# Appendix

We here present additional details on the experimental setting (Section A), investigate the effect of domain shift on motion-based tracking (Section B), report additional results and ablations (Section C), and provide extensive qualitative results on the effectiveness of DARTH (Section D).

An additional video teasing DARTH and its TTA efficacy is attached to this submission.

## A. Experimental Setting

All our models are trained with a total batch size of 16 across 8 GPU NVIDIA RTX 2080Ti.

### A.1. Source Model Training

We train QDTrack on the source dataset using the SGD optimizer and a total batch size of 16, starting from an initial learning rate (lr) of 0.01 which is decayed on a dataset-dependent schedule.

**MOT17/DanceTrack.** We train QDTrack on MOT17 and DanceTrack for 4 epochs, decaying the learning rate by a factor of 10 after 3 epochs. We follow the training hyperparameters provided in MMTracking [11]. Images are first rescaled to a random width within [0.8·1088, 1.2·1088] maintaining the original aspect ratio, and horizontally flipped with a probability of 0.5. We then apply an ordered sequence of the following photometric augmentations, each with probability 0.5, following the MMTracking [11] implementation of the SeqPhotoMetricDistortion class with the default parameters: random brightness, random contrast (mode 0), convert color from BGR to HSV, random saturation, random hue, convert color from HSV to BGR, random contrast (mode 1), randomly swap channels. Images are then cropped to a maximum width of 1088. Finally, we normalize images using the reference ImageNet statistics, i.e. channel-wise mean (123.675, 116.28, 103.53) and standard deviation (58.395, 57.12, 57.375). When generating a training batch, all images are padded with zeros on the bottom-right corner to the size of the largest image in the batch.

**SHIFT.** When training on SHIFT, we train for 5 epochs and decay the learning rate by a factor of 10 after 4 epochs. Images are rescaled to the closest size in the set {(1296, 640), (1296, 672), (1296, 704), (1296, 736), (1296, 768), (1296, 800), (1296, 720)} and horizontally flipped with a probability of 0.5. Finally, images are normalized using the reference ImageNet statistics, i.e. channel-wise mean (123.675, 116.28, 103.53) and standard deviation (58.395, 57.12, 57.375). When generating a training batch, all images are padded with zeros on the bottom-right corner to the size of the largest image in the batch.

**BDD100K.** When training on BDD100K, we train for 12 epochs and decay the learning rate by a factor of 10 after 8 and 11 epochs. Images are rescaled to the closest size in the set {(1296, 640), (1296, 672), (1296, 704), (1296, 736), (1296, 768), (1296, 800), (1296, 720)} and horizontally flipped with a probability of 0.5. Finally, images are normalized using the reference ImageNet statistics, i.e. channel-wise mean (123.675, 116.28, 103.53) and standard deviation (58.395, 57.12, 57.375). When generating a training batch, all images are padded with zeros on the bottom-right corner to the size of the largest image in the batch.

### A.2. Adapting to the Target Domain

We train DARTH on the target domain using the SGD optimizer and a total batch size of 16, starting from an initial lr of 0.001 which is decayed on a dataset-dependent schedule. In particular, we train DARTH on MOT17 and DanceTrack for 4 epochs, decaying the learning rate by a factor of 10 after 3 epochs. When training on BDD100K, we train for 10 epochs and decay the learning rate by a factor of 10 after 8 epochs. For each dataset, we adopt the same image normalization parameters as the one used for the original source model.

During the adaptation phase, the teacher model is updated as an EMA of the student weights with a momentum $\tau = 0.998$.

**Data Augmentation.** We here provide details and hyperparameters for the data augmentation transformations employed in the generation of our target, student and contrastive view. To generate the teacher view, we apply a sequence of *geometric transformations*. Images are first rescaled to a random width within [0.8·1088, 1.2·1088] maintaining the original aspect ratio, and then cropped to a maximum width of 1088 pixels. Random horizontal flipping is also applied with a probability of 0.5. When generating a training batch, all images are padded with zeros on the bottom-right corner to the size of the largest image in the batch. Given the teacher view, we generate the student view by consecutive application of *photometric augmentations*. Generating the student view from the teacher view is necessary to ensure geometric consistency between teacher and student views, as required by our detection consistency losses (Section 3.4). In particular, we apply an ordered sequence of the following augmentations, each with probability 0.5, following the MMTracking [11] implementation of the SeqPhotoMetricDistortion class with the default parameters: random brightness, random contrast (mode 0), convert color from BGR to HSV, random saturation, random hue, convert color from HSV to BGR, random contrast (mode 1), randomly swap channels. The contrastive view is generated using the same strategy as the student view but from independently sampled parameters of the geometric and photometric augmentations.

Table 8. **Appearance-based MOT** (QDTrack [45])

| Source | Target | DetA | MOTA | HOTA | IDF1 | AssA |
|---|---|---|---|---|---|---|
| SHIFT | SHIFT | 46.9 | 48.4 | 55.2 | 60.6 | 65.8 |
| | BDD100K | 12.0 | -66.4 | 17.3 | 18.5 | 28.9 |
| MOT17 | MOT17 | 57.2 | 68.2 | 57.1 | 68.5 | 57.4 |
| | DanceTrack | 52.4 | 57.2 | 21.5 | 19.5 | 9.0 |
| | BDD100K | 23.2 | 10.5 | 27.2 | 33.3 | 32.4 |
| MOT17 (+CH) | MOT17 | 59.8 | 71.7 | 59.7 | 71.6 | 58.7 |
| | DanceTrack | 61.8 | 74.0 | 31.1 | 29.6 | 15.8 |
| | BDD100K | 32.4 | 28.3 | 33.7 | 41.7 | 35.4 |
| DanceTrack | DanceTrack | 68.5 | 79.2 | 43.5 | 42.3 | 28.0 |
| | MOT17 | 24.7 | 23.3 | 32.6 | 35.4 | 43.5 |
| | BDD100K | 9.3 | -16.0 | 14.1 | 12.3 | 21.8 |
| BDD100K | BDD100K | 36.5 | 14.2 | 39.6 | 48.2 | 43.3 |
| | MOT17 | 28.6 | 31.4 | 36.0 | 43.5 | 45.8 |
| | DanceTrack | 41.9 | 41.6 | 18.0 | 17.0 | 7.9 |

Table 9. **Motion-based MOT** (ByteTrack† [75])

| Source | Target | DetA | MOTA | HOTA | IDF1 | AssA |
|---|---|---|---|---|---|---|
| SHIFT | SHIFT | 46.7 | 46.6 | 55.1 | 60.6 | 65.7 |
| | BDD100K | 11.8 | -70.5 | 15.2 | 14.8 | 23.4 |
| MOT17 | MOT17 | 56.7 | 65.8 | 57.5 | 68.9 | 58.9 |
| | DanceTrack | 52.2 | 62.2 | 31.6 | 35.5 | 19.4 |
| | BDD100K | 22.6 | -12.0 | 21.3 | 22.4 | 20.5 |
| MOT17 (+ CH) | MOT17 | 60.0 | 70.3 | 58.8 | 71.4 | 58.1 |
| | DanceTrack | 61.1 | 75.2 | 36.1 | 38.9 | 21.5 |
| | BDD100K | 32.9 | 8.2 | 27.9 | 30.4 | 24.0 |
| DanceTrack | DanceTrack | 65.9 | 77.8 | 40.4 | 41.5 | 25.0 |
| | MOT17 | 25.3 | 21.6 | 34.4 | 38.2 | 47.3 |
| | BDD100K | 7.6 | -19.2 | 13.1 | 10.0 | 22.9 |
| BDD100K | BDD100K | 35.8 | 9.4 | 29.1 | 31.9 | 24.0 |
| | MOT17 | 31.0 | 29.5 | 36.3 | 43.8 | 43.2 |
| | DanceTrack | 43.7 | 44.6 | 25.2 | 27.1 | 14.7 |

Table 10. **Domain shift in MOT.** We assess the impact of domain shift on appearance-based (QDTrack [45], left), and motion-based (ByteTrack [75], right) MOT. † indicates that we use the motion-only version of ByteTrack. We compare both trackers using a Faster R-CNN [48] object detector with a ResNet-50 [24] backbone and FPN [34]. In green the performance on the source domain. The SHIFT $\rightarrow$ BDD100K metrics are averaged across all object categories; only the pedestrian category is considered for other experiments. CH: CrowdHuman. The in-domain performance is aligned for both trackers, although QDTrack excels on the complex BDD100K [70]. Domain shift affects equally the DetA of both trackers, while threatening more the AssA of appearance-based MOT.

## B. Domain Shift in Motion-based MOT

We here study the effect of domain shift on motion-based MOT, and justify the importance of solving domain adaptation for appearance-based tracking. Motion- [3, 4, 18, 6, 75], appearance- [31, 66, 1, 45], and query-based [38, 59, 71] trackers are commonly used to associate instances detected by an object detector. Appearance-based tracking has proven the most versatile formulation, showing SOTA performance on a variety of benchmarks [19] and complementing motion cues for superior tracking performance [75]. On the other hand, motion-based tracking achieves competitive performance on datasets with high frame rates and low relative speed of tracked objects, while failing on complex datasets (e.g. BDD100K [19]) or on any domain at lower frame rates ([19], Fig. 3).

### B.1. Domain Shift in Appearance- and Motion-based Multiple Object Tracking

Intuitively, all categories (appearance-, motion-, and query-based) suffer from domain shift in their detection stage. Moreover, query-based tracking can be seen as an instance of appearance-based, where the queries serve as appearance representation. We study in Table 10 the effect of domain shift on appearance- and motion-based tracking.

We choose QDTrack [45] as representative of appearance-only tracking as it provides the most effective formulation [19] to learn appearance representations for downstream instance association. We choose ByteTrack [75] as representative of motion-only tracking, as its motion-based matching scheme reports state-of-the-art performance. Although ByteTrack can also be extended to use appearance-cues, for the scope of this comparison we only use its motion component, as we intend to disentangle the effect of domain shift on appearance-only and motion-only MOT. In our experiments, we compare both tracking algorithms using a Faster R-CNN [48] object detector with a ResNet-50 [24] backbone and FPN [34]. We choose the same detector for a fair comparison.

**In-domain Comparison.** Table 10 shows that both QDTrack (left) and the motion-only version of ByteTrack (right) obtain comparable in-domain performance (green rows) on almost all datasets. However, motion-based tracking suffers from the complexity and low frame rate of BDD100K, making a case for the use of appearance-based trackers in complex scenarios.

**Domain Shift Comparison.** Despite the superior versatility of appearance-based trackers, we find (Table 10, left) that appearance-based tracking suffers from domain shift in both its detection and instance association stage, due to the learning-based nature of the object detector and the appearance embedding head. On the other hand, motion-based tracking is affected less by domain shift in its data association stage. In particular, we observe that (1) the in-domain performance is aligned for both trackers, except on BDD100K, highlighting that appearance-based trackers work best in complex scenarios; (2) the drop in DetA under domain shift is comparable for both types of trackers; (3) except when shifting to BDD100K, the motion-based ByteTrack generally retains higher AssA than the appearance-based QDTrack under domain shift. This highlights the importance of domain adaptation for appearance-based MOT. Although appearance-based MOT achieves SOTA perfor-

| Method | Source | Target | DetA | MOTA | HOTA | IDF1 | AssA |
|---|---|---|---|---|---|---|---|
| QDTrack [45] | SHIFT | BDD100K | 12.0 | -66.4 | 17.3 | 18.5 | 28.9 |
| ByteTrack [75] | | | 11.8 | -70.5 | 15.2 | 14.8 | 23.4 |
| DARTH | | | **15.2** | **8.3** | **20.6** | **23.7** | **33.1** |
| QDTrack [45] | MOT17 | DT | 52.4 | 57.2 | 21.5 | 19.5 | 9.0 |
| ByteTrack [75] | | | 52.2 | 62.2 | **31.6** | **35.5** | **19.4** |
| DARTH | | | **57.2** | **70.1** | **31.6** | 32.8 | 17.7 |
| QDTrack [45] | MOT17 (+ CH) | DT | 61.8 | 74.0 | 31.1 | 29.6 | 15.8 |
| ByteTrack [75] | | | 61.1 | 75.2 | **36.1** | **38.9** | **21.5** |
| DARTH | | | **64.7** | **78.9** | 35.4 | 35.3 | 19.6 |
| QDTrack [45] | DT | MOT17 | 24.7 | 23.3 | 32.6 | 35.4 | 43.5 |
| ByteTrack [75] | | | 25.3 | 21.6 | **34.4** | **38.2** | **47.3** |
| DARTH | | | **26.4** | **25.5** | 34.3 | 37.9 | 45.2 |
| QDTrack [45] | BDD100K | MOT17 | 28.6 | 31.4 | 36.0 | 43.5 | 45.8 |
| ByteTrack [75] | | | **31.0** | 29.5 | 36.3 | 43.8 | 43.2 |
| DARTH | | | 29.4 | **32.6** | **36.6** | **44.4** | **45.9** |
| QDTrack [45] | BDD100K | DT | 41.9 | 41.6 | 18.0 | 17.0 | 7.9 |
| ByteTrack [75] | | | 43.7 | 44.6 | **25.2** | **27.1** | **14.7** |
| DARTH | | | **45.1** | **50.2** | 21.5 | 21.4 | 10.4 |
| QDTrack [45] | DT | BDD100K | 9.3 | -16.0 | 14.1 | 12.3 | 21.8 |
| ByteTrack [75] | | | 7.6 | -19.2 | 13.1 | 10.0 | 22.9 |
| DARTH | | | **12.8** | **-1.5** | **17.8** | **17.4** | **25.1** |
| QDTrack [45] | MOT17 | BDD100K | 23.2 | 10.5 | 27.2 | 33.3 | 32.4 |
| ByteTrack [75] | | | 22.6 | -12.0 | 21.3 | 22.4 | 20.5 |
| DARTH | | | **31.6** | **21.4** | **32.4** | **40.4** | **33.6** |
| QDTrack [45] | MOT17 (+ CH) | BDD100K | 32.4 | 28.3 | 33.7 | 41.7 | 35.4 |
| ByteTrack [75] | | | 32.9 | 8.2 | 27.9 | 30.4 | 24.0 |
| DARTH | | | **36.3** | **23.4** | **36.3** | **44.4** | **36.8** |

Table 11. **Comparison of appearance- and motion-based MOT under domain shift.** We compare the performance under domain shift of appearance-based (QDTrack), motion-based (ByteTrack), and domain adaptive appearance-based (DARTH, ours) MOT. We use the motion-only version of ByteTrack. Both trackers use a Faster R-CNN [48] object detector with a ResNet-50 [24] backbone and FPN [34]. The SHIFT → BDD100K metrics are averaged across all categories; only the pedestrian category is considered in other experiments. DT: DanceTrack; CH: CrowdHuman.

mance in-domain, it suffers significantly more from domain shift, making a solution to the adaptation problem desirable.

**Recovering Appearance-based MOT.** We now investigate whether our proposed method (DARTH) can recover the performance of appearance-based trackers under domain shift, closing the gap with motion-based trackers under domain shift or even outperforming them. Table 11 compares the performance of QDTrack (appearance-based), ByteTrack (motion-based), and DARTH (domain-adaptive QDTrack) on the shifted domain. DARTH consistently outperforms DetA and MOTA of both QDTrack and ByteTrack. Moreover, it considerably recovers the AssA of QDTrack, outperforming also ByteTrack on shifts to BDD100K and reporting competitive performance to it on pedestrian datasets. Such results highlight the effectiveness of our proposed method DARTH, making a case for the use of our domain adaptive appearance-based tracker under domain shift instead of motion-based ones.

# C. Additional Results

We extend Section 4 with additional results.

## C.1. Extension of the Ablation Study

**SHIFT → BDD100K (Overall).** We here complement the main manuscript results by reporting the Overall performance on the SHIFT → BDD100K experiments. By Overall we mean that for each metric we report the results over all the identities available in the dataset and across all categories, as opposed to the Average results reported in the main paper which are averaged over the category-specific metrics. We make the choice of reporting the Average performance in the main paper because we believe that it is significant towards the evaluation of TTA in a class-imbalanced setting. Nevertheless, we here report the absolute performance over the whole dataset for completeness. Table 12 confirms the superiority of DARTH over the considered baselines; Table 13 confirms that our chosen augmentation policy outperforms all possible alternatives; Table 14 confirms the effectiveness and complementarity of each of our method components.

**MOT17 → DanceTrack.** We extend the ablations on method components (Table 15) and data augmentation settings (Table 16) to the MOT17 → DanceTrack setting, further confirming the findings reported in Section 4.3.

## C.2. Ablation on Confidence Threshold

We ablate on the sensitivity to the confidence threshold value in SFOD and DARTH on SHIFT → BDD100K and MOT17 → DanceTrack. Notice that SFOD and DARTH use the threshold differently. SFOD uses it to only retain high-confidence detections as pseudo-labels for self-training the detector. DARTH leverages a confidence threshold over the teacher detections to identify the object regions used in our patch contrastive learning formulation, as described in Section 3.3 and illustrated in Figure 3.

**SFOD.** We report the average (Table 17) and overall (Table 18) performance of SFOD under different thresholds on the SHIFT → BDD100K setting, and find that SFOD is highly sensitive to the confidence threshold choice. In particular, the average performance always worsens except when the threshold is set at 0.7, while the overall performance improves also with a threshold of 0.5. This indicates that domain shift impacts differently each category and a unique threshold for all categories is suboptimal.

**DARTH.** First, we report the average (Table 17) and overall (Table 18) performance of DARTH under different thresholds on the SHIFT → BDD100K setting, and find that DARTH is highly sensitive to the confidence threshold choice. Table 19 Table 20 The same trend is confirmed on the MOT17 → DanceTrack setting (Table 21).

| Method | Source | Target | DetA | MOTA | HOTA | IDF1 | AssA |
|---|---|---|---|---|---|---|---|
| No Adap. | SHIFT | BDD100K | 27.2 | 20.4 | 35.1 | 39.5 | 46.4 |
| Tent [63] | | | 0.3 | 0.2 | 1.9 | 0.5 | 14.8 |
| SFOD [33] | | | 27.7 | 22.7 | 35.7 | 40.0 | 47.1 |
| Ours | | | **36.5** | **33.3** | **43.1** | **50.9** | **51.8** |
| Oracle | BDD100K | BDD100K | 55.9 | 58.5 | 59.7 | 69.2 | 64.6 |

Table 12. **State of the art on SHIFT → BDD100K (Overall).** We benchmark DARTH (ours) against baseline test-time adaptation methods for adapting a MOT model from the synthetic driving dataset SHIFT to the real-world BDD100K. For each metric we report the overall result across all categories.

| Teacher | Student | Contrastive | DetA | MOTA | HOTA | IDF1 | AssA |
|---|---|---|---|---|---|---|---|
| - | - | - | 27.2 | 20.4 | 35.1 | 39.5 | 46.4 |
| - | - | - | 26.8 | 12.5 | 27.7 | 25.8 | 29.7 |
| g | - | g | 31.4 | 28.5 | 39.2 | 45.2 | 50.0 |
| g | - | g + p | 31.2 | 28.8 | 39.0 | 45.1 | 49.6 |
| g + p | - | g + p | 30.3 | 27.9 | 38.5 | 44.3 | 49.8 |
| g | p | g | **37.0** | 32.8 | 43.2 | 50.8 | 51.6 |
| g | p | g + p | 36.5 | **33.3** | **43.1** | **50.9** | **51.8** |

Table 13. **Ablation study on different data augmentation settings for DARTH (Overall).** We analyze the effect of different data augmentation settings on DARTH on SHIFT → BDD100K. We report the augmentations applied on the Teacher, Student and Contrastive view, chosen from geometric (g) and photometric (p) augmentations as detailed in Section 3.2. For each metric we report the overall result across all categories. No Adap. is in gray.

| EMA | DC | PCL | DetA | MOTA | HOTA | IDF1 | AssA |
|---|---|---|---|---|---|---|---|
| - | - | - | 27.2 | 20.4 | 35.1 | 39.5 | 46.4 |
| - | - | ✓ | 23.8 | 8.3 | 29.6 | 34.7 | 37.6 |
| - | ✓ | - | 28.0 | 23.0 | 36.1 | 40.6 | 47.6 |
| ✓ | ✓ | - | 33.8 | 32.0 | 40.8 | 46.9 | 50.3 |
| ✓ | ✓ | ✓ | **36.5** | **33.3** | **43.1** | **50.9** | **51.8** |

Table 14. **Ablation study on the impact of different method components on DARTH (Overall).** We analyze the effect of different method components on DARTH (ours) on SHIFT → BDD100K. We report with a ✓whether exponential moving average (EMA), detection consistency (DC) and Patch Contrastive Learning (PCL) are applied. For each metric we report the overall result across all categories. No Adap. is in gray.

| EMA | DC | PCL | DetA | MOTA | HOTA | IDF1 | AssA |
|---|---|---|---|---|---|---|---|
| - | - | - | 52.4 | 57.2 | 21.5 | 19.5 | 9.0 |
| - | - | ✓ | 51.2 | 54.1 | 28.3 | 28.6 | 16.0 |
| - | ✓ | - | 52.7 | 58.0 | 21.8 | 19.7 | 9.2 |
| ✓ | ✓ | - | 55.3 | 62.0 | 23.3 | 21.4 | 10.0 |
| ✓ | ✓ | ✓ | **57.2** | **70.1** | **31.6** | **32.8** | **17.7** |

Table 15. **Ablation study on the impact of different method components on DARTH (MOT17 → DanceTrack).** We analyze the effect of different method components on DARTH (ours) on MOT17 → DanceTrack. We report with a ✓whether exponential moving average (EMA), detection consistency (DC) and Patch Contrastive Learning (PCL) are applied. No Adap. is in gray.

| Teacher | Student | Contrastive | DetA | MOTA | HOTA | IDF1 | AssA |
|---|---|---|---|---|---|---|---|
| - | - | - | 52.4 | 57.2 | 21.5 | 19.5 | 9.0 |
| - | - | - | 52.5 | 29.9 | 12.4 | 9.2 | 3.1 |
| g | - | g | 54.7 | 66.9 | 30.8 | 32.2 | 17.6 |
| g | - | g + p | 54.7 | 66.9 | 31.5 | **33.6** | **18.3** |
| g + p | - | g + p | 54.6 | 66.7 | 30.7 | 32.2 | 17.5 |
| g | p | g + p | **57.2** | **70.1** | **31.6** | 32.8 | 17.7 |

Table 16. **Ablation study on different data augmentation settings for DARTH (MOT17 → DanceTrack).** We analyze the effect of different data augmentation settings on DARTH on MOT17 → DanceTrack. We report the augmentations applied on the Teacher, Student and Contrastive view, chosen from geometric (g) and photometric (p) augmentations as detailed in Section 3.2. No Adap. is in gray.

| Conf. Thr. | DetA | MOTA | HOTA | IDF1 | AssA |
|---|---|---|---|---|---|
| - | 12.0 | -66.4 | 17.3 | 18.5 | 28.9 |
| 0.0 | 7.9 | -841.7 | 12.8 | 10.8 | 28.5 |
| 0.3 | 11.2 | -258.2 | 16.2 | 16.2 | 29.2 |
| 0.5 | 12.0 | -135.1 | 17.2 | 17.8 | **29.6** |
| 0.7 | **12.4** | -57.3 | **17.7** | 19.0 | 29.1 |
| 0.9 | 11.9 | **-5.4** | 17.5 | **19.3** | 28.7 |

Table 17. **Ablation study on confidence thresholds for SFOD [33] (Average).** We analyze the sensitivity of SFOD to different confidence thresholds for the detection pseudo labels filtering on SHIFT → BDD100K. For each metric we report its average across all object categories. No Adap. is in gray.

| Conf. Thr. | DetA | MOTA | HOTA | IDF1 | AssA |
|---|---|---|---|---|---|
| - | 27.2 | 20.4 | 35.1 | 39.5 | 46.4 |
| 0.0 | 19.4 | -81.4 | 27.8 | 26.3 | 41.9 |
| 0.3 | 27.0 | 1.9 | 34.4 | 37.5 | 45.3 |
| 0.5 | **27.8** | 15.2 | 35.6 | 39.5 | 46.7 |
| 0.7 | 27.7 | 22.7 | **35.7** | **40.0** | 47.1 |
| 0.9 | 25.0 | **25.2** | 34.4 | 37.7 | **48.1** |

Table 18. **Ablation study on confidence thresholds for SFOD [33] (Overall).** We analyze the sensitivity of SFOD to different confidence thresholds for the detection pseudo labels filtering on SHIFT → BDD100K. For each metric we report the overall result across all categories. No Adap. is in gray.

| Conf. Thr. | DetA | MOTA | HOTA | IDF1 | AssA |
|---|---|---|---|---|---|
| - | 12.0 | -66.4 | 17.3 | 18.5 | 28.9 |
| 0.0 | 14.6 | 5.2 | 19.8 | 22.2 | 31.4 |
| 0.3 | 14.9 | 7.8 | 20.0 | 22.8 | 31.7 |
| 0.5 | **15.2** | 7.6 | 20.3 | 23.0 | 32.2 |
| 0.7 | **15.2** | **8.3** | **20.6** | **23.7** | **33.1** |
| 0.9 | 14.7 | 7.5 | 19.6 | 22.3 | 31.4 |

Table 19. **Ablation study on confidence thresholds for DARTH (Average).** We analyze the sensitivity of DARTH (Ours) to different confidence thresholds for filtering detection in our self-matching stage on SHIFT → BDD100K. For each metric we report its average across all object categories. No Adap. is in gray.

| Conf. Thr. | DetA | MOTA | HOTA | IDF1 | AssA |
|---|---|---|---|---|---|
| - | 27.2 | 20.4 | 35.1 | 39.5 | 46.4 |
| 0.0 | 35.2 | 32.5 | 42.2 | 49.4 | 51.7 |
| 0.3 | 36.2 | 33.2 | **43.2** | **50.9** | **52.5** |
| 0.5 | **36.6** | **33.3** | 43.0 | 50.8 | 51.7 |
| 0.7 | 36.5 | **33.3** | 43.1 | **50.9** | 51.8 |
| 0.9 | 36.4 | 32.7 | 42.8 | 50.2 | 51.2 |

Table 20. **Ablation study on confidence thresholds for DARTH (Overall).** We analyze the sensitivity of DARTH (Ours) to different confidence thresholds for filtering detection in our self-matching stage on SHIFT → BDD100K. For each metric we report the overall result across all categories. No Adap. is in gray.

| Conf. Thr. | DetA | MOTA | HOTA | IDF1 | AssA |
|---|---|---|---|---|---|
| - | 52.4 | 57.2 | 21.5 | 19.5 | 9.0 |
| 0.0 | 56.4 | 68.4 | 30.1 | 30.8 | 16.3 |
| 0.3 | 56.6 | 69.5 | 31.6 | 33.0 | 17.9 |
| 0.5 | 56.8 | 69.4 | 31.7 | 32.9 | 17.9 |
| 0.7 | **57.2** | **70.1** | 31.6 | 32.8 | 17.7 |
| 0.9 | 57.0 | **70.1** | **32.0** | **33.5** | **18.2** |

Table 21. **Ablation study on confidence thresholds for DARTH (MOT17 → DanceTrack).** We analyze the sensitivity of DARTH (Ours) to different confidence thresholds for filtering detections in our self-matching stage on MOT17 → DanceTrack. No Adap. is in gray.

| Momentum | DetA | MOTA | HOTA | IDF1 | AssA |
|---|---|---|---|---|---|
| - | 12.0 | -66.4 | 17.3 | 18.5 | 28.9 |
| 1.0 | 12.8 | -32.1 | 17.9 | 19.4 | 28.5 |
| 0.998 | **15.2** | **8.3** | **20.6** | **23.7** | **33.1** |
| 0.98 | 5.9 | -21.6 | 9.1 | 9.3 | 17.5 |

Table 22. **Ablation study on EMA momentum for DARTH (Average).** We analyze the sensitivity of DARTH (Ours) to different values of the EMA momentum used to update the teacher on SHIFT → BDD100K. For each metric we report its average across all object categories. No Adap. is in gray.

| Momentum | DetA | MOTA | HOTA | IDF1 | AssA |
|---|---|---|---|---|---|
| - | 27.2 | 20.4 | 35.1 | 39.5 | 46.4 |
| 1.0 | 28.2 | 23.5 | 36.3 | 41.1 | 47.8 |
| 0.998 | **36.5** | **33.3** | **43.1** | **50.9** | **51.8** |
| 0.98 | 17.3 | -102.9 | 26.8 | 26.0 | 43.4 |

Table 23. **Ablation study on EMA momentum for DARTH (Overall).** We analyze the sensitivity of DARTH (Ours) to different values of the EMA momentum used to update the teacher on SHIFT → BDD100K. For each metric we report the overall result across all categories. No Adap. is in gray.

### C.3. Ablation on EMA Momentum.

We ablate on the effect on DARTH of different momentum choices for the EMA update of the teacher model, as described in Section 3.2. We report the average (Table 17) and overall (Table 18) performance of DARTH under different momentum values on the SHIFT → BDD100K setting. We find that, while DARTH improves the baseline performance also with a frozen teacher (momentum 1.0), a suitable choice of the momentum (momentum 0.998) allows to incorporate in the teacher model the improved student weights and provide better targets for the detection consistency loss, remarkably boosting the overall performance. However, if the update to the teacher is too fast (momentum 0.98), we hypothesize that the encoder and its adapted representations may update the teacher too quickly and deviate from the expected distribution to the detection head.

## D. Qualitative Results

We provide extensive qualitative results on the effectiveness of DARTH on the MOT17 → DanceTrack and SHIFT → BDD100K settings. In particular, we compare the No Adap. baseline and DARTH by visualizing representative examples of their tracking results, their false negative detections, and their ID switches. For each method, we show 5 adjacent frames.

### D.1. MOT17 → DanceTrack

We compare the No Adap. baseline and DARTH on the MOT17 → DanceTrack setting, providing qualitative results on how DARTH can recover false negative detections and ID switches.

**Recovering False Negative Detections.** We analyze two crowded scenes and visualize for each the tracking results, the false positive detections, and the ID switches: (Figures 5 to 7), and (Figures 8 to 10). It appears evident in Figure 6 and Figure 9 how DARTH drastically recovers false negative detections (orange) by identifying correct matches (green). At the same time, even though DARTH is able to detect and track more objects, also the number of ID switches reduces (Figures 7 and 10), hinting at the improved association performance.

**Recovering ID Switches.** We further consider a variety of scenes with a reduced amount of objects where the No Adap. baseline already does not suffer from false negatives, and show how DARTH drastically reduces ID switches. This can be seen on the following pairs of tracking results and visualizations of ID switches: (Figures 11 and 12), (Figures 13 and 14), (Figures 15 and 16), and (Figures 17 and 18). In most of these cases, DARTH does not suffer ID switches in the considered frames, as opposed to the No Adap. baseline. Nevertheless, an example of ID switch (blue) with DARTH can be identified in Figure 18 at $t{=}\hat{t} + k$, where an ID switches when two dancers switch position and overlap with each other.

### D.2. SHIFT → BDD100K

We compare the No Adap. baseline and DARTH on the SHIFT → BDD100K setting, providing qualitative results on how DARTH can recover false negative detections and ID switches.

**Recovering False Negative Detections.** We show examples of tracking results and the respective visualization of false negative detections in (Figures 19 and 20), (Figures 21 and 22), (Figures 23 and 24), and (Figures 25 and 26). DARTH is able to recover a large amount of false negative detections, especially on the road side vehicles, and correctly track them through time.

**Recovering ID Switches.** We show examples of tracking results and the respective visualization of ID switches in (Figures 27 and 28), (Figures 29 and 30), and (Figures 31 and 32). DARTH reduces the number of ID switches, consistently detect objects through time and correctly assigns them to the same tracklet.

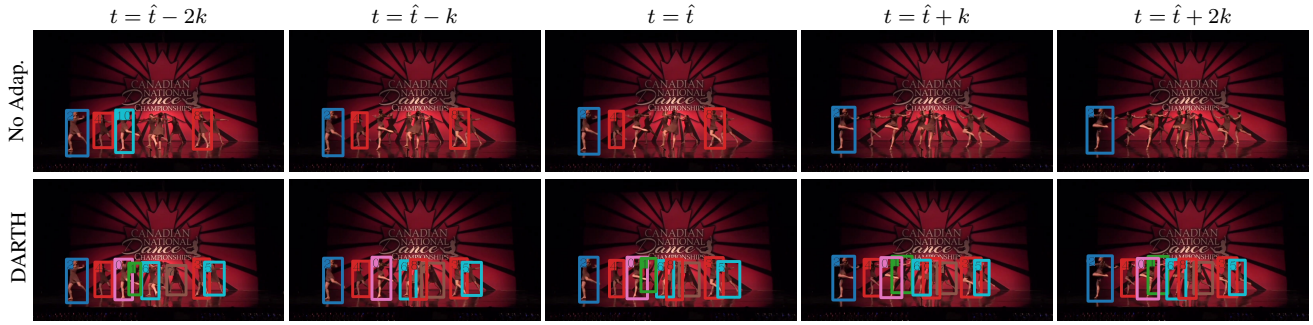


Figure 5. Tracking results on the sequence *0025* of the DanceTrack validation set in the adaptation setting MOT17 $\rightarrow$ DanceTrack. We analyze 5 consecutive frames centered around the frame #28 at time $\hat{t}$ and spaced by $k = 0.05$ seconds. We visualize the No Adap. baseline (top row) and DARTH (bottom row). On each row, boxes of the same color correspond to the same tracking ID.

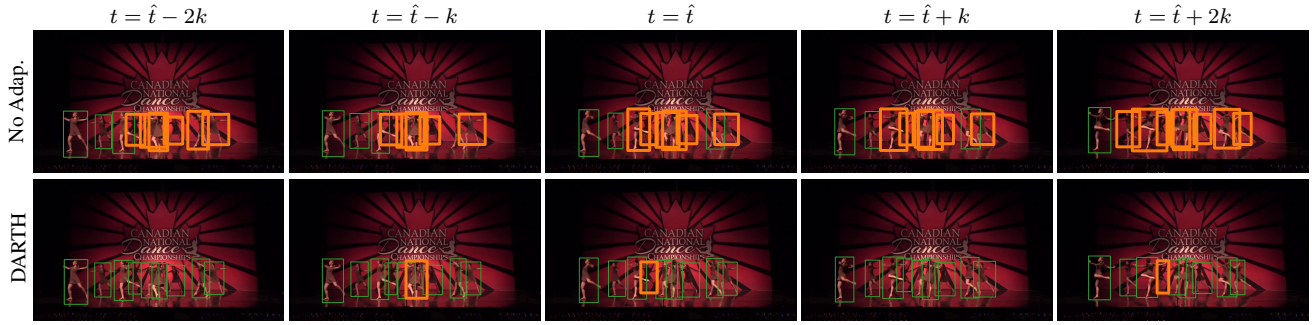


Figure 6. Tracking results on the sequence *0025* of the DanceTrack validation set in the adaptation setting MOT17 $\rightarrow$ DanceTrack. We analyze 5 consecutive frames centered around the frame #28 at time $\hat{t}$ and spaced by $k = 0.05$ seconds. We visualize the No Adap. baseline (top row) and DARTH (bottom row). On each row, green boxes represent correctly tracked objects, and orange boxes represent false negatives. We omit false positive boxes and ID switches for ease of visualization.

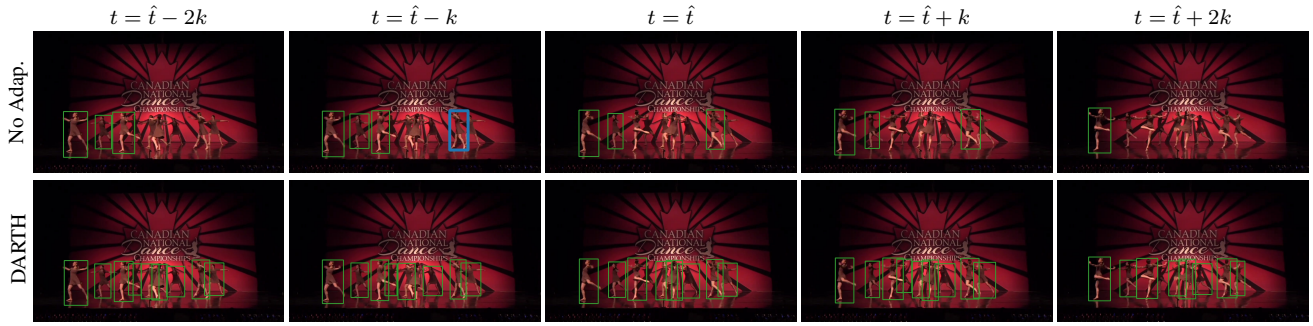


Figure 7. Tracking results on the sequence *0025* of the DanceTrack validation set in the adaptation setting MOT17 $\rightarrow$ DanceTrack. We analyze 5 consecutive frames centered around the frame #28 at time $\hat{t}$ and spaced by $k = 0.05$ seconds. We visualize the No Adap. baseline (top row) and DARTH (bottom row). On each row, green boxes represent correctly tracked objects, and blue boxes represent ID switches. We omit false positive and false negative boxes for ease of visualization.

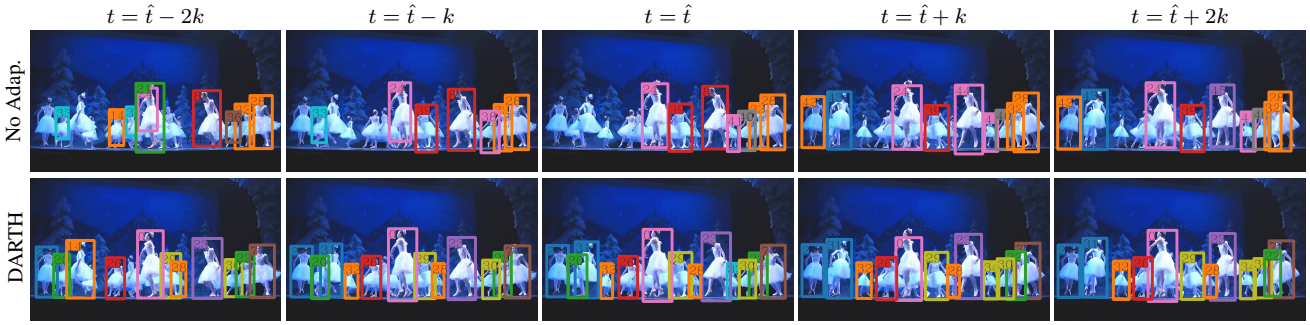


Figure 8. Tracking results on the sequence *0026* of the DanceTrack validation set in the adaptation setting MOT17 $\rightarrow$ DanceTrack. We analyze 5 consecutive frames centered around the frame #54 at time $\hat{t}$ and spaced by $k = 0.05$ seconds. We visualize the No Adap. baseline (top row) and DARTH (bottom row). On each row, boxes of the same color correspond to the same tracking ID.

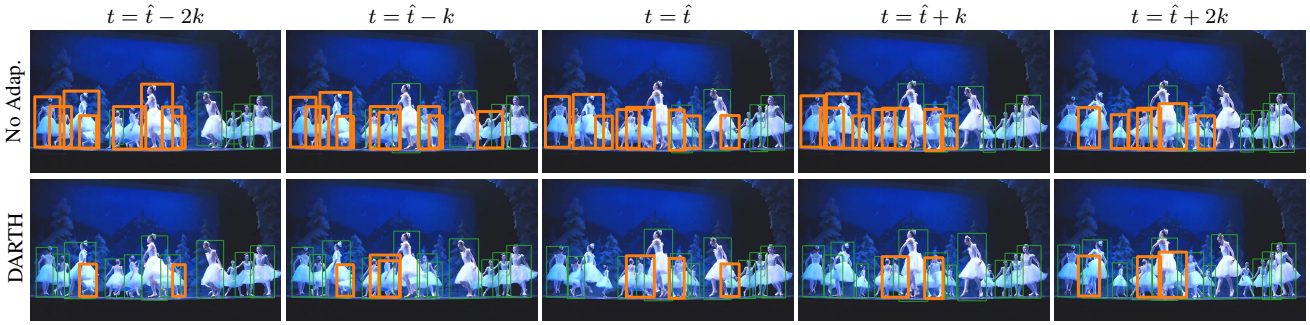


Figure 9. Tracking results on the sequence *0026* of the DanceTrack validation set in the adaptation setting MOT17 $\rightarrow$ DanceTrack. We analyze 5 consecutive frames centered around the frame #54 at time $\hat{t}$ and spaced by $k = 0.05$ seconds. We visualize the No Adap. baseline (top row) and DARTH (bottom row). On each row, green boxes represent correctly tracked objects, and orange boxes represent false negatives. We omit false positive boxes and ID switches for ease of visualization.

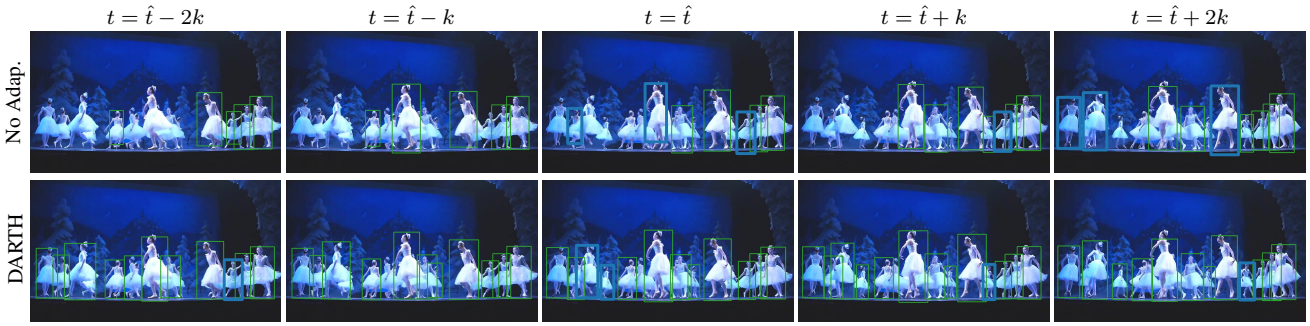


Figure 10. Tracking results on the sequence *0026* of the DanceTrack validation set in the adaptation setting MOT17 $\rightarrow$ DanceTrack. We analyze 5 consecutive frames centered around the frame #54 at time $\hat{t}$ and spaced by $k = 0.05$ seconds. We visualize the No Adap. baseline (top row) and DARTH (bottom row). On each row, green boxes represent correctly tracked objects, and blue boxes represent ID switches. We omit false positive and false negative boxes for ease of visualization.

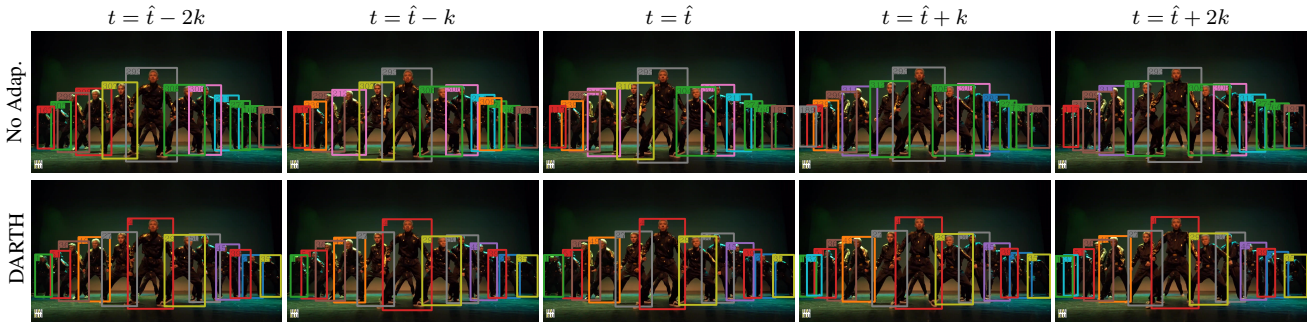


Figure 11. Tracking results on the sequence *0034* of the DanceTrack validation set in the adaptation setting MOT17 $\rightarrow$ DanceTrack. We analyze 5 consecutive frames centered around the frame #143 at time $\hat{t}$ and spaced by $k = 0.05$ seconds. We visualize the No Adap. baseline (top row) and DARTH (bottom row). On each row, boxes of the same color correspond to the same tracking ID.

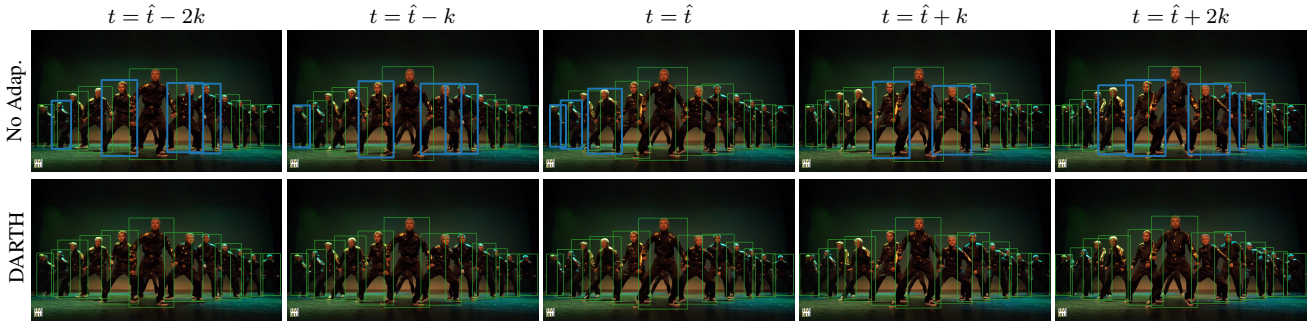


Figure 12. Tracking results on the sequence *0034* of the DanceTrack validation set in the adaptation setting MOT17 $\rightarrow$ DanceTrack. We analyze 5 consecutive frames centered around the frame #143 at time $\hat{t}$ and spaced by $k = 0.05$ seconds. We visualize the No Adap. baseline (top row) and DARTH (bottom row). On each row, green boxes represent correctly tracked objects, and blue boxes represent ID switches. We omit false positive and false negative boxes for ease of visualization.

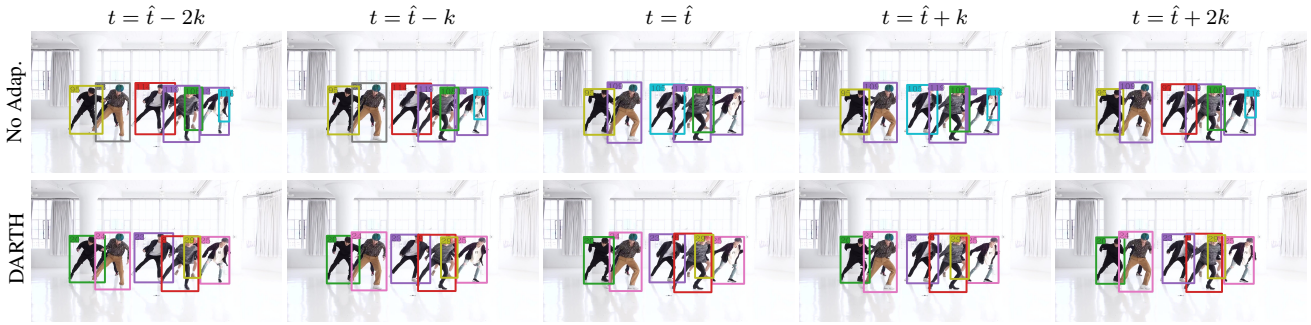


Figure 13. Tracking results on the sequence *0058* of the DanceTrack validation set in the adaptation setting MOT17 $\rightarrow$ DanceTrack. We analyze 5 consecutive frames centered around the frame #783 at time $\hat{t}$ and spaced by $k = 0.05$ seconds. We visualize the No Adap. baseline (top row) and DARTH (bottom row). On each row, boxes of the same color correspond to the same tracking ID.

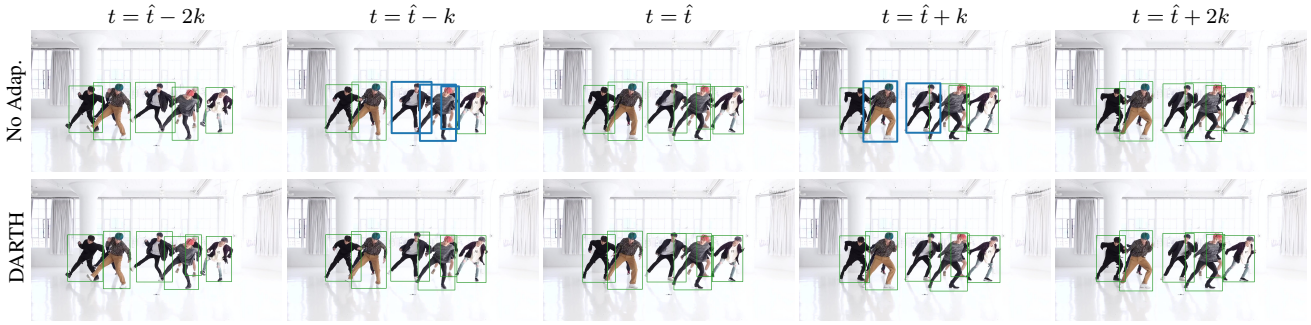


Figure 14. Tracking results on the sequence *0058* of the DanceTrack validation set in the adaptation setting MOT17 $\rightarrow$ DanceTrack. We analyze 5 consecutive frames centered around the frame #783 at time $\hat{t}$ and spaced by $k = 0.05$ seconds. We visualize the No Adap. baseline (top row) and DARTH (bottom row). On each row, green boxes represent correctly tracked objects, and blue boxes represent ID switches. We omit false positive and false negative boxes for ease of visualization.

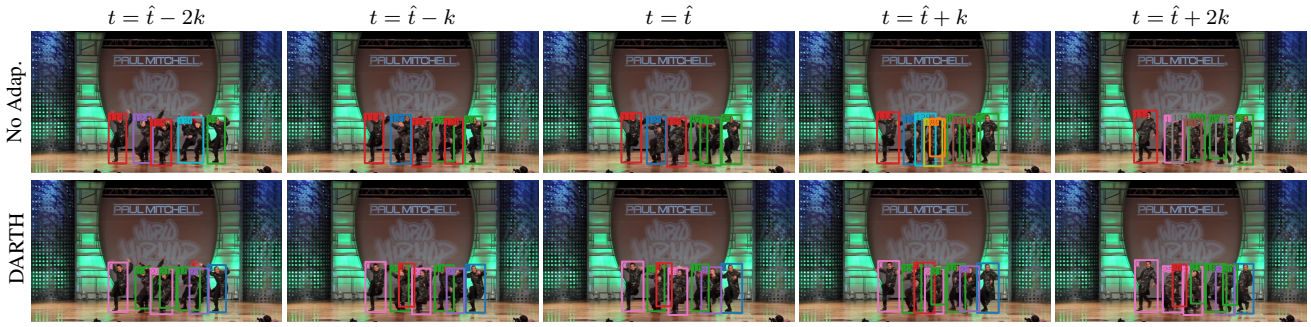


Figure 15. Tracking results on the sequence *0035* of the DanceTrack validation set in the adaptation setting MOT17 $\rightarrow$ DanceTrack. We analyze 5 consecutive frames centered around the frame #248 at time $\hat{t}$ and spaced by $k = 0.05$ seconds. We visualize the No Adap. baseline (top row) and DARTH (bottom row). On each row, boxes of the same color correspond to the same tracking ID.

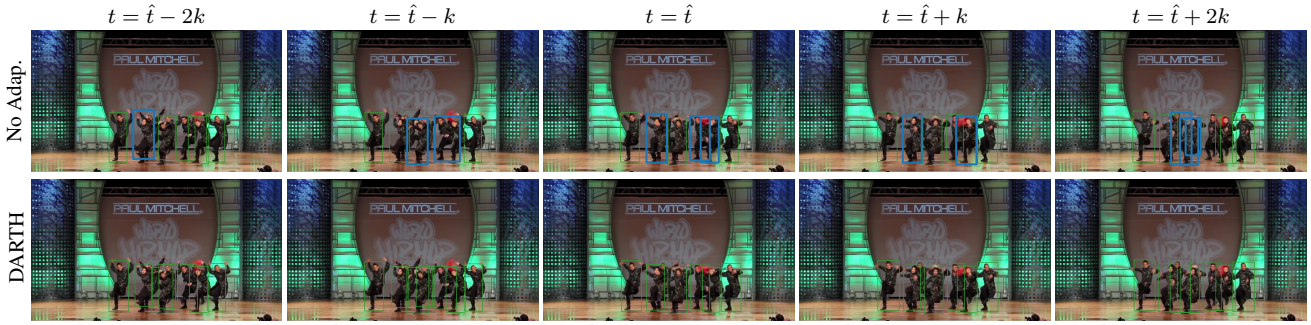


Figure 16. Tracking results on the sequence *0035* of the DanceTrack validation set in the adaptation setting MOT17 $\rightarrow$ DanceTrack. We analyze 5 consecutive frames centered around the frame #248 at time $\hat{t}$ and spaced by $k = 0.05$ seconds. We visualize the No Adap. baseline (top row) and DARTH (bottom row). On each row, green boxes represent correctly tracked objects, and blue boxes represent ID switches. We omit false positive and false negative boxes for ease of visualization.

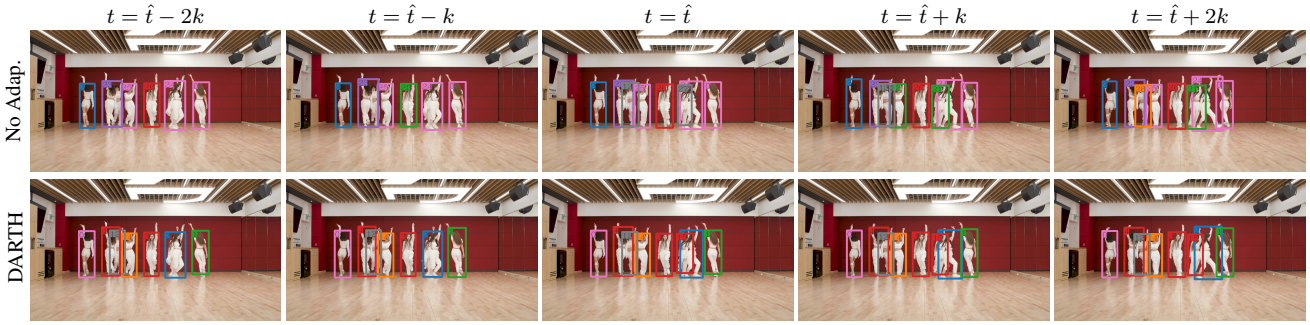


Figure 17. Tracking results on the sequence *0007* of the DanceTrack validation set in the adaptation setting MOT17 $\rightarrow$ DanceTrack. We analyze 5 consecutive frames centered around the frame #143 at time $\hat{t}$ and spaced by $k = 0.05$ seconds. We visualize the No Adap. baseline (top row) and DARTH (bottom row). On each row, boxes of the same color correspond to the same tracking ID.


Figure 18. Tracking results on the sequence *0007* of the DanceTrack validation set in the adaptation setting MOT17 $\rightarrow$ DanceTrack. We analyze 5 consecutive frames centered around the frame #143 at time $\hat{t}$ and spaced by $k = 0.05$ seconds. We visualize the No Adap. baseline (top row) and DARTH (bottom row). On each row, green boxes represent correctly tracked objects, and blue boxes represent ID switches. We omit false positive and false negative boxes for ease of visualization.

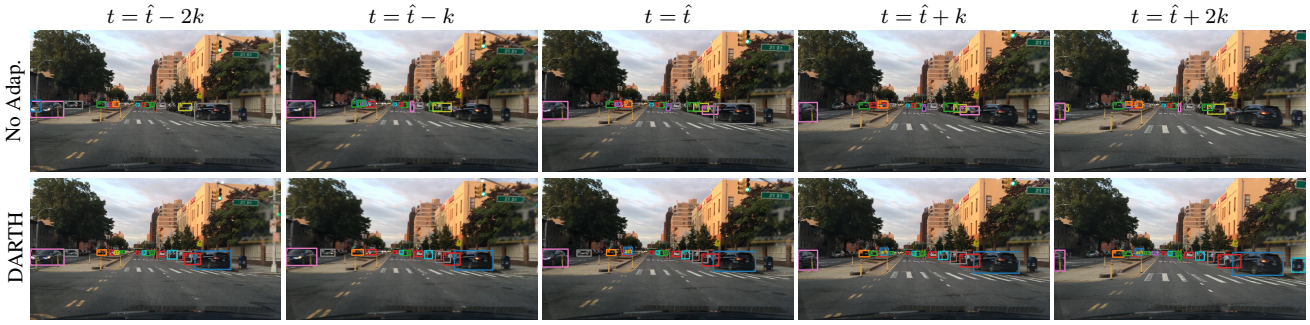


Figure 19. Tracking results on the sequence *b1c66a42-6f7d68ca* of the BDD100K validation set in the adaptation setting SHIFT $\rightarrow$ BDD100K. We analyze 5 consecutive frames centered around the frame #7 at time $\hat{t}$ and spaced by $k{=}0.2$ seconds. We visualize the No Adap. baseline (top row) and DARTH (bottom row). On each row, boxes of the same color correspond to the same tracking ID.


Figure 20. Tracking results on the sequence *b1c66a42-6f7d68ca* of the BDD100K validation set in the adaptation setting SHIFT $\rightarrow$ BDD100K. We analyze 5 consecutive frames centered around the frame #7 at time $\hat{t}$ and spaced by $k{=}0.2$ seconds. We visualize the No Adap. baseline (top row) and DARTH (bottom row). On each row, green boxes represent correctly tracked objects, and orange boxes represent false negatives. We omit false positive boxes and ID switches for ease of visualization.

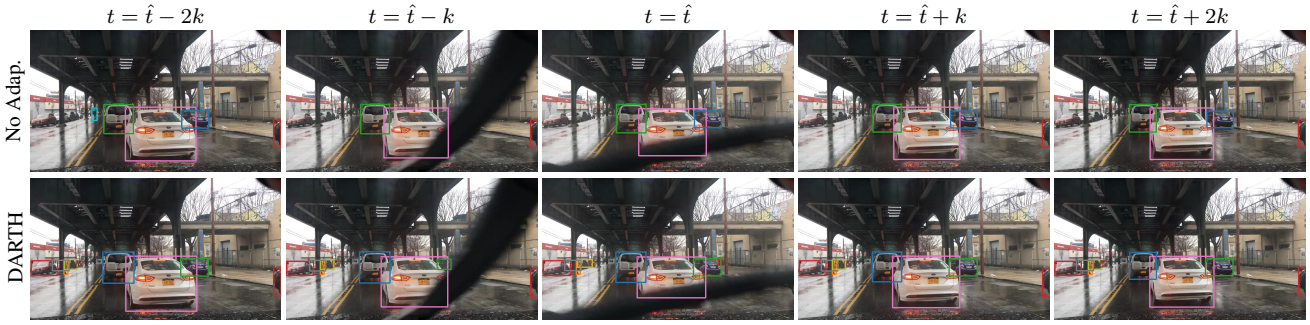


Figure 21. Tracking results on the sequence *b1cac6a7-04e33135* of the BDD100K validation set in the adaptation setting SHIFT $\rightarrow$ BDD100K. We analyze 5 consecutive frames centered around the frame #44 at time $\hat{t}$ and spaced by $k{=}0.2$ seconds. We visualize the No Adap. baseline (top row) and DARTH (bottom row). On each row, boxes of the same color correspond to the same tracking ID.

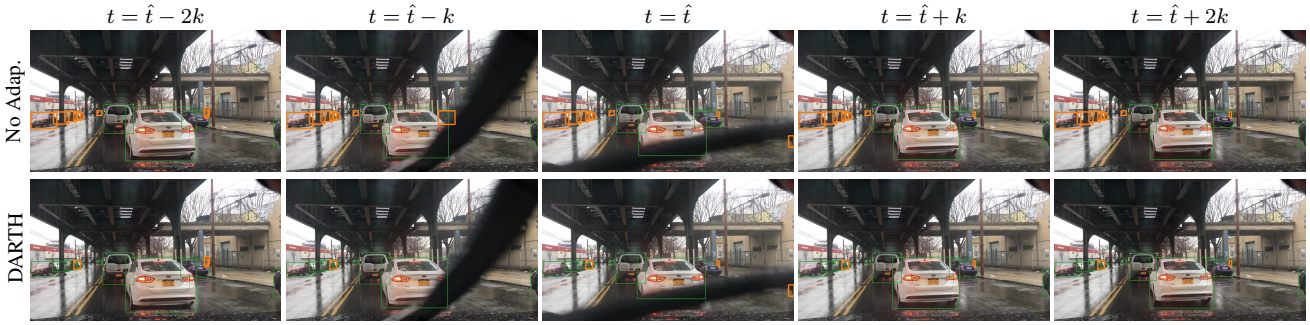


Figure 22. Tracking results on the sequence *b1cac6a7-04e33135* of the BDD100K validation set in the adaptation setting SHIFT $\rightarrow$ BDD100K. We analyze 5 consecutive frames centered around the frame #44 at time $\hat{t}$ and spaced by $k{=}0.2$ seconds. We visualize the No Adap. baseline (top row) and DARTH (bottom row). On each row, green boxes represent correctly tracked objects, and orange boxes represent false negatives. We omit false positive boxes and ID switches for ease of visualization.


Figure 23. Tracking results on the sequence *b250fb0c-01a1b8d3* of the BDD100K validation set in the adaptation setting SHIFT $\rightarrow$ BDD100K. We analyze 5 consecutive frames centered around the frame #114 at time $\hat{t}$ and spaced by $k{=}0.2$ seconds. We visualize the No Adap. baseline (top row) and DARTH (bottom row). On each row, boxes of the same color correspond to the same tracking ID.


Figure 24. Tracking results on the sequence *b250fb0c-01a1b8d3* of the BDD100K validation set in the adaptation setting SHIFT $\rightarrow$ BDD100K. We analyze 5 consecutive frames centered around the frame #114 at time $\hat{t}$ and spaced by $k{=}0.2$ seconds. We visualize the No Adap. baseline (top row) and DARTH (bottom row). On each row, green boxes represent correctly tracked objects, and orange boxes represent false negatives. We omit false positive boxes and ID switches for ease of visualization.

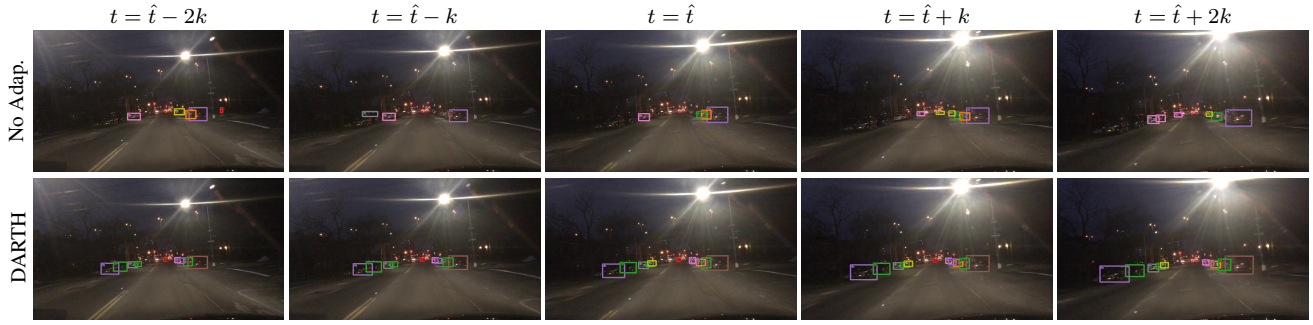


Figure 25. Tracking results on the sequence *b2064e61-2beadd45* of the BDD100K validation set in the adaptation setting SHIFT $\rightarrow$ BDD100K. We analyze 5 consecutive frames centered around the frame #100 at time $\hat{t}$ and spaced by $k{=}0.2$ seconds. We visualize the No Adap. baseline (top row) and DARTH (bottom row). On each row, boxes of the same color correspond to the same tracking ID.

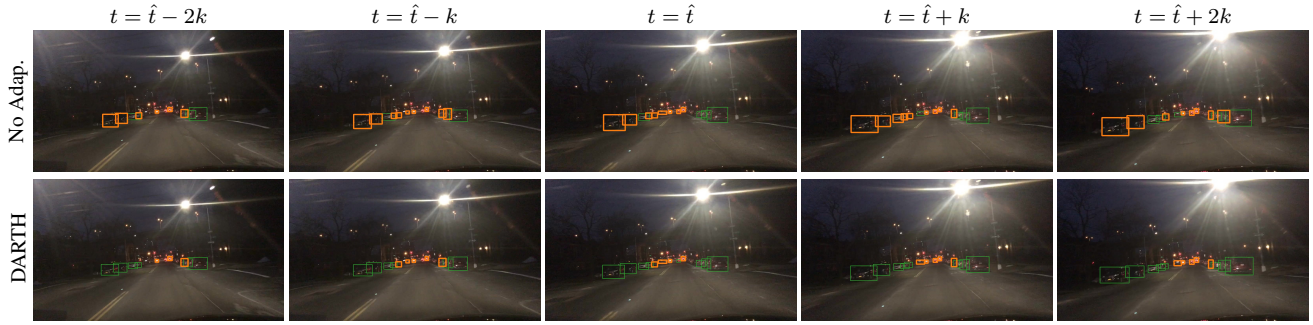


Figure 26. Tracking results on the sequence *b2064e61-2beadd45* of the BDD100K validation set in the adaptation setting SHIFT $\rightarrow$ BDD100K. We analyze 5 consecutive frames centered around the frame #100 at time $\hat{t}$ and spaced by $k{=}0.2$ seconds. We visualize the No Adap. baseline (top row) and DARTH (bottom row). On each row, green boxes represent correctly tracked objects, and orange boxes represent false negatives. We omit false positive boxes and ID switches for ease of visualization.

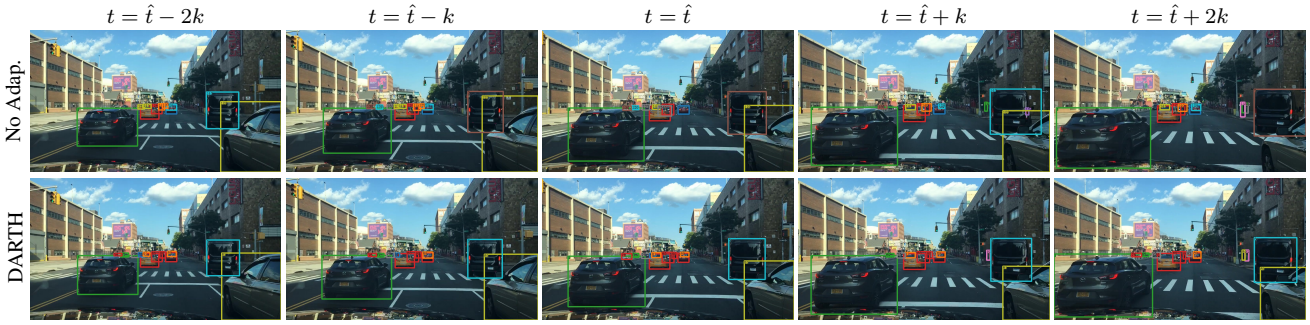


Figure 27. Tracking results on the sequence *b23493b1-3200de1c* of the BDD100K validation set in the adaptation setting SHIFT $\rightarrow$ BDD100K. We analyze 5 consecutive frames centered around the frame #99 at time $\hat{t}$ and spaced by $k{=}0.2$ seconds. We visualize the No Adap. baseline (top row) and DARTH (bottom row). On each row, boxes of the same color correspond to the same tracking ID.


Figure 28. Tracking results on the sequence *b23493b1-3200de1c* of the BDD100K validation set in the adaptation setting SHIFT $\rightarrow$ BDD100K. We analyze 5 consecutive frames centered around the frame #99 at time $\hat{t}$ and spaced by $k{=}0.2$ seconds. We visualize the No Adap. baseline (top row) and DARTH (bottom row). On each row, green boxes represent correctly tracked objects, and blue boxes represent ID switches. We omit false positive and false negative boxes for ease of visualization.

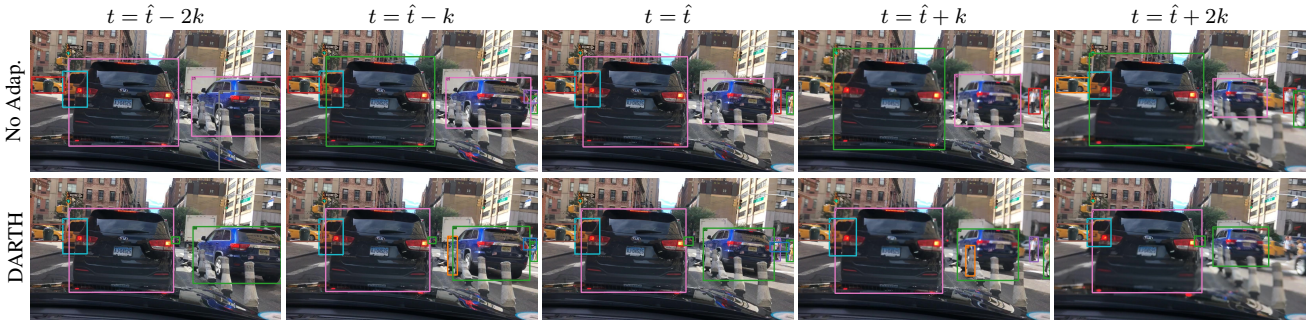


Figure 29. Tracking results on the sequence *b1f4491b-97465266* of the BDD100K validation set in the adaptation setting SHIFT $\rightarrow$ BDD100K. We analyze 5 consecutive frames centered around the frame #32 at time $\hat{t}$ and spaced by $k{=}0.2$ seconds. We visualize the No Adap. baseline (top row) and DARTH (bottom row). On each row, boxes of the same color correspond to the same tracking ID.


Figure 30. Tracking results on the sequence *b1f4491b-97465266* of the BDD100K validation set in the adaptation setting SHIFT $\rightarrow$ BDD100K. We analyze 5 consecutive frames centered around the frame #32 at time $\hat{t}$ and spaced by $k{=}0.2$ seconds. We visualize the No Adap. baseline (top row) and DARTH (bottom row). On each row, green boxes represent correctly tracked objects, and blue boxes represent ID switches. We omit false positive and false negative boxes for ease of visualization.

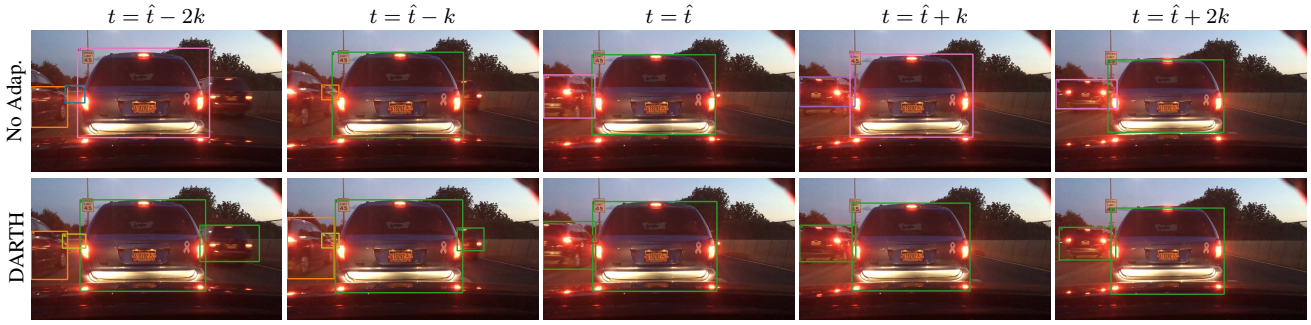


Figure 31. Tracking results on the sequence *b1e8ad72-c3c79240* of the BDD100K validation set in the adaptation setting SHIFT $\rightarrow$ BDD100K. We analyze 5 consecutive frames centered around the frame #107 at time $\hat{t}$ and spaced by $k = 0.2$ seconds. We visualize the No Adap. baseline (top row) and DARTH (bottom row). On each row, boxes of the same color correspond to the same tracking ID.

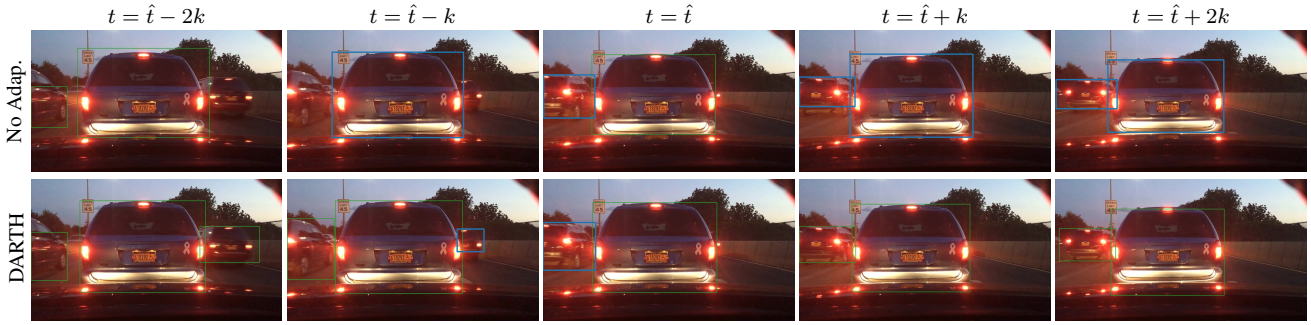


Figure 32. Tracking results on the sequence *b1e8ad72-c3c79240* of the BDD100K validation set in the adaptation setting SHIFT $\rightarrow$ BDD100K. We analyze 5 consecutive frames centered around the frame #107 at time $\hat{t}$ and spaced by $k = 0.2$ seconds. We visualize the No Adap. baseline (top row) and DARTH (bottom row). On each row, green boxes represent correctly tracked objects, and blue boxes represent ID switches. We omit false positive and false negative boxes for ease of visualization.