BoT-FaceSORT: Bag-of-Tricks for Robust Multi-Face Tracking in Unconstrained Videos

Jonghyeon Kim[1⋆], Chan-Yang Ju[1⋆], Gun-Woo Kim[2], and Dong-Ho Lee[1†]

[1] Department of Applied Artificial Intelligence, Hanyang University, South Korea
{jonghyeon,karunogi,dhlee72}@hanyang.ac.kr
[2] Department of Computer Science and Engineering, Gyeongsang National University, South Korea
gunwoo.kim@gnu.ac.kr
https://github.com/bellhyeon/BoT-FaceSORT

Abstract. Multi-face tracking (MFT) is a subtask of multi-object tracking (MOT) that focuses on detecting and tracking multiple faces across video frames. Modern MOT trackers adopt the Kalman filter (KF), a linear model that estimates current motions based on previous observations. However, these KF-based trackers struggle to predict motions in unconstrained videos with frequent shot changes, occlusions, and appearance variations. To address these limitations, we propose BoT-FaceSORT, a novel MFT framework that integrates shot change detection, shared feature memory, and an adaptive cascade matching strategy for robust tracking. It detects shot changes by comparing the color histograms of adjacent frames and resets KF states to handle discontinuities. Additionally, we introduce MovieShot, a new benchmark of challenging movie clips to evaluate MFT performance in unconstrained scenarios. We also demonstrate the superior performance of our method compared to existing methods on three benchmarks, while an ablation study validates the effectiveness of each component in handling unconstrained videos.

Keywords: Multi-Face Tracking · SORT · Kalman Filter

1 Introduction

Multi-Object Tracking (MOT) is a computer vision task that tracks multiple objects of interest across image or video sequences, such as pedestrians, faces, sports players on the court, cars, signs on the road, and even cells or animals [20]. Among these, multi-face tracking (MFT) specifically focuses on detecting and tracking multiple faces across video frames. MFT poses additional challenges, such as deformations and occlusions due to pose, expression, and lighting changes. These challenges are further exacerbated in unconstrained videos with frequent shot changes like movies, TV shows, and web videos, where faces undergo drastic appearance changes and unexpected motions.

⋆Major in Bio Artificial Intelligence
†Corresponding Author

Fig. 1: An example of a shot change situation. In this situation, the three tracklets in the previous frame are lost in the current frame.

With recent advances in object detection and association algorithms, tracking-by-detection (TBD) has become the most popular paradigm for many MOT tasks. TBD involves two main steps: detecting objects with a well-made object detector (*e.g.*, YOLOX [10]) and associating detections across frames to assign identities using assignment algorithms like the Hungarian Algorithm [16]. Many TBD-based trackers usually integrate either a classic motion model, such as the Kalman Filter (KF) [14], or a re-identification (Re-ID) module to address the most challenging problem in MOT, such as occlusion or detection miss [1, 3, 4, 21, 30, 31, 33, 36, 37].

KF assumes that object motion follows a linear Gaussian model, where measurements are corrupted by Gaussian noise. While KF provides simple and fast trajectory estimations, it can encounter challenges in unconstrained videos, particularly at shot boundaries. Shot changes can cause noise into the motion model, leading poor predictions. Due to these limitations, most KF-based trackers typically overlook potential shot changes and only consider continuous scenes.

The Re-ID module re-identifies objects using visual features such as body, face, and clothing. These appearance-based methods provide a straightforward way to associate targets over time [21]. However, training a robust Re-ID module typically requires large amounts of labeled data covering various scenarios and appearances, which is expensive to collect and annotate. Additionally, these methods struggle in situations where objects have similar appearances or are occluded [28].

As illustrated in Figure 1, when a shot change occurs, the tracked objects in the previous frame may disappear or reappear in a different locations, resulting in significant discrepancies between the predicted motion and the actual measurements. In such cases, directly applying the KF can lead to tracklet drifts and association failures.

Existing MFT methods without using KF in unconstrained videos approach the problem in two ways: offline and online. Offline methods assume that the entire video is available simultaneously and perform global optimization to link face detections across frames. While these methods can handle long-term occlusions and shot changes by exploiting future information, they are unsuitable for

real-time applications such as live video streaming and real-time video analysis. In contrast, online methods use only past and current information to make local decisions for matching faces. Modern online MFT trackers utilize feature memory to capture past features, responding appearance changes. However, these methods depend entirely on the Re-ID module, which requires more computation costs than KF-based methods and may suffer from potential distributional mismatches between pretrained data and target sequences.

To address these limitations of existing MFT methods, we propose BoT-FaceSORT, which consists of a bag of tricks to enhance tracking robustness. Our key ideas are threefold: (1) introducing a shot change detection module based on RGB color histogram differences of adjacent frames to detect shot boundaries and reset KF states to avoid motion error propagation; (2) maintaining appearance features of tracked faces via a shared feature memory; (3) adaptively adjusting the matching strategy in a cascade manner based on motion continuity.

Additionally, to the best of our knowledge, no existing benchmark is specifically designed to evaluate MFT methods in unconstrained videos. Therefore, we build a new benchmark named MovieShot, which consists of 10 highly diverse movie clips with dense face annotations. The MovieShot dataset covers various challenging aspects of unconstrained videos, including frequent shot changes, large-scale variation, severe occlusions, and diverse subject appearances and demographics.

We evaluate the proposed BoT-FaceSORT on three benchmark datasets: MovieShot, Music [34], and ChokePoint [32] to demonstrate its efficiency and effectiveness. We utlize a comprehensive set of evaluation metrics, including Higher Order Tracking Accuracy (HOTA), Multiple Object Tracking Accuracy (MOTA), Identification F1 Score (IDF1), ID-switching (IDSW), and the speed of the tracker, Frame Per Second (FPS) [2, 19, 26]. We also demonstrate the effectiveness and efficiency of each proposed component in handling unconstrained videos through an ablation study.

The main contributions of our work can be summarized as follows:

- We propose BoT-FaceSORT, simple but robust framework that adaptively adjusts motion and appearance information to tackle the challenges of MFT in unconstrained videos.
- We introduce MovieShot, the first benchmark dataset specifically designed for unconstrained MFT with extensive annotations.
- We conduct various experiments across three challenging scenarios to demonstrate the superior performance of BoT-FaceSORT and validate the effectiveness and efficiency of each proposed component with an ablation study.

2 Related Work

2.1 Motion-based Multi-Object Tracking

Motion-based multi-object-tracking (MOT) is a straightforward way to tracking objects online based on their previous motions. Many multi-object trackers

typically adopt the Kalman Filter (KF) for estimating object motion due to its simplicity and speed. KF is a linear model that provides current state estimations through previous observations. SORT [3] is the first work to demonstrate the effectiveness of using KF for object motion prediction and data association following the tracking-by-detection (TBD) paradigm. It calculates Intersection over Union (IoU) between the KF estimation and the actual detection, then solves the linear assignment problem using Hungarian algorithm [16] to assign identities.

Recently, several KF-based trackers have expanded upon their key ideas of SORT. ByteTrack [36] observes that low-confidence detections sometimes indicate the presence of objects and incorporates them into the KF-based assignments. OC-SORT [4] proposes an object observation model to adjust for the accumulated noise in KF during occlusions. HybridSORT [33] estimates the confidence state of tracklets through an improved KF and proposes a new IoU for matching, Height Modulated IoU, to distinguish overlapping objects better. These methods have achieved state-of-the-art performance on several MOT benchmarks, such as MOT15, MOT16, MOT20, and DanceTrack [5, 17, 22, 28].

2.2 Appearance-based Multi-Object Tracking

Motion-based trackers have achieved great success in MOT, but they still struggle to estimate non-linear or occluded motions. To address these challenges, many multi-object trackers additionally utilize appearance information to overcome the limitations of KF. This appearance information is extracted as features by deep learning-based re-identification (Re-ID) modules and used to re-assign identities based on the distance or similarity of previous features.

DeepSORT [31] integrates appearance features into the motion-only tracking framework, significantly enhancing motion-based methods. It effectively alleviates occlusion issues of SORT [3] by considering both motion and appearance information for association in a cascade manner. Based on this, BoT-SORT [1] improves the motion vector of KF by adding width and height states for accurate estimation. It also proposes a fusion of motion IoU and appearance Re-ID similarity to achieve more robust associations. Similarly, HybridSORT [33], also integrates both appearance and motion information for better tracking performance.

2.3 Multi-Face Tracking in Unconstrained Videos

Appearance often changes when shot switches occur in unconstrained videos, making real-time tracking of multiple faces challenging. Due to these issues, most previous studies perform offline tracking, utilizing all frames for the entire video sequence rather than frame-by-frame online tracking [9, 18, 34, 35].

Existing multi-face-tracking (MFT) methods in unconstrained videos do not utilize motion information due to the limitations of KF. Instead, they perform association through memory-based appearance matching for tracking in an online manner which relies too heavily on the Re-ID module [24, 25].

Algorithm 1: Pseudo-code of BoT-FaceSORT

Input: Input sequence I; track set $\mathcal{T}$; face detector `Det`; Re-ID module `AdaFace`; shot change detector χ^2; shot change threshold θ
Output: Updated track set $\mathcal{T}$

Initialize $\mathcal{T} \leftarrow \emptyset$
for *frame* f_k *in* I **do**
  $\mathcal{D}_k \leftarrow$ `Det`(f_k)
  $\mathcal{F} \leftarrow \emptyset$
  for d *in* $\mathcal{D}_k$ **do**
    $\mathcal{F} \leftarrow \mathcal{F} \cup$ `AdaFace`(d) `// extract deep appearance features`
  `// shot change detection`
  `shot_changed` $\leftarrow \chi^2(f_{k-1}, f_k) > \theta$
  `// adaptive cascade matching based on shot change detection`
  if `shot_changed` **then**
    $C_{emb} \leftarrow \cos(\mathcal{T}.\text{memory.features}, \mathcal{F})$
    $C_{1st} \leftarrow C_{emb}$
    Linear assign. by Hungarian's alg. with C_{1st}
    $\mathcal{F}_{remain} \leftarrow$ remaining face features from $\mathcal{F}$
    $C_{2nd} \leftarrow \cos(\mathcal{T}.\text{memory.features}, \mathcal{F}_{remain})$
    Linear assign. by Hungarian's alg. with C_{2nd}
    $\mathcal{T}_{matched} \leftarrow$ matched tracks from $\mathcal{T}$
    `// reset KF state for matched tracks`
    for τ *in* $\mathcal{T}_{matched}$ **do**
      $\tau \leftarrow$ `KalmanFilter`(τ).initiate()
  else
    for t *in* $\mathcal{T}$ **do**
      $t \leftarrow$ `KalmanFilter`(t)
    $C_{iou} \leftarrow \text{IoU}(\mathcal{T}.\text{boxes}, \mathcal{D})$
    $C_{emb} \leftarrow \text{FusionDist}(\mathcal{T}.\text{memory.features}, \mathcal{F}, C_{iou})$
    $C_{1st} \leftarrow \min(C_{iou}, C_{emb})$
    Linear assign. by Hungarian's alg. with C_{1st}
    $\mathcal{F}_{remain} \leftarrow$ remaining face features from $\mathcal{F}$
    $C_{2nd} \leftarrow \cos(\mathcal{T}.\text{memory.features}, \mathcal{F}_{remain})$
    Linear assign. by Hungarian's alg. with C_{2nd}
    $\mathcal{T}_{matched} \leftarrow$ matched tracks from $\mathcal{T}$
  $\mathcal{F}_{matched} \leftarrow$ matched features from $\mathcal{F}$
  `// update shared feature memory for matched features and tracks`
  for f, t *in* $zip(\mathcal{F}_{matched}, \mathcal{T}_{matched})$ **do**
    $\mathcal{T}$.memory $\leftarrow \mathcal{T}$.memory $\cup$ (f, t.ID)
Return: updated track $\mathcal{T}$

Please note that the track rebirth, remove, and initialization [1,31,36] are not shown in the algorithm for simplicity. Our main contributions are highlighted in **green**.

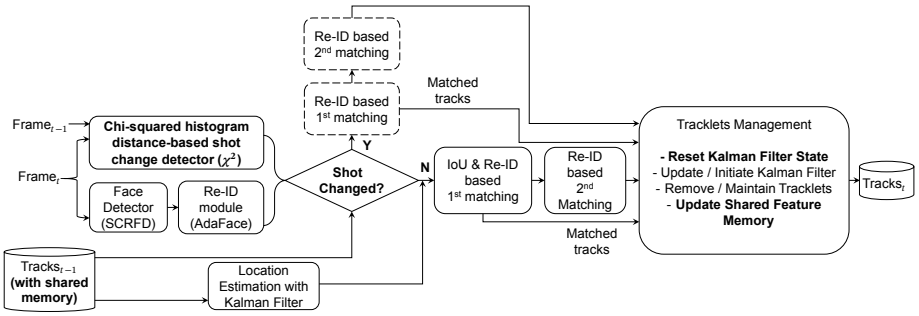


Fig. 2: Pipeline of BoT-FaceSORT. The cascade matching strategy is inspired from BoT-SORT [1], and key components of ours are in **bold**.

3 BoT-FaceSORT

In this section, we present our novel bag of tricks for multi-face tracking in unconstrained videos, BoT-FaceSORT. Our framework addresses the challenges of frequent shot changes, occlusions, and appearance variations in unconstrained videos. The key components of BoT-FaceSORT are as follows: (1) chi-squared RGB histogram distance based shot change detection module, (2) shared feature memory, and (3) adaptive cascade matching. The pseudo-code and pipeline for BoT-FaceSORT are illusrtrated in Algorithm 1 and Figure 2, respectively.

3.1 Shot Change Detection Module

Existing motion-based MOT trackers have demonstrated promising results in certain scenarios. However, they often require additional information, such as camera motion estimation, and still rely on the linear motion assumptions of KF. Therefore, these trackers are difficult to apply in unconstrained videos where frequent shot changes lead to unexpected motion variations.

To address these limitations of linear motion models in unconstrained videos, we simply detect shot boundaries based on the chi-squared distance (χ^2) [23] of normalized RGB histograms between adjacent frames as follows:

$$\chi^2(x, y) = \frac{1}{2} \sum_{i=1}^{n} \frac{(x_i - y_i)^2}{(x_i + y_i)}, \tag{1}$$

where x and y are the normalized histograms of the previous and current frames, respectively, and n is the number of bins in the histogram. When the difference between the two histograms exceeds the distance threshold θ, we consider that a shot change has occured, as in Line 7 of Algorithm 1 and in Eq. 2.

$$\texttt{shot\_changed} = \begin{cases} 1, & \chi^2 > \theta \\ 0, & otherwise \end{cases}. \tag{2}$$

This module resets the KF state for all active tracks to prevent noise at shot boundaries, where objects may disappear or appear in entirely different locations. This reset allows for new motion observations (Line 16 to 17 in Algorithm 1).

3.2 Shared Feature Memory Module

Since it is difficult to deal with shot change and occlusions using only previous feature information, we design a shared feature memory that stores deep appearance features and IDs for all tracks. This memory module is updated using a First-In-First-Out strategy, which enhances efficiency and reduces computational overhead compared to memory-per-track approaches commonly used in other MOT trackers. Although this strategy may not capture long-term appearance changes, it provides a good balance of efficiency and performance in unconstrained scenarios with frequent shot changes. The shared face memory M at time t can be defined as:

$$M_t = \{(f, \mathrm{ID})_l\}_{l=1}^{N(t)}, \tag{3}$$

where f is the face feature vector, ID is the unique identity of the track, and $N(t)$ is the total memory length at the time t.

To compute the cost matrix C for appearance matching with the face memory, we obtain the memory feature set F_i associated with each track T_i from the set of candidate tracks T. The criteria for associating features is based on the ID of each track and the corresponding ID in the memory feature, as in Eq. 4.

$$F_i = \{f_k | (f_k, \mathrm{ID}_k) \in M_t, \mathrm{ID}_k = \mathrm{ID}_i\}, \tag{4}$$

where f_k is the k-th feature vector in the face memory, and ID_k is the corresponding ID associated with each memory feature vector, matching the ID of each candidate track T_i.

And then, calculate the distance matrix G_{ij} between the associated memory set F_i and each detection D_j in the detection set D. The distances are computed using cosine distance to recognize faces among feature vectors as shown in Eq. 5.

$$G_{ij} = \{\cos(f_k, f_j) | f_k \in F_i, f_j \in D_j\}, \tag{5}$$

where f_j is the j-th feature vector of current detection. We only use the minimum distances for the cost matrix C_{ij} to find the closest and the most similar candidates between the existing memory features and the current detections as follows:

$$C_{ij} = \max(0, \min(G_{ij})). \tag{6}$$

Unlike existing Re-ID-based trackers that rely solely on previous appearance information, BoT-FaceSORT adopts this shared feature memory module to enhance the robustness of face re-identification across shot changes and long-term occlusions. We limit the maximum memory length to balance computational efficiency and tracking robustness. This module maintains temporal features for face appearances, allowing for more reliable matching even when faces undergo significant appearance changes.

3.3 Adaptive Cascade Matching

Our method follows a cascade matching strategy commonly used in modern KF-based MOT trackers. This strategy consists of two steps: the first step attempts to match through detections, active tracks that are candidates with a high probability of matching, and lost tracks that were previously active but are waiting to be removed. The second step attempts to match through detections and active tracks that are not matched in the first match and unidentified tracks in a temporary state with a low probability of matching. It allows for more precise associations by matching track candidates from high to low probability of matching in a cascade manner.

However, our BoT-FaceSORT adaptively determines which information to use for matching based on the shot change detection module. It only considers high-confidence detections and high-quality appearance information for robust face tracking. In the first matching step, lost tracks that were previously active but are waiting to be removed, active tracks, and all detections are included. This step combines motion and appearance information for matching when no shot changes are detected. The second matching step attempts to match using detections that failed to match in the first step, active tracks that are not matched in the first step, and unidentified tracks that are not yet active.

When a shot change occurs, BoT-FaceSORT attempts the first and second matching step using only appearance information based on the Re-ID module, omitting motion information that may be noisy due to unexpected shot change. And then, resets the KF state for the tracks matched in both the first and second steps (Line 8 to 17 in Algorithm 1 and the dotted line in Figure 2).

If no shot change occurs, we integrate KF motion and appearance information for the first matching step, similar to BoT-SORT [1]. Then, perform a second matching step using only appearance information, as in the shot change situation (Line 18 to 28 in Algorithm 1 and N divergence at shot changed in Figure 2).

We named this matching strategy as adaptive cascade matching. It adaptively utilizes motion and appearance information based on whether a shot change has occured. This adaptive strategy maintains notable tracking performance in both unexpected shot changes and continuous segments with no-shot changes.

4 Experiments

In this section, we present a comprehensive evaluation of our proposed tracker, BoT-FaceSORT. We conduct experiments on three challenging MFT datasets: MovieShot, Music [34] and ChokePoint [32]. Our evaluation aims to demonstrate the effectiveness of BoT-FaceSORT in handling unconstrained videos with frequent shot changes, occlusions, and appearance variation. We compare our method with state-of-the-art KF-based trackers, perform an ablation study to validate the contribution of each component, and provide both quantitative and qualitative analyses of the results.

4.1 Datasets

MovieShot. To address the limitations of existing MFT datasets, we introduce MovieShot, a new diverse dataset specifically designed for evaluating MFT performance in challenging scenarios.

Most publicly available MFT datasets are either unannotated [32] or lack precise annotations [34], unlike existing MOT datasets such as MOT15, MOT16, MOT20, DanceTrack, and SoccerNet [5, 11, 17, 22, 28], which primarily focus on person tracking. This discrepancy makes it difficult to effectively benchmark MFT trackers.

MovieShot aims to fill this gap by providing a comprehensive and precisely annotated dataset for MFT in challenging movie environments. The dataset consists of 10 diverse movie clips sourced from YouTube. We carefully selected and annotated clips to represent real-world scenarios and tracking challenges, prioritizing diversity in genres, shot changes, demographics, and challenging scenarios such as occlusions and varying face scales. The final dataset comprises 32,552 frames across the 10 movie clips, totaling 69,204 unique face detection and 291 tracks. For more details of our dataset, please refer our supplementary material.

Music [34]. The Music dataset consists of 3 live vocal concert recordings from multiple cameras with different views and 5 music video clips, that feature faster and more frequent shot changes than MovieShot. However, it may be unsuitable for precise MFT benchmarking since it only provides annotations for main casts.

ChokePoint [32]. The ChokePoint dataset consists of 48 real-world surveillance without shot changes. Among them, we selected 6 sequences that represent crowded scenarios with numerous variations and occlusions to test the generalization ability of our method against more challenging real-world surveillance problems.

4.2 Metrics

For evaluation, we follow the metrics widely adopted in MOT: Higher-Order Tracking Accuracy (HOTA), Multiple Object Tracking Accuracy (MOTA), ID Switching (IDSW), Identity F1 Score (IDF1), and Frame Per Second (FPS) [2, 19, 26]. HOTA is a metric that considers both accuracy and consistency of object tracking. MOTA considers the false positives and false negatives of detection. IDSW considers the number of times identity changes while tracking the object to evaluate the accuracy of multiple object tracking. IDF1 is a metric that accurately maintains the identity of tracked objects, considering the precision and recall of identity to calculate the F1-score. FPS is a metric that indicates the speed of the tracker to evaluate the impact of the proposed methods on the execution time.

Table 1: HOTA performance comparison of different face detectors on the MovieShot and Music datasets. For a fair comparison, we use the same hyperparameters for each model and tracker.

Dataset	Detector	DeepSORT [31]	BoT-SORT [1]	ByteTrack [36]	OC-SORT [4]	Deep OC-SORT [21]	StrongSORT [8]	HybridSORT [33]	BoT-FaceSORT (ours)
MovieShot	RetinaFace [6]	33.53	36.75	37.09	39.13	36.99	38.19	41.27	**42.44**
	YOLOv7 [29]	35.86	38.36	38.2	40.39	38.33	39.64	42.18	**44.21**
	SCRFD [12]	36.84	40.9	40.68	42.43	40.19	42.28	45.88	**47.39**
Music [34]	RetinaFace [6]	9.5	10.49	10.65	10.87	10.4	10.43	11.81	**18.2**
	YOLOv7 [29]	9.48	10.64	10.7	11.07	10.33	10.59	11.96	**19.07**
	SCRFD [12]	9.72	11.14	11.11	11.28	10.66	10.95	12.36	**19.62**

4.3 Implementation Details

Face Detector. In MOT, the primary objects being tracked are bodies, whereas in MFT, the focus is on faces. Therefore, we utilize a novel face detector, SCRFD [12], instead of the person detector YOLOX [10]. We employ pre-trained weights by a face detection dataset, CrowdHuman [27] in our experiments.

Deep Feature Extractor for Face Re-Identification. We utilized AdaFace [15], which provides high-quality face recognition for more accurate face recognition in various situations. The backbone of the module is ResNet100 [7], which improves the residual structure of ResNet [13]. The pre-trained weights were also adopted trained on WebFace12M [38], a million-scale face dataset that considered various situations and races.

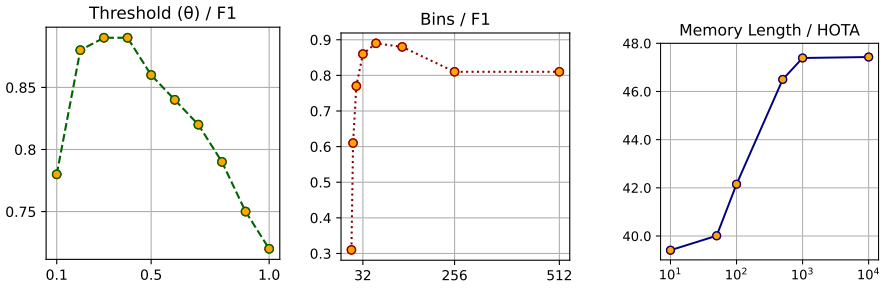
Configuration. For the face detector, the detection threshold is set to 0.5, and the non-maximum suppression threshold is 0.7. The IoU and Re-ID thresholds for cascade matching are 0.35 and 0.3, respectively. The shot change detection threshold is 0.4, and the number of bins for calculating the histogram to detect shot changes is 64. We set the maximum shared memory length to 1000 and the maximum age of all lost tracks to 100 to respond shot-change situations. BoT-FaceSORT rejects the matching if the similarity between detection and tracklet is below 0.2 during the linear assignment step, the same as BoT-SORT [1]. All experiments are implemented with PyTorch on a desktop with an AMD Ryzen 9 7950X @ 4.5GHz processor and an NVIDIA GeForce RTX 4090 GPU.

4.4 Ablation Study

Different Face Detectors. BoT-FaceSORT follows a tracking-by-detection paradigm, where detection performance directly affects tracking performance. Therefore, we evaluate two additional face detectors YOLOv7 [29] and RetinaFace [6] alongside SCRFD [12] on the MovieShot and Music [34] datasets. As shown in Table 1, our tracking method achieves the highest HOTA performance on both datasets across all detectors. This result demonstrates that BoT-FaceSORT is robust regardless of detection quality.

Table 2: Ablation study on the MovieShot and Music datasets.

Dataset	SC	SM	HOTA↑	MOTA↑	IDSW↓	IDF1↑	FPS↑
MovieShot			42.31	83.59	1042	32.85	**84.41**
	✓		45.62	83.94	946	36.39	84.26
		✓	42.56	83.54	1014	33.56	81.4
	✓	✓	**47.39**	**84.03**	**902**	**39.33**	81.05
Music [34]			12.9	58.91	1900	9.63	80.3
	✓		15.77	61.19	1611	13.06	**80.69**
		✓	13.44	58.65	1824	10.33	76.87
	✓	✓	**19.62**	**62.17**	**1367**	**19.16**	76.75



(a) (b) (c)

Fig. 3: Performance of key hyperparameters for our components on the MovieShot dataset.

Components Analysis. We conduct an ablation study on the MovieShot and Music datasets to evaluate the contribution of each component: the Shot Change Detection Module (SC) and the Shared Memory Module (SM). As shown in Table 2, adopting either SC or SM individually improves overall tracking performance on both datasets. Using SC alone enhances performance without significantly affecting tracking speed. Although using SM alone results in a minor frame drop and slightly decreases MOTA, but it improves the other metrices. The integration of both SC and SM shows the most notable improvements in both datasets, particularly in the HOTA and IDF1 metrics. Specifically, we observe improvements of over 5% for HOTA and 6% for IDF1 in the MovieShot dataset, and about 7% for HOTA and 9% for IDF1 in the Music dataset. These improvements indicate that more accurate and consistent identity tracking is possible in complex, unconstrained video sequences.

Effects of Key Hyperparameters. We perform some experiments to determine the default values for key hyperparameters in our components: the shot change detection threshold (θ), the number of bins for calculating the color his-

Table 3: Comparison with KF-based trackers on the MovieShot dataset.

Tracker	Re-ID	HOTA↑	MOTA↑	IDSW↓	IDF1↑	FPS↑
DeepSORT [31]	✓	36.84	60.63	3025	25.36	63.48
BoT-SORT [1]	✓	40.9	82.76	1007	30.61	73.6
ByteTrack [36]		40.68	82.53	1009	31.03	129.57
OC-SORT [4]		42.43	80.37	923	31.62	**129.72**
Deep OC-SORT [21]	✓	40.19	74.27	1164	28.49	66.22
StrongSORT [8]	✓	42.28	81.25	2711	30.75	64.33
HybridSORT [33]	✓	45.88	82.92	**803**	34.97	49.4
BoT-FaceSORT (ours)	✓	**47.39**	**84.03**	902	**39.33**	81.05

togram of adjacent frames (`bins`), and the maximum shared memory length (l). First, we perform a baseline experiment by setting the value of θ to 0.5 and `bins` to 32 (Figure 3a and 3b). Then, we set the default values for θ and `bins` to 0.4 and 64, respectively, based on the point where each F1 score of shot change detection reaches its maximum. Finally, the default value for l is selected based on the HOTA score across different memory lengths with default values for θ and `bins` on the MovieShot dataset. As shown in Figure 3c, we observe that the HOTA score does not significantly change beyond 10^3, so we set l to 1000 for computational efficiency.

4.5 Benchmarks Evaluation

We compare BoT-FaceSORT with several state-of-the-art Kalman filter-based multi-object trackers [1, 4, 8, 21, 31, 33, 36] using MovieShot, Music [34], and ChokePoint [32] datasets. For a fair comparison, we employ the same face detector and face recognition module and the default hyperparameters provided for each tracker. Please note that we did not pre-train or fine-tune the detector and Re-ID models for the benchmarks.

MovieShot dataset. Table 3 demonstrates the comparison results of BoT-FaceSORT with existing state-of-the-art MOT methods on the MovieShot dataset. BoT-FaceSORT achieves superior performance, achieving scores of 47.39 for HOTA, 84.03 for MOTA, and 39.33 for IDF1, outperforming the existing state-of-the-art methods. In particular, the IDF1 score shows a significant improvement, indicating that our method maintains more consistent identifications across shot changes. However, the processing speed of our method is slower compared to methods without Re-ID module, ByteTrack [36] and OC-SORT [4]. Nevertheless, it is the fastest tracker among Re-ID-based trackers.

Music dataset. We also evaluate the Music dataset to validate our proposed methods across a wider range of scenarios. As shown in Table 4, most state-of-the-art methods perform poorly. It may be associated with imprecise annotations

Table 4: Comparison with KF-based trackers on the Music dataset.

Tracker	Re-ID	HOTA↑	MOTA↑	IDSW↓	IDF1↑	FPS↑
DeepSORT [31]	✓	9.72	27.66	4298	5.66	64.03
BoT-SORT [1]	✓	11.14	57.24	2332	7.11	73.19
ByteTrack [36]		11.11	58.71	2372	7.08	131.09
OC-SORT [4]		11.28	**63.33**	2279	7.61	**132.55**
Deep OC-SORT [21]	✓	10.66	48.03	2478	6.82	65.79
StrongSORT [8]	✓	10.95	60.99	4102	6.99	62.37
HybridSORT [33]	✓	12.36	61.95	2008	8.83	49.63
BoT-FaceSORT (ours)	✓	**19.62**	62.17	**1367**	**19.16**	76.75

Table 5: Comparison with KF-based trackers on the ChokePoint dataset.

Tracker	Re-ID	HOTA↑	MOTA↑	IDSW↓	IDF1↑	FPS↑
DeepSORT [31]	✓	72.33	65.67	1612	73.39	78.82
BoT-SORT [1]	✓	79.52	80.87	69	85.05	90.16
ByteTrack [36]		79.45	82.07	69	85.49	161.96
OC-SORT [4]		88.77	**88.51**	**64**	88.76	**164.52**
Deep OC-SORT [21]	✓	82.98	80.95	107	82.77	85.9
StrongSORT [8]	✓	81.63	82.91	728	79.92	83.14
HybridSORT [33]	✓	**88.93**	88.43	65	**88.79**	70.85
BoT-FaceSORT (ours)	✓	82.91	85.2	66	87.35	90.5

or the capability to respond to fast and frequent shot-change situations. Nevertheless, our method performs the best on all metrics except MOTA and FPS, with huge gaps. These results indicate that our method is also robust in various situations with fast shot changes.

ChokePoint dataset. We also evaluate the ChokePoint dataset to address the curiosity about the effectiveness of the proposed methods in the general scenarios of no-shot changes. For videos without shot changes, our method uses only the shared feature memory and almost follows the matching method of BoT-SORT [1]. As shown in Table 5, although BoT-FaceSORT does not achieve state-of-the-art performance, it still performs competitively on the ChokePoint dataset.

5 Conclusion

In this paper, we propose BoT-FaceSORT, a Kalman filter-based tracker for multi-face tracking in unconstrained videos. Our method consists of a simple and efficient bag of tricks: a shot change detection module, a shared feature memory module, and an adaptive cascade matching strategy. Additionally, we build a new benchmark dataset named MovieShot, for multi-face tracking in

Fig. 4: Qualitative results of BoT-FaceSORT on each of the three videos from MovieShot and Music [34]. The frame number is in the upper left corner of each figure, and different colored bounding boxes represent different identities.

unconstrained videos due to the lack of publicly available datasets for multi-face tracking. Our proposed methods are particularly effective in unconstrained videos with frequent shot changes. However, they may be slightly less effective in videos without shot changes, as our approach is primarily designed around shot change detection. We will further explore other memory management approaches for the shared feature memory module like Least Recently Used algorithm for better generalization.

Acknowledgments. This work was partly supported by Institute of Information communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2022-00155885, Artificial Intelligence Convergence Innovation Human Resources Development (Hanyang University ERICA)) and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No.NRF-2022R1F1A1073208).

References

1. Aharon, N., Orfaig, R., Bobrovsky, B.Z.: Bot-sort: Robust associations multi-pedestrian tracking. arXiv preprint arXiv:2206.14651 (2022)
2. Bernardin, K., Stiefelhagen, R.: Evaluating multiple object tracking performance: the clear mot metrics. EURASIP Journal on Image and Video Processing **2008**, 1–10 (2008)
3. Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B.: Simple online and realtime tracking. In: 2016 IEEE International Conference on Image Processing (ICIP). pp. 3464–3468 (2016)
4. Cao, J., Pang, J., Weng, X., Khirodkar, R., Kitani, K.: Observation-centric sort: Rethinking sort for robust multi-object tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9686–9696 (June 2023)
5. Dendorfer, P., Rezatofighi, H., Milan, A., Shi, J., Cremers, D., Reid, I., Roth, S., Schindler, K., Leal-Taixé, L.: Mot20: A benchmark for multi object tracking in crowded scenes. arXiv preprint arXiv:2003.09003 (2020)
6. Deng, J., Guo, J., Ververas, E., Kotsia, I., Zafeiriou, S.: Retinaface: Single-shot multi-level face localisation in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
7. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
8. Du, Y., Zhao, Z., Song, Y., Zhao, Y., Su, F., Gong, T., Meng, H.: Strongsort: Make deepsort great again. IEEE Transactions on Multimedia **25**, 8725–8737 (2023)
9. Fang, Y., Ko, S., Jo, G.S.: Robust visual tracking based on global-and-local search with confidence reliability estimation. Neurocomputing **367**, 273–286 (2019)
10. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: Yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430 (2021)
11. Giancola, S., Amine, M., Dghaily, T., Ghanem, B.: Soccernet: A scalable dataset for action spotting in soccer videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (June 2018)
12. Guo, J., Deng, J., Lattas, A., Zafeiriou, S.: Sample and computation redistribution for efficient face detection. arXiv preprint arXiv:2105.04714 (2021)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
14. Kalman, R.E.: A New Approach to Linear Filtering and Prediction Problems. Journal of Basic Engineering **82**(1), 35–45 (03 1960)
15. Kim, M., Jain, A.K., Liu, X.: Adaface: Quality adaptive margin for face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18750–18759 (June 2022)
16. Kuhn, H.W.: The hungarian method for the assignment problem. Naval Research Logistics Quarterly **2**(1-2), 83–97 (1955)
17. Leal-Taixé, L., Milan, A., Reid, I., Roth, S., Schindler, K.: Motchallenge 2015: Towards a benchmark for multi-target tracking. arXiv preprint arXiv:1504.01942 (2015)
18. Lin, C.C., Hung, Y.: A prior-less method for multi-face tracking in unconstrained videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)

19. Luiten, J., Osep, A., Dendorfer, P., Torr, P., Geiger, A., Leal-Taixé, L., Leibe, B.: Hota: A higher order metric for evaluating multi-object tracking. International journal of computer vision **129**, 548–578 (2021)
20. Luo, W., Xing, J., Milan, A., Zhang, X., Liu, W., Kim, T.K.: Multiple object tracking: A literature review. Artificial Intelligence **293**, 103448 (2021)
21. Maggiolino, G., Ahmad, A., Cao, J., Kitani, K.: Deep oc-sort: Multi-pedestrian tracking by adaptive re-identification. In: 2023 IEEE International Conference on Image Processing (ICIP). pp. 3025–3029 (2023)
22. Milan, A., Leal-Taixé, L., Reid, I., Roth, S., Schindler, K.: Mot16: A benchmark for multi-object tracking. arXiv preprint arXiv:1603.00831 (2016)
23. Pele, O., Werman, M.: The quadratic-chi histogram distance family. In: Computer Vision – ECCV 2010. pp. 749–762. Springer Berlin Heidelberg, Berlin, Heidelberg (2010)
24. Pernici, F., Bartoli, F., Bruni, M., Del Bimbo, A.: Memory based online learning of deep representations from video streams. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
25. Pernici, F., Bruni, M., Del Bimbo, A.: Self-supervised on-line cumulative learning from video streams. Computer Vision and Image Understanding **197-198**, 102983 (2020)
26. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: Computer Vision – ECCV 2016 Workshops. pp. 17–35. Springer International Publishing, Cham (2016)
27. Shao, S., Zhao, Z., Li, B., Xiao, T., Yu, G., Zhang, X., Sun, J.: Crowdhuman: A benchmark for detecting human in a crowd. arxiv 2018. arXiv preprint arXiv:1805.00123 (2018)
28. Sun, P., Cao, J., Jiang, Y., Yuan, Z., Bai, S., Kitani, K., Luo, P.: Dancetrack: Multi-object tracking in uniform appearance and diverse motion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20993–21002 (June 2022)
29. Wang, C.Y., Bochkovskiy, A., Liao, H.Y.M.: Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7464–7475 (June 2023)
30. Wang, Z., Zheng, L., Liu, Y., Li, Y., Wang, S.: Towards real-time multi-object tracking. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. pp. 107–122. Springer International Publishing, Cham (2020)
31. Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. In: 2017 IEEE International Conference on Image Processing (ICIP). pp. 3645–3649 (2017)
32. Wong, Y., Chen, S., Mau, S., Sanderson, C., Lovell, B.C.: Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition. In: CVPR 2011 WORKSHOPS. pp. 74–81. IEEE (2011)
33. Yang, M., Han, G., Yan, B., Zhang, W., Qi, J., Lu, H., Wang, D.: Hybrid-sort: Weak cues matter for online multi-object tracking. Proceedings of the AAAI Conference on Artificial Intelligence **38**(7), 6504–6512 (Mar 2024)
34. Zhang, S., Gong, Y., Huang, J.B., Lim, J., Wang, J., Ahuja, N., Yang, M.H.: Tracking persons-of-interest via adaptive discriminative features. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Computer Vision – ECCV 2016. pp. 415–433. Springer International Publishing, Cham (2016)

35. Zhang, S., Huang, J.B., Lim, J., Gong, Y., Wang, J., Ahuja, N., Yang, M.H.: Tracking persons-of-interest via unsupervised representation adaptation. International Journal of Computer Vision **128**, 96–120 (2020)
36. Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., Wang, X.: Bytetrack: Multi-object tracking by associating every detection box. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022. pp. 1–21. Springer Nature Switzerland, Cham (2022)
37. Zhang, Y., Wang, C., Wang, X., Zeng, W., Liu, W.: Fairmot: On the fairness of detection and re-identification in multiple object tracking. International Journal of Computer Vision **129**, 3069–3087 (2021)
38. Zhu, Z., Huang, G., Deng, J., Ye, Y., Huang, J., Chen, X., Zhu, J., Yang, T., Lu, J., Du, D., Zhou, J.: Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10492–10502 (June 2021)