoder uses these object queries and the global feature map from the encoder, leveraging cross-attention to locate objects in the current frame and output detection results.

The other decoder, the Tracking Decoder, is responsible for tracking objects from the previous frame. It takes track queries as input, which are generated from the objects detected in the previous frame. These queries contain the positional and appearance information of already detected objects. Similarly, the Tracking Decoder uses cross-attention to combine these track queries with the feature map of the current frame, identifying and tracking objects that existed in the previous frame, and outputting their bounding boxes in the current frame.

The encoder in TransTrack takes as input the feature maps of both the current frame and the previous frame. A CNN (Convolutional Neural Network) is used in the backbone to extract the global feature maps of the images. These feature maps are then fed into the Transformer's encoder, which outputs a combined feature map that fuses information from both frames. This fusion enables the subsequent decoders to leverage temporal information, improving the recognition of objects in the current frame and providing valuable data for tracking objects across subsequent frames.

TransTrack uses a simple IoU (Intersection over Union) matching algorithm to combine detection boxes and tracking boxes. [6] This process employs the Hungarian algorithm to calculate the IoU values between each detection box and tracking box, and based on these values, objects are matched to produce the final tracking result. Specifically, the calculation of IoU is as follows:

$$IoU = \frac{\text{Area of Intersection}}{\text{Area of Union}} \tag{1}$$

The value of IoU ranges from 0 to 1, where 1 indicates that the two bounding boxes are perfectly aligned, and 0 means there is no overlap.

While learning and replicating it, we also tested its robustness. Since TransTrack is a multi-object tracking method designed for scenarios with high real-time requirements, our approach is to test its robustness. The main method involves using OpenCV to simulate different conditions in various scenes, such as snowy weather, lighting effects, and motion blur.

III. Methodology

A. Complex scenario simulation and data augmentation

To simulate various real-world challenges in object detection and tracking, a series of effects were applied to video frame sequences using OpenCV. (Visual in Fig.2) These augmentations simulate different lighting conditions, motion blur, and weather effects like snow. Each effect is applied to individual frames to create complex scenarios that test the robustness of tracking algorithms.

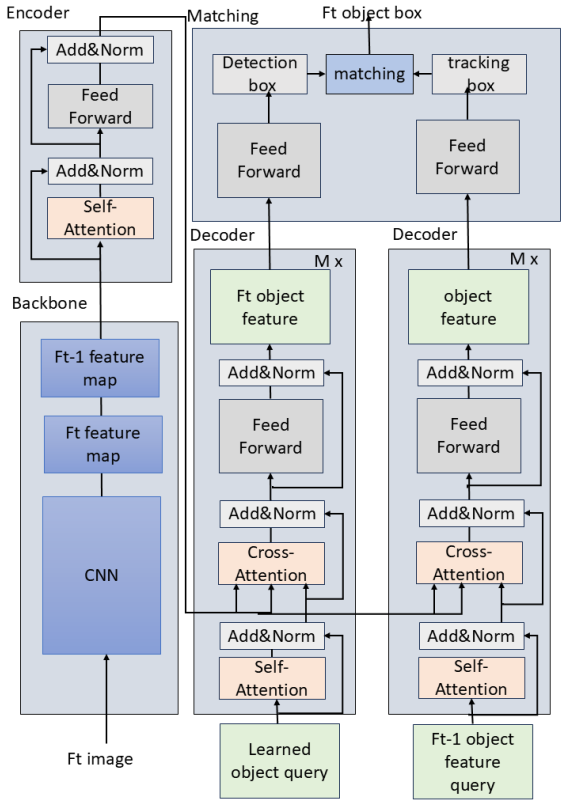

Fig. 1. The Basic Architecture of TransTrack

Lighting and Shadow Effects: Random brightness adjustments (±100) are introduced along with simulated shadow occlusions. A random rectangular region is darkened with a shadow intensity of 0.8, mimicking real-world challenges such as changing lighting conditions and partial occlusions caused by obstacles like trees or buildings.

Motion Blur Effects: Motion blur is simulated by applying a blur kernel with intensity 15 to each frame. This mimics the effect of fast-moving objects, requiring the tracking system to handle blurred or distorted object appearances.

Snow Effects: A snow effect is simulated by randomly generating white circles of varying sizes across the image, with a density of 0.01 to represent falling snow. This tests the tracking system's ability to maintain accuracy under weather conditions that obscure visibility.

Table

Sequence	Effect	Length	dataloader
MOT17-02	None	300	val
MOT17-04	Snow	1050	train
MOT17-05	Light	300	val
MOT17-09	Blur	625	train
MOT17-10	Blur	300	val
MOT17-11	Light	900	train
MOT17-13	Snow	300	val

TABLE I
AUGMENTED DATASET

Fig. 2. Visual effects of environmental changes: top left is the original image, top right shows lighting variations, bottom left depicts motion blur, and bottom right represents snowy conditions.

B. Transfer Learning and Fine Tuning

Following the augmentation of video sequences with lighting, motion blur, and snow effects, transfer learning techniques are employed to enhance the performance of multi-object tracking systems. The Transtrack model is selected for this purpose due to its robustness in handling complex tracking scenarios. The pre-trained Transtrack model is fine-tuned using the augmented datasets to adapt to the specific challenges posed by the simulated environments. This process involves retraining the model on the newly created datasets, allowing it to learn from both the original and augmented images, improving its accuracy and resilience against occlusions and varying environmental conditions. During fine-tuning, specific hyperparameters are utilized:

Learning Rate	Batch Size	Epochs	Early Stopping
0.0001	2	20	5 epochs

TABLE II
HYPER PARAMETERS

In object tracking tasks, a minimum batch size of 2 is required to capture temporal associations between consecutive frames. This allows the model to observe object movements and calculate differences in position, size, and appearance, which are essential for learning dynamic features and calculating tracking errors.

To prevent overfitting and maintain the integrity of the learned features, certain layers are frozen during the training process. The following layers are typically frozen:

- **Backbone Layers:** The initial layers responsible for feature extraction are frozen to retain their learned weights.
- **Intermediate Layers:** Some intermediate layers that capture more general features are also frozen, while higher layers, which are more task-specific, remain trainable.

The fine-tuning process is optimized through techniques such as learning rate scheduling and data augmentation, ensuring that the model can effectively track multiple objects under diverse and challenging conditions.

C. Evaluation criteria

- **IDF1 (ID F1 Score)**
  The harmonic mean of ID precision and ID recall, measuring the balance between correctly tracked objects and total tracked objects.
$$\text{IDF1} = \frac{2 \times \text{IDP} \times \text{IDR}}{\text{IDP} + \text{IDR}} \tag{2}$$
- **Recall (Rcll)**
  The ratio of correctly tracked objects to the total number of ground truth objects.
$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$
- **MOTA (Multiple Object Tracking Accuracy)**
  A comprehensive metric that accounts for false positives, false negatives, and identity switches in tracking results.
$$\text{MOTA} = 1 - \frac{\text{FP} + \text{FN} + \text{ID Sw}}{\text{GT}} \tag{4}$$
- **MOTP (Multiple Object Tracking Precision)**
  The average distance between predicted and ground truth locations of tracked objects, focusing on localization accuracy.
$$\text{MOTP} = \frac{\sum_i d_i}{\text{TP}} \tag{5}$$
  where d_i is the distance for each correctly predicted object.
- **IDP (ID Precision)**
  The ratio of correctly tracked objects to all objects tracked (including false positives).
$$\text{IDP} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{6}$$

Notes:

- **TP**: True Positives
- **FP**: False Positives
- **FN**: False Negatives
- **ID Sw**: Identity Switches
- **GT**: Total Ground Truth Objects

IV. RESULTS AND DISCUSSION

A. Original performance on MOT17

Sequence	IDF1	IDP	Recall	MOTA
MOT17-02	46.4%	75.3%	41.9%	38.8%
MOT17-05	59.3%	71.6%	68.1%	64.9%
MOT17-10	61.3%	69.4%	69.5%	59.1%
MOT17-13	72.2%	75.0%	78.9%	63.4%

TABLE III
EVALUATION METRICS FOR ORIGINAL MOT17 SEQUENCES

B. Performance on MOT17 with different effects

Sequence	IDF1	IDP	Recall	MOTA
MOT17-02-origin	46.4%	75.3%	41.9%	38.8%
MOT17-05-light	41.6%	54.1%	59.9%	55.6%
MOT17-10-blur	50.5%	79.2%	41.8%	36.7%
MOT17-13-snow	38.1%	47.7%	50.0%	30.0%

TABLE IV
EVALUATION METRICS FOR MOT17 SEQUENCES WITH DIFFERENT EFFECTS

The performance comparison across the three scenarios reveals distinct challenges: under varying lighting conditions, target recognizability diminishes due to changes in appearance, resulting in lower accuracy; in scenarios with motion blur, tracking consistency is significantly affected, leading to a higher rate of missed detections; and during snowy weather, the complex background and obscured target features result in a substantial decrease in overall detection precision. Each of these factors contributes to increased false negatives, highlighting the need for robust model training to address these environmental challenges effectively.

Here are some examples of the challenges encountered under different effects: (Fig.3-5)

Fig. 3. Visual of missing under light effect

Fig. 4. Visual of ID switch under motion blur

Fig. 5. Visual of missing under snow weather

C. Performance on MOT17 with Different Effects After Transfer Learning and Fine Tuning

Sequence	IDF1	IDP	Recall	MOTA
MOT17-02-origin	45.8%	71.5%	43.1%	38.6%
MOT17-05-light	44.6%	57.3%	60.5%	55.3%
MOT17-10-blur	51.8%	65.6%	52.5%	39.1%
MOT17-13-snow	43.2%	44.9%	60.7%	24.5%

TABLE V
EVALUATION METRICS FOR MOT17 SEQUENCES WITH DIFFERENT EFFECTS

- **MOT17-02-origin:** The model maintained a stable IDF1 of 45.8%, slightly down from 46.4% before fine-tuning, with a MOTA of 38.6%, indicating effective target identification under standard conditions.
- **MOT17-05-light:** The IDF1 improved from 41.6% to 44.6%, demonstrating better target identification in challenging visibility; however, MOTA declined to 55.3%, suggesting a need for further refinement to maintain tracking consistency.
- **MOT17-10-blur:** The model showed improvement with IDF1 rising from 50.5% to 51.8% and MOTA increasing from 36.7% to 39.1%, indicating effective adaptation to motion blur challenges due to transfer learning.
- **MOT17-13-snow:** The IDF1 improved from 38.1% to 43.2%, but MOTA dropped to 24.5%, highlighting capability for target identification while revealing significant tracking challenges in adverse weather.

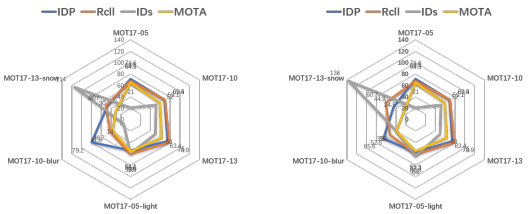

Fig. 6. Performance before(left) and after(right) transfer learning

V. CONCLUSION

REFERENCES

[1] G. Ciaparrone, F. Luque Sánchez, S. Tabik, L. Troiano, R. Tagliaferri, and F. Herrera, "Deep learning in video multi-object tracking: A survey," *Neurocomputing*, vol. 381, pp. 61-88, 2020. doi: 10.1016/j.neucom.2019.11.023. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] H. Gupta, O. Kotlyar, H. Andreasson and A. J. Lilienthal, "Robust Object Detection in Challenging Weather Conditions," in 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 2024, pp. 7508-7517, doi: 10.1109/WACV57701.2024.00735.

[3] Q. Guo, Z. Cheng, F. Juefei-Xu, L. Ma, X. Xie, Y. Liu, and J. Zhao, "Learning to adversarially blur visual object tracking," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10839-10848, 2021.

[4] P. Sun, J. Cao, Y. Jiang, R. Zhang, E. Xie, Z. Yuan, C. Wang, and P. Luo, "TransTrack: Multiple Object Tracking with Transformer," arXiv, arXiv:2012.15460, May 2021. doi: 10.48550/arXiv.2012.15460.

[5] OpenCV, "OpenCV: Open Source Computer Vision Library," [Online]. Available: https://opencv.org/.

[6] Ren, Shaoqing, et al, "Faster R-CNN: Towards real-time object detection with region proposal networks," Adv. Neural Inf. Process. Syst., vol. 28, pp. 91-99, 2015.

[7] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[8] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.