ROBUST UNSUPERVISED MULTI-OBJECT TRACKING IN NOISY ENVIRONMENTS

C.-H. Huck Yang†, Mohit Chhabra2, Y.-C. Liu1, Quan Kong2, Tomoaki Yoshinaga2, Tomokazu Murakami2

1Georgia Institute of Technology, Atlanta, GA, USA
2Lumada Data Science Lab. Hitachi, Ltd., Tokyo, Japan

ABSTRACT

Physical processes, camera movement, and unpredictable environmental conditions like the presence of dust can induce noise and artifacts in video feeds. We observe that popular unsupervised MOT methods are dependent on noise-free inputs. We show that the addition of a small amount of artificial random noise causes a sharp degradation in model performance on benchmark metrics. We resolve this problem by introducing a robust unsupervised multi-object tracking (MOT) model: AttU-Net. The proposed single-head attention model helps limit the negative impact of noise by learning visual representations at different segment scales. AttU-Net shows better unsupervised MOT tracking performance over variational inference-based state-of-the-art baselines. We evaluate our method in the MNIST-MOT and the Atari game video benchmark. We also provide two extended video datasets: “Kuzushiji-MNIST MOT” which consists of moving Japanese characters and “Fashion-MNIST MOT” to validate the effectiveness of the MOT models.

Index Terms— Multi-object tracking, Unsupervised learning, Video tracking, Robust representation learning.

1. INTRODUCTION

Multi-object tracking (MOT) is a challenging video processing task, where the aim is to locate and track one or more visual pattern(s) from an input video. MOT has many essential real-world applications, such as multimedia content analysis [1] and industrial automation [2]. However, the time and pecuniary cost of labeling large-scale video data constrains the effectiveness of current methods. Unsupervised Multi-Object Tracking (UMOT) is an appealing approach to further improve performance without the requirement of labeled training data.

Real-world applications often require data processing in the presence of environmental noise as shown in Figure 1, which is different from UMOT evaluations of video data in carefully controlled laboratory settings. The robustness of UMOT performance against environmental noise also remains relatively unexplored. In this work, we first evaluate the robustness of the state-of-the-art UMOT model against artificial noise. We propose a multi-scale tracker based on attention U-Net to improve model generalization and to avoid over-fitting on background pixels. We also show ablation studies on different visual and optimization setups in the MNIST-MOT benchmark and two extended video datasets.

Our contributions include:

• Investigation of the effect of noise on UMOT performance.
• A new DNN video tracking backbone to leverage upon attention U-Net for robust UMOT.
• Evaluation of the proposed method on MNIST-MOT [3], Atari gaming video [4], and two newly created UMOT datasets with complex patterns: (1) Kuzushiji-MNIST MOT derived from [5] and (2) Fashion-MNIST MOT derived from [6].

2. RELATED WORK

2.1. Unsupervised Multi-Object Tracking

Unsupervised object tracking often incorporates visual reconstruction and geometric rendering to utilize spatial information for predicting bounding boxes. Tracking by anima-
tion [3] (TBA) showed the first competitive UMOT results on both the MNIST-MOT and DukeMTMC [8] dataset. It used a recurrent neural network (RNN) to learn the representations of object motion and a convolutional neural network (CNN) based autoencoder to reconstruct frames and visual patterns from geometrical tracker arrays. Spatially invariant attend-repeat (SPAIR) models [2] used CNN feature extractors and a local spatial object specification scheme to conduct UMOT video processing with variational inference. Building upon the SPAIR method, spatially invariant label-free object tracking (SILOT) [4] incorporated VAE based architecture from geometrical tracker arrays. Spatially invariant attention [3] (TBA) showed the first competitive UMOT results on MNIST-MOT and Atari tracking (SILOT) [4] incorporated VAE based architecture.

3. ENHANCED UNSUPERVISED MULTI-OBJECT TRACKING

3.1. Noisy Background Setup

The loss function $l_t$ at time $t$ is defined similarly to the loss function in TBA (refer to Eq. (9) in [3]):

$$l_t = \text{MSE} \left( X_t, X_t^C \right) + \lambda \frac{1}{I} \sum_{i=1}^{I} (s_{t,i}^x, s_{t,i}^y),$$

where the first term is the mean squared error (MSE) between $X_t$, a ground truth frame and $X_t^C$, reconstructed frame generated by DNN. The second term is tightness constraint on the bounding box size and is computed by $\lambda$, a scaling coefficient; $I$, the number of trackers and $s_{t,i}^x, s_{t,i}^y$ object poses.

To simulate noisy conditions, we consider random noise $\delta_t \sim \mathcal{N}(0,1)$ sampled from Gaussian distribution, which has been used previously in robust video learning studies [16, 17]. The total training frames of a video input in Eq. (1) are modified to $\sum_{t=1}^{T} X_t' = \sum_{t=1}^{T} (X_t + \beta \times \delta_t)$ as a noisy setup in testing for total time step $t \in \{1, 2, ..., T\}$, where $\beta \in \{0\% , 10\% , 20\% , 30\% \}$ scales the noise magnitude to a percentage of maximum intensity (refer to [16, 17]).

3.2. Attention U-Net Feature Encoder

We design a spatial feature encoder consisting of transformer-based single-head attention [18, 19] module. ResNet$_18$ [20] encoder extracts the features map $(m_t)$ from inputs $X_t$. The feature map $m_t$ is also used to extract key $(k_t)$, value $(v_t)$ and query $(q_t)$ tensors. The attention map $(A_t)$ is computed by scaling the value $(v_t)$ tensor by the inner product of the key $(k_t)$ and query $(q_t)$ tensors along the spatial dimensions (width and height) of the feature map. Here, the view operation is used to reshape the tensors.

$$m_t = f_{ResNet, \theta_1} (X_t); \quad v_t = \text{view} (f_{q, \theta_2} (m_t));$$

$$k_t = \text{view} (f_{k, \theta_3} (m_t));$$

$$A_t = \text{view} \left( \frac{q_t k_t^T}{\sqrt{d_k}} \right) v_t;$$

where $f_{ResNet}, f_q, f_k, f_v$ are individual DNNs with hyper-parameters $(\theta_1, \theta_2, \theta_3)$ updated by end-to-end gradient descent training. As shown in Figure 2 (a), a U-Net [11] encoder is used to learn multi-scale features from the attention map: $C_t = \text{U-Net}(A_t; \theta_5)$. The attention U-Net comprises of channel sizes $\{32, 64, 128, 64, 32\}$. Each convolutional layer is followed by a rectified linear unit and a 2 $\times$ 2 max-pooling operation with stride 2 for down-sampling.

3.3. Tracker Arrays and Rendering

After obtaining the feature representation $C_t$, we use tracker arrays to extract the individual location of the object(s) from a frame. As shown in Figure 2 (b), we use a latent state $H_{(t,i)}$ feature extractor for $i^{th}$ tracker to convert spatial information by neural sequence modeling:

$$H_{(t,i)} = f_{seq, \theta_6} (H_{t-1}, C_t),$$

where $f_{seq, \theta_6}$ consists of bidirectional GRU [21] networks. To obtain the location of objects in an unsupervised manner, we incentivize a jointly parameterized tracker network $(f_{\text{joint}})$ to predict a single object location $(Y_{t,i})$.

$$Y_{t,i} = f_{\text{joint}} \left( h_{t,i}; \theta_{\text{out}} \right)$$

where the output $Y_{t,i}$ is a mid-level representation of objects on 2D image planes, including: (a) probability of having tracked an object; (b) object pose $(s_{t,i}^x, s_{t,i}^y)$ in Eq. (1); (c) shape (with channel size); (d) object appearance. These mid-level output locations are used to validate model performance by MOT evaluation. We follow the layer-wise composition setup from TBA [3] and use Spatial Transformer Network (STN) [22] to scale $Y_{t,i}$ to multiple layer foreground $(L_t^{j})$
Fig. 2: Overview of the U-Net enhanced animation tracking framework: (a) multi-scale learning; (b) sequence learning; (c) loss computation. This enhanced scale-free learning method is motivated by the tracking-by-animation benchmark [3].

![U-Net Diagram](image)

Fig. 3: Four training datasets used in our experiments: (a) MNIST-MOT [3, 7], (b) Atari gaming video [4], (c) Kuzushiji-MNIST MOT, and (d) Fashion-MNIST MOT. A test dataset with scrambled objects (e) is also provided to evaluate generalization ability.

![Dataset Images](image)

4. EXPERIMENTS

4.1. Datasets and Baseline Setup

As shown in Figure 3, we select two baseline video datasets from the previous UMOT studies: (a) MNIST-MOT containing 2M training frames and 25k validation frames, where each frame is of size 128 x 128 x 1 and (b) Atari gaming video on Space-Invader (Atari-SI) containing 128k training frames and 1k validation frames, where each frame is converted to gray-scale with a input size of 210 x 160 x 1 for testing.

MOT Video Datasets: To study the effects of tracking on patterns other than MNIST characters, we provide two new MOT video datasets with more intricate patterns: (c) Japanese cursive characters (denoted as Kuzushiji from [5]) and (d) Fashion-MNIST MOT (denoted as Fashion derived from [6]). These new datasets have been chosen to study complex visual patterns and could be useful for different UMOT applications (e.g., pedestrian tracking).

Object-Scrambling Test: To verify whether the unsupervised model indeed learns semantically meaningful patterns, we provide an object-scrambling video test-set containing 1k frames as shown in Figure 3 (e). The test-set was constructed from the MNIST-MOT video dataset by randomly permuting the pixels in object bounding boxes. Ideally, a robust UMOT model [7, 2] should learn semantically consistent discriminative representations for predicting object location. A model which learns such discriminative patterns on the MNIST-MOT dataset is expected to have degraded performance on the object scramble test-set.

Baseline Models: We select two major benchmark UMOT models as reproducible studies as shown in Table 1.

- Tracking by animation (TBA) [3]: a model combines with RNNs for sequence modeling with attention features and uses CNN feature extraction for representation learning on frame rendering.
- Spatially invariant label-free object tracking (SILOT) [4]: a model built upon VAE for frame-wise reconstruction and using YOLO-based CNN feature extraction for representation learning.

U-Net based UMOT backbones: We design two different type of U-Net based feature encoder backbones: residual U-Net network with a ResNet50 encoder (denoted as ResU-Net) and propose self-attention U-Net (denoted as AttU-Net) as discussed in section 3.2. The tested UMOT models have a comparable number of model parameters $\sim 12M$.

4.2. UMOT Performance in Different Contexts

As shown in Figure 4, we evaluated the MOT accuracy with noise level ranging from 0% (without noise) to 30%. All the models were trained by using noise free frames as shown.
### 4.3. Evaluation of UMOT Model Generalization

To study generalization of different UMOT models with noisy inputs, we evaluated pre-trained UMOT models on MNIST-MOT test set in two different configurations. **(1) Testing with scrambled objects:** As in Table 2, U-Net backbones show major degradation when context-wise representation has been corrupted by the scrambled object settings (S). As discussed in Section 4.1, U-Net backbones generalize better than baseline models. **(2) Training with augmented noise:** We also evaluated as to how models would benefit from training with noisy Fashion-MNIST MOT video dataset. As shown in Table 3, both U-Net based UMOT models show better performance than TBA and SILOT by a noisy training (N) with augmented inputs containing 30% noise ratio. All UMOTs improve in terms of absolute performance but the architecture-specific (e.g., multi-scale learning by U-Net based backbones) differences still significantly impact performance metrics.

#### Table 2: Performance of pretrained UMOT models on MNIST-MOT when tested on scrambled MNIST objects (S).

| Model       | S-MOTA | S-MOTP |
|-------------|--------|--------|
| TBA         | 78.9%  | 87.1%  |
| SILOT       | 91.4%  | 83.9%  |
| ResU-Net    | 65.1%  | 61.4%  |
| AttU-Net    | 61.7%  | 59.8%  |

#### Table 3: Performance of different UMOT models trained with 30% noise (N) in Fashion-MNIST MOT dataset. Higher numbers indicate better generalisation ability.

| Model       | N-MOTA | N-MOTP |
|-------------|--------|--------|
| TBA         | 24.1%  | 46.7%  |
| SILOT       | 76.3%  | 75.2%  |
| ResU-Net    | 81.2%  | 80.1%  |

#### 5. Conclusion

In this work, we studied UMOT performance with noisy inputs and show that pre-existing methods suffer a substantial decrease in MOT accuracy when the noise ratio is increased above 30%. AttU-Net based approach increases the robustness to noise on all evaluated datasets. We also show the ability of AttU-Net to learn semantically relevant features by testing it on the test set of scrambled instances.

---

**Table 1:** Detailed MOT metrics in the clean MNIST-MOT dataset \([3, 4]\) with UMOT models as a reproducible baseline.

| Model       | IDR | IDP | IDR | Rec | Pr | GT | MT | PT | ML | FP | FN | IDs | FM | MOTA | MOTP |
|-------------|-----|-----|-----|-----|----|----|----|----|----|----|----|-----|----|-------|------|
| TBA         | 99.4% | 99.3% | 99.5% | 99.8% | 99.6% | 99 | 99 | 2 | 0 | 9 | 4 | 2 | 1 | 99.3% | 88.4% |
| SILOT       | 98.9% | 98.2% | 99.5% | 99.9% | 98.5% | 100 | 100 | 0 | 0 | 34 | 3 | 7 | 1 | 98.0% | 85.9% |
| ResU-Net    | 99.5% | 99.5% | 99.8% | 100.0% | 99.7% | 99 | 99 | 0 | 4 | 3 | 2 | 1 | 0 | 99.4% | 89.2% |
| AttU-Net    | 99.9% | 99.9% | 100.0% | 99.9% | 99 | 99 | 0 | 3 | 1 | 1 | 0 | 99.8% | 89.6% |

In Figure \(3 \text{ (a) to (d)}\). In the noise free setting (0% noise), MNIST-MOT derived from \([3, 4]\) and Kuzushiji-MOT derived from \([4, 4]\), the evaluated models attained MOT accuracy ranging above 95; Atari-SI \([4b]\) and Fashion-MNIST MOT derived from \([4d]\) were found to be more challenging and the MOT accuracy was below 90% for tested models. When the noise magnitude was increased, TBA model showed a sharp decrease in performance over all tested datasets. MOT accuracy for TBA model fell below ~ 20% in the MNIST-MOT dataset as noise was increased to 30%.

Although SILOT models are more robust than TBA under the presence of small amount of noise. Their performance suffers major degradation when the noise ratio is increased. It decreases to 52.12% in Atari-SI when the noise level is 20%. For U-Net based backbones, both ResU-Net and AttU-Net show robustness in the MNIST-MOT dataset even with the presence 30% noise. For all datasets, AttU-Net based UMOTs show best performance with substantial absolute improvements (~ 3-5%) over the ResU-Net with the exception of Fashion-MNIST MOT where the improvement is (1.01%) as \([4d]\). In short, the performance of existing benchmark UMOTs, TBA and SILOT, are sensitive to noise whereas U-Net based UMOTs perform competitively.

![Fig. 4: MOTA Performance under different noisy level testing in: (a) MNIST-MOT \([7, 3]\), (b) Atari-Space Invader \([4]\) (SI) video, (c) Kuzushiji-MNIST MOT, and (d) Fashion-MNIST MOT tracking datasets.](image-url)
6. REFERENCES

[1] Anders Johansson, Dirk Helbing, and Pradyumn K Shukla, “Specification of the social force pedestrian model by evolutionary adjustment to video tracking data,” *Advances in complex systems*, vol. 10, no. supp02, pp. 271–288, 2007.

[2] Sorin Grigorescu, Bogdan Trasnea, Tiberiu Cocias, and Gigel Macesanu, “A survey of deep learning techniques for autonomous driving,” *Journal of Field Robotics*, vol. 37, no. 3, pp. 362–386, 2020.

[3] Zhen He, Jian Li, Daxue Liu, Hangen He, and David Barber, “Tracking by animation: Unsupervised learning of multi-object attentive trackers,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1318–1327.

[4] Eric Crawford and Joelle Pineau, “Exploiting spatial invariance for scalable unsupervised object tracking,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 3684–3692.

[5] Tarin Clanuwat and et al., “Deep learning for classical japanese literature,” *arXiv preprint arXiv:1812.01718*, 2018.

[6] Han Xiao and et al., “Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms,” *arXiv preprint arXiv:1708.07747*, 2017.

[7] Eric Crawford and Joelle Pineau, “Spatially invariant unsupervised object detection with convolutional neural networks,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 3412–3420.

[8] Mengran Gou and et al., “Dukemtmc4reid: A large-scale multi-camera person re-identification dataset,” in *Proc. CVPR Workshops*, 2017, pp. 10–19.

[9] Antoni Buades, Bartomeu Coll, and Jean-Michel Morel, “A review of image denoising algorithms, with a new one,” *Multiscale Modeling & Simulation*, vol. 4, no. 2, pp. 490–530, 2005.

[10] Hao-Hsiang Yang, Chao-Han Huck Yang, and Yi-Chang James Tsai, “Y-net: Multi-scale feature aggregation network with wavelet structure similarity loss function for single image dehazing,” in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 3107–3111.

[11] Ashish Vaswani and et al., “Attention is all you need,” *Advances in Neural Information Processing Systems*, vol. 30, pp. 5998–6008, 2017.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[13] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling,” *arXiv preprint arXiv:1412.3555*, 2014.

[14] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al., “Spatial transformer networks,” *Advances in neural information processing systems*, vol. 28, pp. 2017–2025, 2015.