Robust Unsupervised Multi-Object Tracking in Noisy Environments

C.-H. Huck Yang[1], Mohit Chhabra[2], Y.-C. Liu[1], Quan Kong[2], Tomoaki Yoshinaga[2], Tomokazu Murakami[2]
[1]*Georgia Institute of Technology, USA*, [2]*Lumada Data Science Lab. Hitachi, Ltd., Tokyo, Japan*

Overview

In this work, we first evaluate the robustness of the state-of-the-art UMOT model against artificial random noise. We then propose a multi-scale tracker based on attention U-Net to improve model generalization and to avoid over-fitting on irrelevant pixels.

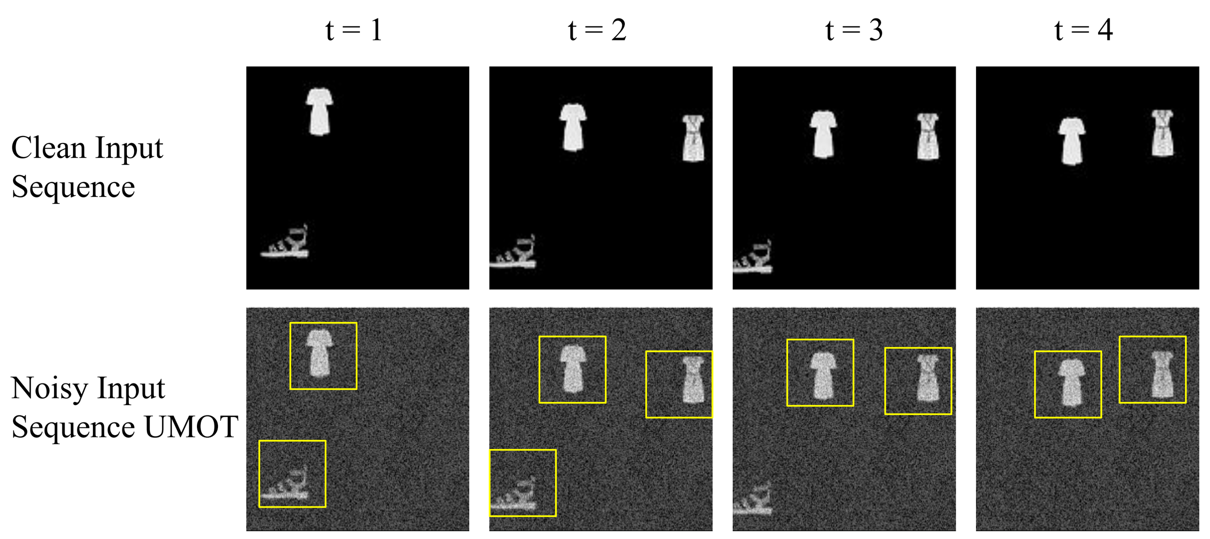
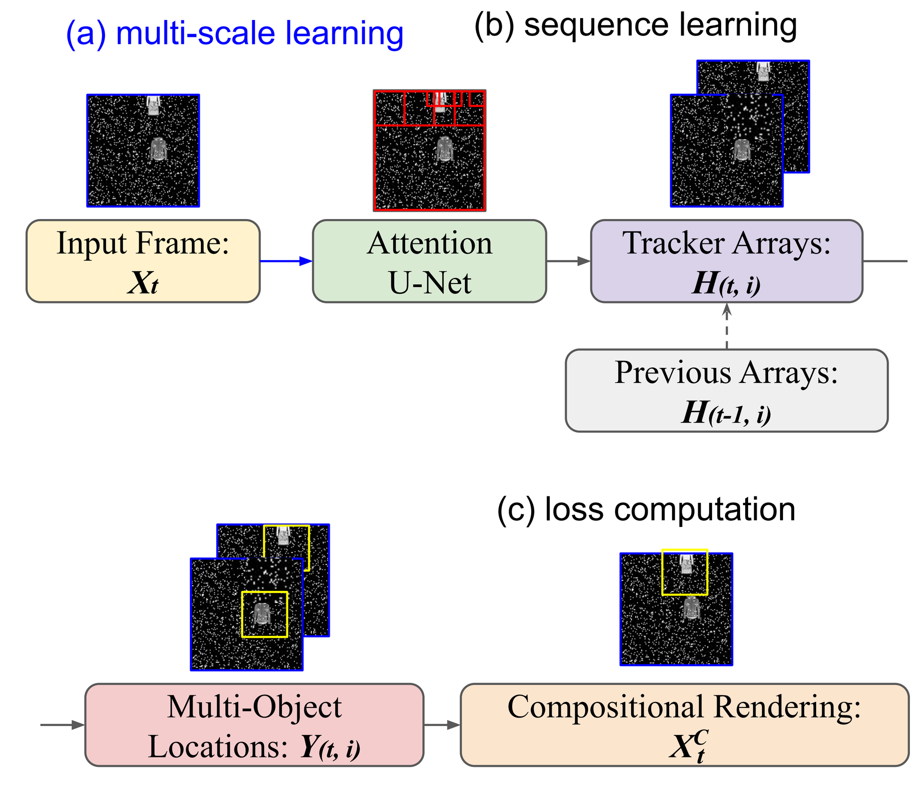


Figure 1: Unsupervised multi-object tracking (UMOT) on a new fashion context video dataset.

Proposed UMOT Model



- Investigate the effect of noise on UMOTs.
- Propose a new attention U-Net for robust UMOT and attain competitive results.
- Two newly created UMOT datasets with complex patterns: (1) Japanese cursive characters (Kuzushiji) and (2) fashion images.

Enhanced Unsupervised Multi-Object Tracking

- Noisy Background Setup.

1. For reproducible studies, we first define a loss objective l_t at time t in a standard setup UMOT model, TBA.
2. $l_t = \mathrm{MSE}\left(\boldsymbol{X}_t, \boldsymbol{X}_t^C\right) + \lambda \cdot \frac{1}{I}\sum_{i=1}^{4}(s_{t,i}^x, s_{t,i}^y)$,
3. where the first term is the reconstruction mean squared error (MSE) between $\boldsymbol{X}_t$ (a grounded truth frame) and $\boldsymbol{X}_t^C$ (generated by DNN reconstruction), and the second term is tightness constraint on the bounding boxe size computed by λ (a scaling coefficient), I (a number of trackers), and $s_{t,i}^x, s_{t,i}^y$ (object poses).
4. To simulate remaining environment noise received from the image sensors, we consider a random noise $\delta_t \sim \mathcal{N}(0,1)$ sampled from Gaussian distribution.
5. The total training frames of a video input are modified to $\sum_{t=1}^{T} \boldsymbol{X'}_t = \sum_{t=1}^{T}(\boldsymbol{X}_t + \beta \times \delta_t)$ as a **noisy setup in testing** for total time step $t \in \{1, 2, ..., T\}$, where $\beta \in \{0\%, 10\%, 20\%, 30\%\}$ refers to a noise ratio.

Attention U-Net Feature Encoder: A spatial feature encoder consisting of transformer-based attention (A_t) is computed by a feature map (m_t) with a ResNet_{18} encoder extracting from $\boldsymbol{X_t}$ feeding into keys (k_t) and value (v_t) with queries (q_t):

$$m_t = f_{ResNet,\theta_1}(\boldsymbol{X_t}); \quad q_t = \mathrm{unroll}\left(f_{q,\theta_2}(m_t)\right);$$
$$k_t = \mathrm{unroll}\left(f_{k,\theta_3}(m_t)\right); v_t = \mathrm{unroll}\left(f_{v,\theta_4}(m_t)\right);$$
$$A_t = \mathrm{softmax}\left(\frac{q_t k_t^T}{\sqrt{d_k}}\right) v_t,$$

Video Dataset

Four training datasets (a) to (d) and one test dataset (e) with scrambled objects is provided to evaluate context learning effects.

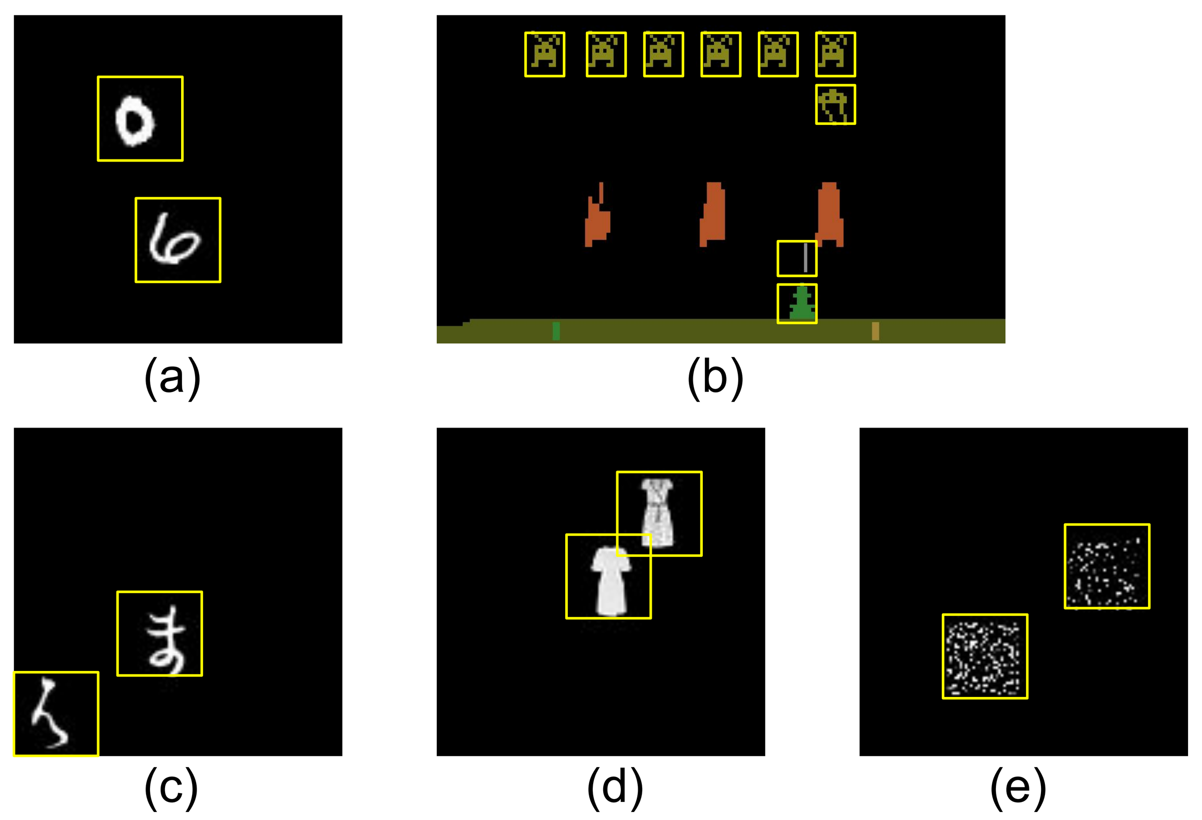


Figure 2: (c) Kuzushiji video, and (d) Fashion video.

Generalization Results

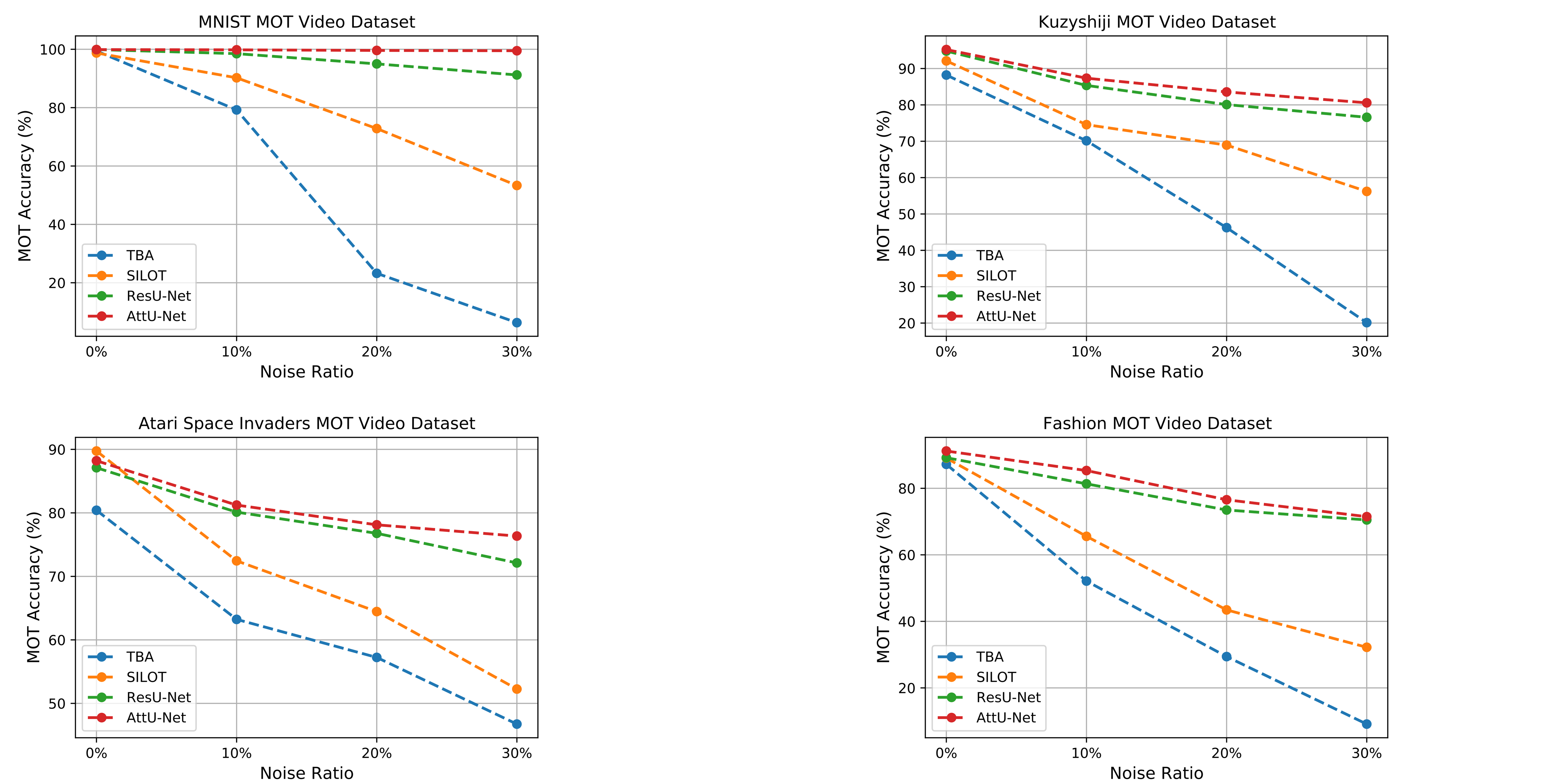
	TBA	SILOT	ResU	AttU
S-MOTA	91.4%	83.9%	65.1%	**61.7%**
S-MOTP	90.9%	82.1%	61.4%	**59.8%**

Table 1: Performance (↓) of pretrained UMOT models in MNIST but testing with the scrambled MNIST objects (S) .

	TBA	SILOT	ResU	AttU
N-MOTA	24.1%	46.7%	76.3%	**81.2%**
N-MOTP	23.2%	44.8%	75.2%	**80.1%**

Table 2: Performance (↑) of different UMOT models trained with a 30% noise (N) in Fashion video dataset.

MOT Accuracy Performance Discussion



ICIP 2021

Paper Link Presentation

[1] Yang et al. "robust unsupervised multi-object tracking in noisy environments". *IEEE ICIP*, 2021.

Conclusion and Future Research

We show the capability of AttU-Net UMOT to learn useful discriminative features by testing it on the test set of scrambled instances.

Our future work includes incorporating noise augmentation training studies and investigate adversarial robustness for UMOT models.

Dataset

Data https://github.com/huckiyang/MOT-Kuzushiji-Fashion-Video

Email huckiyang@gatech.edu