DetectSCI: Toward Object-Guided ROI Reconstruction for High-Resolution Video Snapshot Compressive Imaging

Supplementary Material

In our supplementary material, we include more experiment results, visualization analysis to further support the motivations and contributions of DetectSCI. The supplementary material is organized as follows:

- Sec. 1 presents expanded preliminaries, including architectural details of CNN-based and Transformer-based detectors, and an extended explanation of the SS2D mechanism that employed for our linear-complexity global modeling.
- Sec. 2 provides deeper feature-degradation analysis under SCI conditions through visualize the extracted features of measurements and original frames across three sports scenarios.
- Sec. 3 reports comprehensive dataset statistics and full training configurations of our model.
- Sec. 4 and Sec. 5 extends the comparative study of CNN-based detectors, and provides qualitative visualizations of detection results to highlight failure modes of existing detectors and the superior performance of our detector under SCI conditions.

1. Preliminaries

CNN-based architecture. CNN-based one-stage object detectors *i.e*. YOLOs have a unified paradigm that directly predicts object categories and bounding boxes from feature representations. They first use an efficient backbone to extract multi-scale features which are then processed by an aggregation module to propagate information across different scales. Finally, detection heads perform joint classification and bounding-box regression at multiple feature levels. Although YOLOv12 [26] incorporates attention mechanisms within its aggregation module, the overall detector remains fundamentally convolutional-based.

Transformer-based architecture. Transformer-based detectors, *i.e*. DETRs, generally follow a similar architectural design. First, an input image is passed through a CNN backbone to extract multi-scale features that are flattened into feature tokens and concatenated as input to the Transformer encoder. The encoder progressively transforms these tokens from low-level features into high-level semantic representations. Second, a fixed set of learnable object queries $\mathbf{Q} = \{q_0, q_1, \ldots, q_n\}$ is initialized and fed into the decoder, where they interact with encoder features through cross-attention to produce refined queries that correspond to potential objects. Finally, separate classification and box regression heads operate on the refined queries to generate the final detection results including categories $\mathbf{C} = \{c_0, c_1, \ldots, c_n\}$ and bounding boxes $\mathbf{B} = \{b_0, b_1, \ldots, b_n\}$.

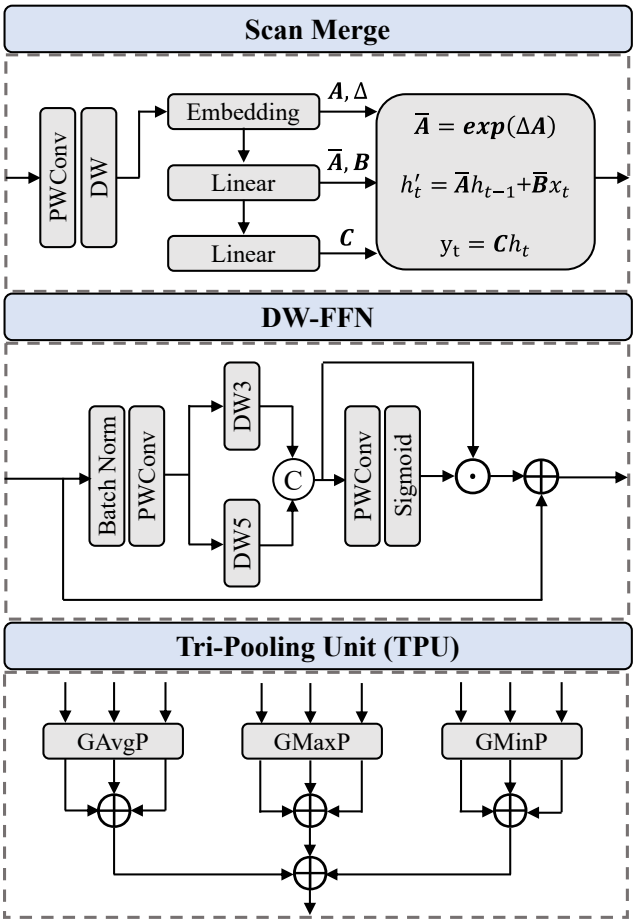

Figure 1. Detailed illustration of the architectural components used in the DetectSCI detector. This figure visualizes the internal structures of the modules within the Mamba-Implicit Module (MIM) and Frequency Mamba (FM), including the Scan Merge block in SS2D, the Depth-wise Feed-Forward Network (DW-FFN), and the Tri-Pooling Unit (TPU) used in MFCA. These diagrams complement Figure 2 by showing how selective scan sequences are merged and propagated, and how frequency pooling is performed at the module level.

2D Selective Scan (SS2D). State Space Models (SSMs) have recently re-emerged as a compelling alternative to self-attention because they model global relationship with linear complexity. But original Mamba is inherently one-dimensional since its selective scan assumes a temporal order that is incompatible with the 2D images that are non-sequential and contain spatial information[17]. To close this gap, VMamba[17] introduces an input-independent 2D Se-

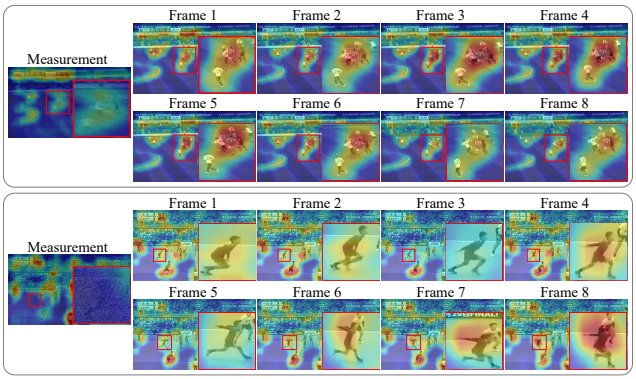

Figure 2. Illustration of feature degradation under snapshot compressive imaging (SCI). Zoom in for better view. Each row compares the heatmaps of features extracted from the SCI measurement (left) with those extracted from the corresponding original video frames (right) across three representative sports scenes: basketball, football, and volleyball. The measurements exhibit pronounced degradation due to temporal mixing, especially in regions involving object overlap and high-speed motion. This visualization highlights the inherent information loss introduced by SCI and motivates the core problem addressed in DetectSCI: **compression produces frequency-biased degraded visual features that fundamentally limit the performance of existing detection models.**

lective Scan (SS2D) mechanism to restructure a 2D feature map into four token sequence generated from complementary traversal scanning paths. Each path generates a 1D sequence processed by an S6 block, and the four directional states are subsequently fused via a scan merge block to form the output map. This multi-directional traversal ensures that every pixel aggregates contextual evidence from all scan directions, which contributing to yielding 2D global receptive fields. In this paper, we employ the proposed SS2D[17] module to replace the self-attention for global modeling with linear complexity, and integrate it with channel attention to enhance feature discriminability.

2. Feature Analysis

To further substantiate the feature-degradation analysis discussed in Sec. 1 and Sec. 3 of the main paper, we provide an extended examination of how snapshot compressive imaging (SCI) alters visual perception, with visual evidence summarized in Figure 2. Comparing features extracted from SCI measurements with those extracted from the corresponding original video frames, we finds a consistent degradation pattern across basketball, football, and volleyball scenes. Temporal modulation weakens the contrast between objects and background, but more critically, the degradation is highly biased: regions containing high-frequency structures, such as fast-moving limbs, overlapping players, and large frame-to-frame displacement, suffer disproportionately. In these areas, object boundaries merge into the background or another boundary due to compression, which makes object-level feature difficult to separate. It can be seen from the results that motion-intensive regions consistently lose fine contours and object-level details, whereas low-motion regions remain relatively intact. These observations directly reinforce the core motivation behind DetectSCI: **SCI introduces a temporal-spatial coupled, frequency-biased feature degradation that fundamentally limits the performance of conventional detectors**. The proposed Mamba-Implicit Module (MIM) and Frequency Mamba (FM) are thus not merely architectural enhancements but targeted responses to these degradation modes. MIM compensates for spatial discontinuities by multi-scale receptive fields and continuous coordinate cues, while FM restores discriminative frequency information suppressed during coded exposure. Together, these modules form our solution to bridge the perception gap between SCI measurements and their corresponding original frames.

Table 1. Detailed statistics for different dataset splits. Meas denotes SCI measurements.

Split	Basketball	Football	Volleyball	Bboxes/Meas
Train	17287	13829	7298	11
Valid	4105	2965	1164	11
Test	3304	3165	1755	11
Total	24696	19959	10217	33

3. Dataset and Training Details

Dataset. Table 1 summarizes the statistics of our SCI-oriented dataset constructed from SportsMOT [6], which is used in all DetectSCI experiments. The dataset covers three representative sports scenes (basketball, football, and volleyball) and reflects the diverse motion patterns, object densities, and occlusion behaviors commonly observed in high-speed sports scenarios under SCI conditions. As shown in the table, most samples are allocated to the training split to ensure sufficient diversity for learning representations under coded exposure, while the validation and test splits maintain similar distributions in both category composition and average object count ($\approx$11 objects per measurement). Overall, the dataset contains 24,696 basketball, 19,959 football, and 10,217 volleyball instances, forming a balanced yet challenging benchmark with heterogeneous motion dynamics and dense multi-object interactions.

Training Details. We train our detector following the training strategy and hyperparameters settings of RT-DETR[35]. The data augmentation applied during training is the same as [35], including random flipping, resize, expand and crop. We have the optimizer, base learning rate and learning rate of backbone in Sec. 4, the other main hyperparameters are listed in Table 3. Since players are the main class of our dataset, we appropriately increase the bbox loss weight and reduce class loss to emphasize more on the localization.

4. Comparison with More YOLO Detectors

We have conducted experiments to compare our detectors with CNN-based and Transformer-based model in Sec.4. Since the main text has space limitations, we present the experiments of lighter YOLOs in Tab. 2 to make more comprehensive comparisons based on results in Sec. 4. These experiments show that our approach yields consistent and substantial gains across different lightweight YOLOs.

Table 2. Comparison with CNN-based object detectors that are not listed in Tab. 1 on the test set.

Model	AP^{test}	AP^{test}_{50}	AP^{test}_{75}
YOLOv8-S [12]	67.7	91.8	76.1
YOLOv8-M [12]	70.4	92.6	78.7
YOLOv8-L [12]	75.2	92.6	83.7
YOLOv9-S [28]	67.4	91.0	75.4
YOLOv9-M [28]	70.5	92.5	78.8
YOLOv9-c [28]	74.5	92.5	82.4
YOLOv10-S [27]	68.5	91.9	76.9
YOLOv10-M [27]	69.9	92.3	78.1
YOLOv10-L [27]	76.3	92.8	85.4
YOLOv11-S [11]	69.3	93.0	33.9
YOLOv11-M [11]	69.8	92.3	77.5
YOLOv11-L [11]	76.2	92.5	85.6
YOLOv12-S [26]	67.0	92.2	74.5
YOLOv12-M [26]	70.3	92.9	80.7
YOLOv12-L [26]	75.1	92.9	83.5
Ours	**80.9**	**98.5**	**93.1**

Table 3. Training hyper-parameters.

Item	**Value**
batch size	16
num workers	8
freezing BN	False
linear warm-up start factor	0.001
linear warm-up steps	3000
weight decay	5e-5
ema decay	0.9999
embedding dim	256
feedforward dim	1024
number of feature scales	3
number of decoder layers	6
number of queries	300
decoder npoints	4
clip gradient norm	0.1
class cost weight	2.0
bbox cost weight	5.0
GIoU cost weight	2.0
α in class cost	0.25
γ in class cost	2.0
class loss weight	1.0
bbox loss weight	8.0
GIoU loss weight	1.0
α in class loss	0.75
γ in class loss	2.0
denoising number	100
label noise ratio	0.5
box noise scale	1.0

5. Qualitative visualization of Predictions

We provide visualized detection results for all compared models in Figure 3. It can be seen from the visualization results that missed detections consistently cluster in motion-intensive regions, particularly where players exhibit rapid limb movement, large frame-to-frame displacement, or severe spatial overlap. These areas correspond exactly to the heavily degraded regions in Figure 2, where SCI's coded exposure suppresses high-frequency structures such as edges and fine boundaries. In contrast, low-motion body parts and static backgrounds preserve spatial continuity, leading most CNN- and Transformer-based detectors to perform reliably only on these easy regions. For example, In the basketball scene, nearly all detectors collapse two heavily overlapping players in the bottom region into a single prediction. In the football scene *i.e*. the first comparison, the central cluster of five tightly overlapping players represents the most severe detection-missing region where even strong baselines including YOLO-v12x[26], DINO[33], RT-DETR[35] fail to detect all individual instances. In the volleyball scene *i.e*. the second comparison, fast jumping motion causes multiple detectors to entirely miss the leftmost athlete. These results collectively demonstrate that existing detectors are highly vulnerable to SCI-induced high-frequency loss that leads missed detection localized precisely in regions with strong temporal dynamics. In stark contrast, our method successfully separates players that appear visually merged in the compressed measurements which qualitatively indicates a superior performance to address feature degradation under SCI conditions.

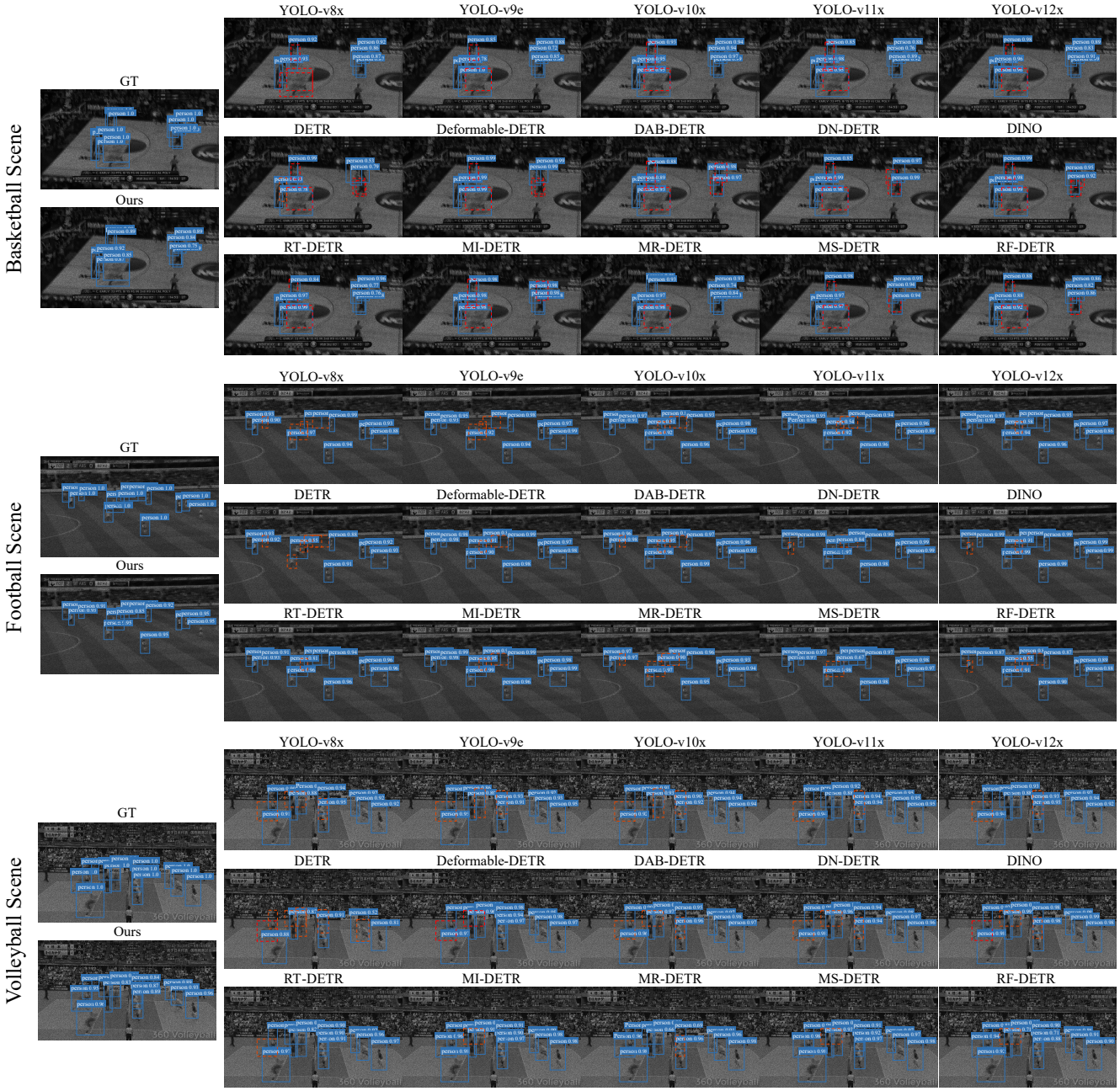

Figure 3. Qualitative comparison of detection results across basketball, football, and volleyball scenes. For each scene, predictions are presented in a consistent left-to-right layout: **Ground Truth**, **Ours**, the best-performing **CNN-based detectors** of each version, and a set of **Transformer-based detectors** (Deformable-DETR, DAB-DETR, DN-DETR, *etc.*). **Red dashed boxes denote missed detections**. In complex sports scenes, our detector demonstrates its capability to detect objects with diverse motion patterns. Notably, our detector shows a stronger robustness to motion overlap and compression-induced ambiguity since it successfully separates players that appear visually merged in the compressed measurements.