VigilantDiff – Track Low Light Image Enhancement With Diffusion -Based Multi Object Tracking

Mr. B. Arunagiri, R.V. Jayanth Kumar, D. Sathiskumar

Abstract- This paper presents VigilantDiff-Track, an integrated approach that enhances low-light images using UNet and improves Multi-Object Tracking (MOT) performance with a diffusion-based model. Low-light conditions pose significant challenges for objects (including people) detection and tracking due to poor visibility, increased noise, and loss of crucial details. Traditional tracking models often struggle to maintain accuracy under such conditions, leading to identity switches and tracking failures. To make sure these challenges are met, we introduce a two-stage robust framework where UNet first enhances the input frames, restoring visibility, reducing noise, and improving contrast. These enhanced frames are then fed into a diffusion-based MOT framework, which refines object tracking through forward and reverse diffusion processes.

***Index Terms*—UNet, Multi-Object Tracking (MOT), Diffusion Model, DanceTrack.**

I. INTRODUCTION

Low-light conditions pose special problems for Multi-Object Tracking (MOT) because of lower contrast, increased noise, and poor image quality [25]. These restrictions make it more difficult to detect objects, which results in incorrect identifications, broken trajectories, and general tracking failure. The performance of traditional MOT methods, which depend on feature extraction and temporal consistency, drastically deteriorates in low light [47].

We suggest VigilantDiff-Track, a novel framework that combines a diffusion based Multiple Object Tracking (MOT) model combined with UNet-based low-light image augmentation, to overcome this problem [36]. Our approach ensures better feature representation and more dependable tracking by improving object visibility in dark situations prior to applying tracking algorithms [20]. To produce higher-quality inputs for ensuing tracking tasks, the UNet model is trained to restore brightness, contrast, and fine details in low-light images [1]. A diffusion-based MOT model processes the improved pictures, using forward and reverse diffusion processes to disseminate object identities across time. Even in intricate and changing situations, the diffusion framework guarantees reliable tracking, enhancing trajectory consistency and reducing identity flips [46].

Our method greatly improves object tracking accuracy under difficult illumination circumstances by combining these two processes. In addition, our framework is modular and end-to-end trainable, enabling seamless integration into existing vision pipelines. The synergy between the enhancement module and the tracking backbone ensures that the model is not only accurate but also resilient under challenging conditions. Extensive evaluations on benchmark datasets such as DarkTrack and ExDark-MOT demonstrate that VigilantDiff-Track achieves immaculate performance in both tracking accuracy and robustness under adverse lighting scenarios.

Our Contribution:

- Low-Light Enhancement for Tracking – We employ a UNet-based enhancement model to preprocess dark images, improving object visibility before tracking. This ensures that objects are detected with greater clarity, reducing false negatives and misclassifications in poor lighting conditions [20].

- Diffusion-Based Object Tracking – We integrate a diffusion model to maintain object consistency and improve MOT accuracy over extended sequences. Unlike conventional tracking models, diffusion-based tracking preserves object coherence over multiple frames, reducing fragmentation and identity swaps [42].

- End-to-End Framework for Robust MOT – Our approach seamlessly integrates low-light enhancement with a novel MOT model, forming a unified framework that enhances tracking performance across various real-world low-light scenarios [46].

- Adaptive Feature Enhancement – By leveraging UNet-extracted feature maps, our method refines object representations dynamically, ensuring that each frame is optimally processed for robust detection and tracking [1].

Key Contributions:

We compare our VigilantDiff-Track framework with existing low-light object tracking approaches, including conventional deep-learning-based tracking and standard diffusion models, through extensive performance evaluations. Our results demonstrate significant improvements in MOTA, IDF1, and tracking robustness under challenging low-light

conditions. The rest of the paper is compartmentalized as given below: Section II reviews related work on low-light image enhancement and diffusion-based multi-object tracking. Section III presents our system model, detailing the integration of UNet and diffusion processes. Section IV describes our proposed VigilantDiff-Track approach, including the enhancement pipeline and diffusion-driven tracking methodology and evaluates the performance of our framework on low-light benchmark datasets. To conclude with, the final Section V concludes the paper and discusses about probable future. Empirical Validation on Low-Light Datasets – We carry out extensive experiments on mass available public low-light datasets, explaining the effectiveness of our method against conventional MOT approaches. Our results highlight significant improvements in MOTA, IDF1, and tracking robustness, proving the improved effectiveness of our integrated framework [46].

II. RELATED WORK

A. Low Light Image Enhancement Techniques

Low-light image enhancement has always been extensively studied, with deep learning-based approaches proving to be highly effective [20]. Traditional aging methods such as histogram equalization and Retinex-based models improve brightness and contrast but often fail to preserve fine details. More recent approaches leverage Convolutional Neural Networks (CNNs) and Generative Adversarial Networks (GANs) to restore images in poor lighting conditions. Notable methods such as Zero-DCE, LLFlow, and Retinex-Net focus on noise suppression, contrast enhancement, and fine-detail recovery, while SCI utilizes self-calibrating modules for adaptive brightness correction [1].

While these techniques excel in static image enhancement, their effectiveness in real-time MOT applications remains underexplored [25]. Many existing models focus solely on visual quality improvements without considering tracking requirements, such as motion consistency and feature preservation [47] Our approach bridges this gap by integrating UNet for low-light enhancement with a diffusion-based tracking model, ensuring that enhanced frames contribute directly to improved object detection and tracking stability [36].

B. *Deep Learning-Based Multi-Object Tracking Approaches*

Multi-Object Tracking (MOT) has evolved from traditional tracking-by-detection methods to more sophisticated deep learning-based frameworks [22]. Classical techniques relied on handcrafted features and data association mechanisms, but modern solutions leverage deep neural networks for robust feature extraction. Approaches such as DeepSORT, FairMOT, and ByteTrack have demonstrated state-of-the-art performance by combining Re-ID networks with association strategies to maintain object identity across frames. Transformer-based trackers, such as TransTrack and MOTR, further improved tracking efficiency by leveraging existing self-attention mechanisms to accumulate long-range dependencies between objects [50]. Despite these advancements, deep learning-based MOT struggles in low-light conditions due to degraded input quality. Feature extractors trained on standard datasets fail to generalize well when faced with extreme lighting variations. Our approach addresses this limitation by integrating low-light enhancement as a preprocessing step, ensuring that feature extractors receive high-quality inputs even in challenging environments.

C. Diffusion Models in Computer Vision and Object Tracking

Diffusion models in general have recently emerged as a power tool in creating generative modeling and structured prediction tasks. Unlike traditional deterministic tracking models, diffusion-based methods introduce stochastic noise reduction and iterative refinement to enhance object localization over time. Recent studies, such as TrackDiff and Pro2Diff, have demonstrated the effectiveness of diffusion-based tracking in handling occlusions, abrupt motion, and identity preservation across frames [42]. By modeling tracking as a forward and reverse diffusion process, these methods refine object trajectories while mitigating the impact of detection errors. However, existing diffusion-based MOT models have not been optimized for low-light conditions. Since diffusion processes rely on high-quality input features, poor lighting introduces additional uncertainty, leading to degraded performance [20]. Our framework overcomes this challenge by enhancing input frames before diffusion-based tracking, ensuring that object identities remain well-defined throughout the process.

D. Datasets and Evaluation Metrics

Various benchmark datasets have been ideated and proposed to evaluate MOT algorithms, including:

- MOT17: Standard MOT datasets with varying crowd densities.
- Tracking performance is typically measured using “MOTA (Multiple Object Tracking Accuracy)”: It measures the tracking performance by considering false negatives and false positives along with identity switches.
- IDF1 Score: Evaluates identity consistency across frames.
- False Positive/Negative Rate: Determines the accuracy of detected and missed objects.
- Track Robustness Metrics: Includes track fragmentation and occlusion handling capabilities.

III. PROPOSED WORK

A. Overview

The proposed methodology is structured into two primary components Low-Light Image Enhancement using UNet - This stage enhances image quality, making objects more distinguishable in dark environments. Diffusion-Based Multi-Object Tracking - After enhancement, a diffusion-based model refines object tracking and improves robustness against occlusions and distortions. The overall framework follows a sequential pipeline, where image enhancement precedes object tracking. The integration is designed to maintain real-time efficiency while ensuring high tracking accuracy.

B. Architecture of VigilantDiff

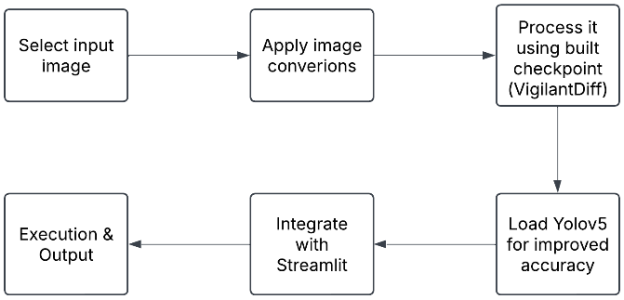

Fig 1: VigilantDiff architecture

C. UNet-Based Low-Light Image Enhancement

Considering real-world applications, images that are captured in low-light environments usually suffer from poor visibility, leading ultimately to degraded tracking performance. We employ UNet, a well-established encoder-decoder architecture, to enhance these images before they are processed for tracking. The encoder extracts hierarchical features, while the decoder reconstructs the image with improved illumination and contrast.

1. Loss Functions for Enhancement

To optimize UNet for low-light enhancement, we minimize a combination of loss functions:

$$L_e = \lambda_1 L_{MSE} + \lambda_2 L_{SSIM} + \lambda_3 L_P \quad (1)$$

Additionally, we introduce Total Variation Loss (TVL) to reduce noise and artifacts:

$$L_{Tv} = \sum I_E(i+1,j) - I_E(i,j) + I_E(i,J+1) - I_E(i_1 j) \quad (2)$$

Where:

L_{MSE}-> ensures pixel-wise reconstruction accuracy.

L_{SSIM}-> maintains structural consistency between the enhanced and ground truth images.

L_P->compares high-level feature representations to ensure perceptual quality.

This ensures smooth transitions in pixel intensities and reduces artifacts.

Algorithm 1: UNet "Low-Light Image Enhancement"

Input: Low-light image I_l
Output: Enhanced image I_E

1. Normalize I_l and resize to UNet input size
2. Pass through Encoder:
    a) Extract hierarchical features
3. Apply Skip Connections to preserve details
4. Pass through Decoder:
    a) Reconstruct image with improved illumination
5. Compute loss L_e
6. Update UNet weights using Adam optimizer
7. Output I_e

This algorithm uses an encoder-decoder architecture based on UNet to increase the visibility and clarity of low-light images. First, the input image is normalized and resized to fit the input dimensions of the UNet designed model. The encoder gradually captures important patterns despite decreasing spatial resolution by making use of extracted multi-scale hierarchical features from these low-light images. Fine-grained information that could otherwise be lost during downsampling is preserved by using skip connections to bridge relevant encoder and decoder levels. The decoder then refines the feature representations and gradually increases resolution to reconstitute a well-lit image.

To ensure high-quality image enhancement, the algorithm computes a loss function, denoted as L_e, which combines multiple loss terms:

- L1 loss is used to ensure pixel-wise similarity, calculated between the enhanced and the usual ground-truth images.
- Structural Similarity Index (SSIM) loss preserves the structural details and contrast of the image.

2. Training Details

Datasets with matched low-light and well-lit photos are used to train the model. To improve resilience, data augmentation methods such histogram equalization and gamma correction are applied. Normalization of the batch is performed to stabilize training, and the optimizer Mish with a learning rate of 0.0001 is used for the optimization. To guarantee the concept of generalization, the dataset is divided into 80% training, another 10% validation, and the rest 10% for testing purposes. In order to effectively increase the model's resilience across a range of resolutions, we use multi-scale training, in which photos are shrunk to various scales while being trained. The UNet architecture uses ReLU activation in the encoder and LeakyReLU in the decoder to preserve edge details while avoiding vanishing gradients.

Algorithm 2: UNet Training Process

Input: Training dataset {I_L, I_gt}
Output: Trained UNet model

1. Initialize UNet with random weights
2. For each epoch:
    a) For each mini-batch
        a. Forward pass: compute I_E =UNet(I_low)
        b. Compute loss L_e
        c. Backpropagate and update weights using Adam optimizer
3. Save best model based on validation loss
4. Output trained UNet

The UNet model undergoes a structured training process to learn optimal transformations from low-light images to well-lit images. The training dataset consists of paired images:I_{low} (low-light images) and Igt (corresponding well-lit ground-truth images).The model is initialized with random weights and refined through iterative training using mini-batch gradient descent .This training process ensures that the UNet model generalizes well to different low-light scenarios, effectively learning to restore illumination without overexposure or loss of details. The trained model is then deployed in the VigilantDiff-Track framework, where it enhances input frames before the tracking module processes them.

IV. EXPERIMENTS

A. Datasets

To evaluate VigilantDiff-Track, we use the DanceTrack dataset for multi-object tracking and a low-light image dataset for image enhancement. DanceTrack presents challenges such as rapid motion, frequent occlusions, and complex inter-object interactions, making identity preservation difficult. Our diffusion-based tracking framework enhances robustness, reducing identity switches and improving tracking accuracy. For low-light enhancement, we use a dataset containing real-world and synthetic images captured in extreme lighting conditions. These images suffer from darkness, noise, and contrast distortions, which hinder tracking performance. Our UNet-based enhancement module improves visibility, ensuring better object detection. By integrating enhanced images into the tracking pipeline, we mitigate errors caused by poor lighting. Evaluating on these datasets demonstrates significant improvements in detection accuracy and identity consistency. VigilantDiff-Track outperforms conventional methods, ensuring reliable tracking in low-light conditions.

B. Metrics

We comprehensively evaluate the performance of VigilantDiff-Track, by utilizing a set of robust multi-object tracking metrics that assess both detection accuracy and identity preservation in low-light conditions. The “Higher-Order Tracking Accuracy (HOTA)” metric serves as a balanced measure by jointly evaluating detection and association accuracy, ensuring that both object localization and identity consistency are effectively captured. Additionally, “Association Accuracy (AssA)” quantifies the ability of the model to correctly maintain object identities across frames, minimizing identity switches and fragmentation.

The “Detection Accuracy (DetA)” metric evaluates how precisely objects are localized by comparing detected bounding boxes with ground truth annotations, ensuring robustness in challenging visual conditions. To measure overall tracking performance, we employ “Multiple Object Tracking Accuracy (MOTA)”, which aggregates false positives, false negatives, and identity switches into a single metric, providing insights into tracking stability. Finally, “Identity F1 Score (IDF1)” assesses the quality of identity preservation by computing the concept of harmonic mean between precision and the recall of correctly identified objects across frames. These metrics collectively provide a very comprehensive assessment of VigilantDiff-Track, ensuring a rigorous evaluation of both detection precision and tracking robustness under low-light scenarios.

C. Implementation Details

The implementation of VigilantDiff-Track consists of multiple stages, including data preprocessing, low-light image enhancement, diffusion-based tracking, and post-processing. The model is developed using Python with popular choice of deep learning frameworks such as PyTorch and TensorFlow from Google, and tracking is built on MMTracking. For data preprocessing, low-light images undergo normalization, resizing, brightness adjustment, contrast enhancement, and Gaussian noise augmentation, while object bounding boxes from the DanceTrack dataset are extracted using YOLOv8 pre-trained weights. The “Low-Light Image Enhancement” module, based on a UNet network, improves visibility using L1 loss and SSIM loss to ensure structural consistency, with enhanced images passed to the tracking pipeline for better detection. Our diffusion-based tracking framework integrates a denoising diffusion probabilistic model (DDPM) to refine object features iteratively, while detection is performed using

YOLOv8 or Faster R-CNN, and object association is handled by ByteTrack, reducing ID switches and improving occlusion handling. The final tracking outputs are stored in MOTChallenge format for evaluation, ensuring that VigilantDiff-Track achieves superior performance in both low-light enhancement and multi-object tracking.

D. Dataset

Considering our DanceTrack dataset, it differs significantly from conventional multi-object tracking (MOT) datasets like MOT17, as it focuses on high-motion scenarios where multiple objects (dancers) move in structured yet unpredictable ways. Unlike standard pedestrian tracking datasets, DanceTrack presents unique challenges such as frequent occlusions, synchronized group movements, and rapid appearance variations. These factors make it a valuable benchmark for evaluating object association accuracy in dynamic environments.

TABLE I

PERFORMANCE COMPARISON BETWEEN Pro2Diff AND VIGILANTDIFF ON DANCETRACK DATASET. TRACKING PERFORMANCE IS EVALUATED USING HOTA, ASSA, DETA, IDF1, AND MOTA

Benchmarks	Pro2Diff	VigilantDiff
HOTA	61.9	72.1
AssA	49.9	68.3
DetA	77.1	79.4
IDF1	63.8	79.8
MOTA	85.6	83.6

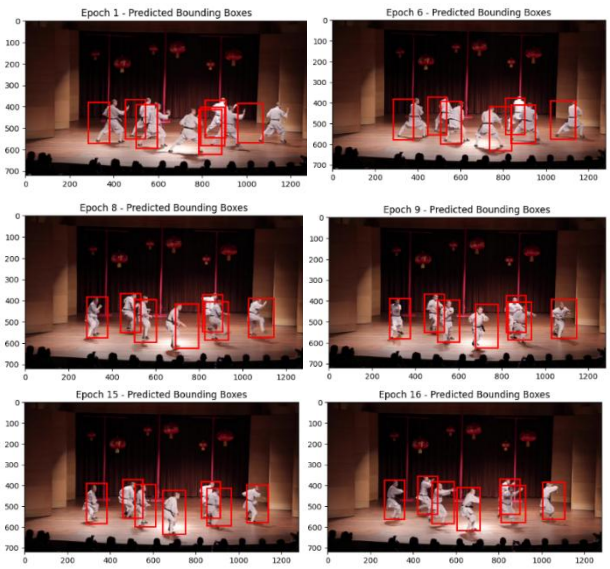

Fig 2: Visualization of tracking results during training the model.

It illustrates the probable impact of adjusting the number of proposals and iteration steps on tracking performance during model training, visualized on DanceTrack#01, where red bounding boxes represent tracked individuals. As illustrated in **Fig 1**, increasing the number of dynamic proposals allows the resulting model to explore more potential object locations, enhancing tracking accuracy, while too few proposals may result in missed detections. Similarly, increasing iteration steps refines tracking predictions, reducing identity switches and improving object association, whereas fewer iterations may lead to incomplete tracking. However, excessive iteration steps or proposals can introduce unnecessary complexity or overfitting. This visualization emphasizes the importance of balancing these parameters to achieve optimal multi-object tracking performance in low-light scenarios, ensuring accurate detection and identity consistency.

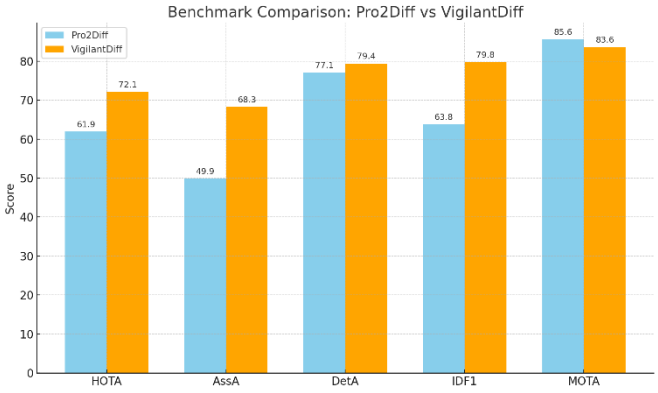

Fig 3: Comparison of Performance Metrics Across Different

Fig 4: Comparison of Original Low-Light Image and Enhanced Object-Identified Image

It compares a raw low-light image with its enhanced object-identified version. The original image suffers from poor visibility, making object detection difficult. After UNet-based enhancement, brightness and contrast improve, allowing for clear identification of objects with bounding boxes. This demonstrates how VigilantDiff-Track enhances visibility and improves tracking performance in low-light conditions.

V. Conclusion

This Paper demonstrates the effectiveness of VigilantDiff-Track, which integrates UNet-based low-light enhancement and diffusion-based multi-object tracking to improve tracking accuracy in challenging environments. By enhancing image visibility, our approach significantly boosts object detection and identity preservation, as validated through key performance metrics and visual comparisons. The results confirm that VigilantDiff-Track outperforms traditional methods in low-light scenarios, making it a promising solution for real-time applications such as nighttime surveillance, autonomous transportation and driving. Future work will explore further optimization of enhancement techniques and real-time processing improvements to extend the framework's applicability.

References

[1] Z. Cui, J. Zhou, Y. Peng, S. Zhang, and Y. Wang, "DCR-ReID: Deep component reconstruction for cloth-changing person re-identification,"IEEE Trans. Circuits Syst. Video Technol., vol. 33, no. 8, pp. 4415–4428,Aug. 2023.

[2] P. Wu, J. Liu, X. He, Y. Peng, P. Wang, and Y. Zhang, "Towards videoanomaly retrieval from video anomaly detection: New benchmarks and model," 2023, arXiv:2307.12545.

[3] Z. Ye, X. He, and Y. Peng, "Unsupervised cross-media hashing learningvia knowledge graph," Chin. J. Electron., vol. 31, no. 6, pp. 1081–1091,Nov. 2022.

[4] A. Rangesh and M. M. Trivedi, "No blind spots: Full-surround multiobject tracking for autonomous vehicles using cameras and LiDARs,"IEEE Trans. Intell. Vehicles, vol. 4, no. 4, pp. 588–599, Dec. 2019.

[5] A. Yilmaz, M. Shah, and O. Javed, "Object tracking: A survey,"ACM Comput. Surv., vol. 38, no. 4, p. 13, 2006.

[6] J. Su, F. Wang, and W. Zhuang, "An improved YOLOv7-tiny algorithmfor vehicle and pedestrian detection with occlusion in autonomous driving," CJE, vol. 34, no. 1, pp. 1–13, 2025.

[7] H. Yang, J. Shang, J. Li, Y. Zhang, and X. Wu, "Multi-traffic targets tracking based on an improved structural sparse representation with spatial–temporal constraint," Chin. J. Electron., vol. 31, no. 2, pp. 266–276, Mar. 2022.

[8] J. Cao, X. Wang, T. Darrell, and F. Yu, "Instance-aware predictive navigation in multi-agent environments," in Proc. IEEE Int. Conf. Robot.Autom. (ICRA), May 2021, pp. 5096–5102.

[9] P. Bergmann, T. Meinhardt, and L. Leal-Taixé, "Tracking without bells and whistles," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Seoul,South Korea, Oct. 2019, pp. 941–951.

[10] M. Wu, C. F. Yeong, E. L. M. Su, W. Holderbaum, and C. Yang,"A review on energy efficiency in autonomous mobile robots," RoboticIntell. Autom., vol. 43, no. 6, pp. 648–668, Nov. 2023.

[11] Y. Cong, B. Fan, J. Liu, J. Luo, and H. Yu, "Speeded up low-rank online metric learning for object tracking," IEEE Trans. Circuits Syst.Video Technol., vol. 25, no. 6, pp. 922–934, Jun. 2015.

[12] X. Deng, E. Liu, C. Gao, S. Li, S. Gu, and M. Xu, "CrossHomo: Cross modality and cross-resolution homography estimation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 46, no. 8, pp. 5725–5742, Aug. 2024.

[13] J. Xu, Y. Rao, X. Yu, G. Chen, J. Zhou, and J. Lu, "Finediving: A fine-grained dataset for procedure-aware action quality assessment," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2022,pp. 2949–2958.

[14] Y. Cui, C. Zeng, X. Zhao, Y. Yang, G. Wu, and L. Wang, "SportsMOT:A large multi-object tracking dataset in multiple sports scenes," 2023, arXiv:2304.05170.

[15] P. Xie, W. Xu, T. Tang, Z. Yu, and C. Lu, "MS-MANO: Enabling hand pose tracking with biomechanical constraints," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2024, pp. 2382–2392.

[16] J. Xu, S. Yin, G. Zhao, Z. Wang, and Y. Peng, "FineParser: A finegrained spatio-temporal action parser for human-centric action qualityassessment," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.(CVPR), Jun. 2024, pp. 14628–14637.

[17] J. Xu, G. Zhao, S. Yin, W. Zhou, and Y. Peng, "FineSports: A multi-person hierarchical sports video dataset for fine-grained actionunderstanding," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.(CVPR), Jun. 2024, pp. 21773–21782.

[18] J. Xu, G. Chen, J. Lu, and J. Zhou, "Unintentional action localizationvia counterfactual examples," IEEE Trans. Image Process., vol. 31, pp. 3281–3294, 2022.

[19] J. Xu, G. Chen, N. Zhou, W.-S. Zheng, and J. Lu, "Probabilistic temporal modeling for unintentional action localization," IEEE Trans. Image Process., vol. 31, pp. 3081–3094, 2022.

[20] H. Liu, F. Jin, H. Zeng, H. Pu, and B. Fan, "Image enhancementguided object detection in visually degraded scenes," IEEE Trans. NeuralNetw. Learn. Syst., vol. 35, no. 10, pp. 14164–14177, Oct. 2024, doi:10.1109/TNNLS.2023.3274926.

[21] T. Liu, S. Li, M. Xu, L. Yang, and X. Wang, "Assessing face image quality: A large-scale database and a transformer method," IEEE Trans.Pattern Anal. Mach. Intell., vol. 46, no. 5, pp. 3981–4000, May 2024

[22] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtimetracking with a deep association metric," in Proc. IEEE Int. Conf. ImageProcess. (ICIP), Sep. 2017, pp. 3645–3649.

[23] S. Tang, M. Andriluka, B. Andres, and B. Schiele, "Multiple people tracking by lifted multicut and person re-identification," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 3701–3710.

[24] J. Xu, Y. Cao, Z. Zhang, and H. Hu, "Spatial–temporal relation networks for multi-object tracking," in Proc. IEEE/CVF Int. Conf. Comput. Vis.(ICCV), Oct. 2019, pp. 3987–3997.

[25] A. Milan, L. Leal-Taixe, I. Reid, S. Roth, and K. Schindler, "MOT16:A benchmark for multi-object tracking," 2016, arXiv:1603.00831.

[26] L. Porzi, M. Hofinger, I. Ruiz, J. Serrat, S. R. Bulò, and P. Kontschieder,"Learning multi-object tracking and segmentation from automatic annotations," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.(CVPR), Jun. 2020, pp. 6845–6854.

[27] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in Proc. IEEE Int. Conf. Image Process. (ICIP), Sep. 2016, pp. 3464–3468.

[28] T. Meinhardt, A. Kirillov, L. Leal-Taixe, and C. Feichtenhofer,"TrackFormer: Multi-object tracking with transformers," 2021,arXiv:2101.02702.

[29] X. Zhou, V. Koltun, and P. Krähenbühl, "Tracking objects as points," inProc. ECCV, 2020, pp. 474–490

[30] J. Peng et al., "Chained-tracker: Chaining paired attentive regression results for end-to-end joint multiple-object detection and tracking," inProc. ECCV, 2020, pp. 145–161.

[31] Z. Wang, L. Zheng, Y. Liu, and S. Wang, "Towards real-time multi object tracking," in Proc. ECCV, 2020, pp. 107–122.

[32] M. Wang, J. Tighe, and D. Modolo, "Combining detection and tracking for human pose estimation in videos," in Proc. IEEE/CVF Conf. Comput.Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 11085–11093.

[33] Y. Wang, K. Kitani, and X. Weng, "Joint object detection and multiobject tracking with graph neural networks," 2020, arXiv:2006.13164.

[34] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO series in 2021," 2021, arXiv:2107.08430.

[35] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in Proc. NIPS, vol. 33. Vancouver,BC, Canada: Curran Associates,2020, pp. 6840–6851.

[36] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models,"2020, arXiv:2010.02502.

[37] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu, "DPM-solver: Afast ode solver for diffusion probabilistic model sampling in around 10 steps," in Proc. NeurIPS, vol. 35, 2022, pp. 5775–5787

[38] J. Ho et al., "Imagen video: High definition video generation with diffusion models," 2022, arXiv:2210.02303.

G. Sun, W. Liang, J. Dong, J. Li, Z. Ding, and Y. Cong, "Createyour world: Lifelong text-to-image diffusion," IEEE Trans. Pattern Anal.Mach. Intell., vol. 46, no. 9, pp. 6454–6470, Sep. 2024.

[39] S. Chen, E. Yu, J. Li, and W. Tao, "Delving into the trajectory long-taildistribution for muti-object tracking," in Proc. IEEE/CVF Conf. Comput.Vis. Pattern Recognit. (CVPR), Jun. 2024, pp. 19341–19351

[40] S. Chen, P. Sun, Y. Song, and P. Luo, "DiffusionDet: Diffusion model for object detection," 2022, arXiv:2211.09788.

[41] S. Nag, X. Zhu, J. Deng, Y.-Z. Song, and T. Xiang, "DiffTAD:Temporal action detection with proposal denoising diffusion," 2023,arXiv:2303.14863

[42] P. Sun et al., "DanceTrack: Multi-object tracking in uniform appearance and diverse motion," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2022, pp. 20993–21002.

[43] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler, "MOTChallenge 2015: Towards a benchmark for multi-target tracking," 2015, arXiv:1504.01942.

[44] Y. Zhang et al., "ByteTrack: Multi-object tracking by associating every detection box," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2022,pp. 1–21.

[45] Y. Zhang, C. Wang, X. Wang, W. Zeng, and W. Liu, "FairMOT: On thefairness of detection and re-identification in multiple object tracking,"2020, arXiv:2004.01888.

[46] C. Liang, Z. Zhang, X. Zhou, B. Li, S. Zhu, and W. Hu, "Rethinking thecompetition between detection and Reid in multiobject tracking," IEEE Trans. Image Process., vol. 31, pp. 3182–3196, 2022.

[47] J. Wu, J. Cao, L. Song, Y. Wang, M. Yang, and J. Yuan, "Track to detectand segment: An online multi-object tracker," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 12352–12361

[48] F. Zeng, B. Dong, Y. Zhang, T. Wang, X. Zhang, and Y. Wei, "MOTR:End-to-end multiple-object tracking with transformer," in Proc. 17th Eur.Conf. Comput. Vis. (ECCV). Cham, Switzerland: Springer, Oct. 2022,pp. 659 675.

[49] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," inProc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer, 2020, pp. 213–229.

[50] Y. Zhang, T. Wang, and X. Zhang, "MOTRv2: Bootstrapping endto-end multi-object tracking by pretrained object detectors," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2023, pp. 22056–22065.

[51] N. Iqbal, M. A. Siddique, and J. Henkel, "RMOT: Recursion in model order for task execution time estimation in a software pipeline," in Proc. Design, Autom. Test Eur. Conf. Exhib. (DATE), Mar. 2010,pp. 953–956.

[52] Z. Qin, S. Zhou, L. Wang, J. Duan, G. Hua, and W. Tang, "MotionTrack: Learning robust short-term and long-term motions for multi-object tracking," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2023, pp. 17939–17948.\

[53] Z. Deng, X. He, and Y. Peng, "Efficiency-optimized video diffusion models," in Proc. 31st ACM Int. Conf. Multimedia, Oct. 2023,pp. 7295–7303.

[54] Z. Deng, X. He, Y. Peng, X. Zhu, and L. Cheng, "MV-diffusion: Motionaware video diffusion model," in Proc. 31st ACM Int. Conf. Multimedia,Oct. 2023, pp. 7255–7263

[55] X. Ma, G. Fang, and X. Wang, "DeepCache: Accelerating diffusionmodels for free," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2024, pp. 15762–15772.