Supplementary Material
Real-Time Tracking of Flexible Markers in Low-Contrast Fluoroscopy Using a Deep Neural Network Trained Solely on Synthetic Data

Tomoki Uchiyama[1], Yukinobu Sakata[1], Ryusuke Hirai[1], Hitoshi Ishikawa[2], Shinichiro Mori[2]
[1]Toshiba Corporation, Japan
{tomoki.uchiyama.x22, yukinobu.sakata.r73, ryusuke.hirai.r37}@mail.toshiba
[2]National Institutes for Quantum Science and Technology, Japan
{ishikawa.hitoshi, mori.shinichiro}@qst.go.jp

Table 1. Detailed architecture of the shared backbone.

Block	Output Size	Operation	Configuration
Conv1	$32 \times 31 \times 31$	Conv2D	5×5, stride 2, pad 0
		BN + ReLU	–
Conv2	$32 \times 27 \times 27$	Conv2D	5×5, stride 1, pad 0
		BN + ReLU	–
Conv3	$64 \times 12 \times 12$	Conv2D	5×5, stride 2, pad 0
		BN + ReLU	–
Conv4	$64 \times 10 \times 10$	Conv2D	3×3, stride 1, pad 0
		BN + ReLU	–
Conv5	$64 \times 8 \times 8$	Conv2D	3×3, stride 1, pad 0
		BN + ReLU	–

A. Network Architecture Details

This section provides the detailed architecture of the proposed Siamese CNN. The network features a lightweight design consisting of a shared backbone and two prediction heads, optimized for real-time tracking applications.

A.1. Shared Backbone Architecture

The shared backbone takes a $1 \times 65 \times 65$ image as input and consists of five convolutional blocks. Each block is composed of a convolutional (Conv) layer, a batch normalization (BN) layer [2], and a ReLU activation, applied in sequence. The detailed configuration of the shared backbone is shown in Tab. 1.

A.2. Similarity Head Architecture

The similarity head takes $64 \times 8 \times 8$ feature maps from the template and candidate images as input and evaluates the similarity between them. It begins by computing the element-wise (Hadamard) product of the two feature maps to create a correlation map. The detailed configuration is shown in Tab. 2.

Table 2. Detailed architecture of the similarity head.

Block	Output Size	Operation	Configuration
Corr	$64 \times 8 \times 8$	Hadamard ($\odot$)	–
Conv1	$64 \times 8 \times 8$	Conv2D	3×3, stride 1, pad 1
		BN + ReLU	–
Flatten	4096	Flatten	–
FC1	512	Fully Connected	–
		ReLU	–
FC2	2	Fully Connected	–
		Softmax	–

Table 3. Detailed architecture of the detection head.

Block	Output Size	Operation	Configuration
Conv1	$64 \times 8 \times 8$	Conv2D	3×3, stride 1, pad 1
		BN + ReLU	–
Flatten	4096	Flatten	–
FC1	512	Fully Connected	–
		ReLU	–
FC2	2	Fully Connected	–
		Softmax	–

A.3. Detection Head Architecture

The detection head takes the feature map of only the candidate image as input and outputs the probability of the marker presence at the image center. This architecture is detailed in Tab. 3.

B. Synthetic Data Examples

This section presents examples of synthetic images demonstrating the effects of different data generation parameters. These examples visually illustrate the importance of diversity and quality control in our training data generation pipeline.

B.1. Effect of Marker Contrast Factor (a)

The marker contrast factor a controls the contrast of the marker against its local background. Fig. 1 shows synthetic image examples with varying values of a.

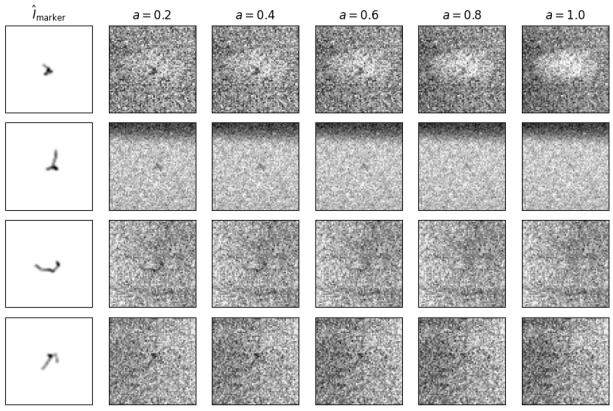


Figure 1. Effect of marker contrast factor a for fixed noise factor b = 1.5. As a increases (from left to right), the contrast between the marker and the background decreases, reducing its visibility.

B.2. Effect of Noise Factor (b)

The noise factor b controls the intensity of the Gaussian noise added to simulate the X-ray imaging process. Fig. 2 shows synthetic image examples with varying values of b.

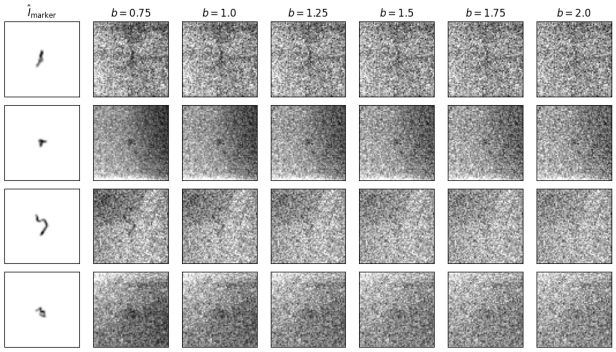


Figure 2. Effect of noise factor b for fixed marker contrast factor a = 0.8. As b increases (from left to right), the image noise increases, making the marker more difficult to detect.

C. Clinical Data Examples

This section presents examples of clinical images for the pancreatic cancer cases, categorized by visibility level. Marker visibility was classified into the following three levels:

- **High Visibility:** The marker is clearly identifiable.
- **Medium Visibility:** The marker presence can be confirmed, but it is difficult to identify in some frames.
- **Low Visibility:** The marker is barely visible and requires careful observation, even by experts.

Examples of fluoroscopic images for each visibility level are shown in Fig. 3.

Table 4. Hyperparameters and settings for network training.

Parameter	Value
Training Images	500,000
Validation Images	50,000
Batch Size	2,048
Epochs	50
Optimizer	Adam [3]
Learning Rate	1e-3
Weight Decay	1e-3
Data Augmentation	Random vertical flip
	Random horizontal flip
	Random brightness
	Random contrast
	Random gamma correction

D. Implementation Details

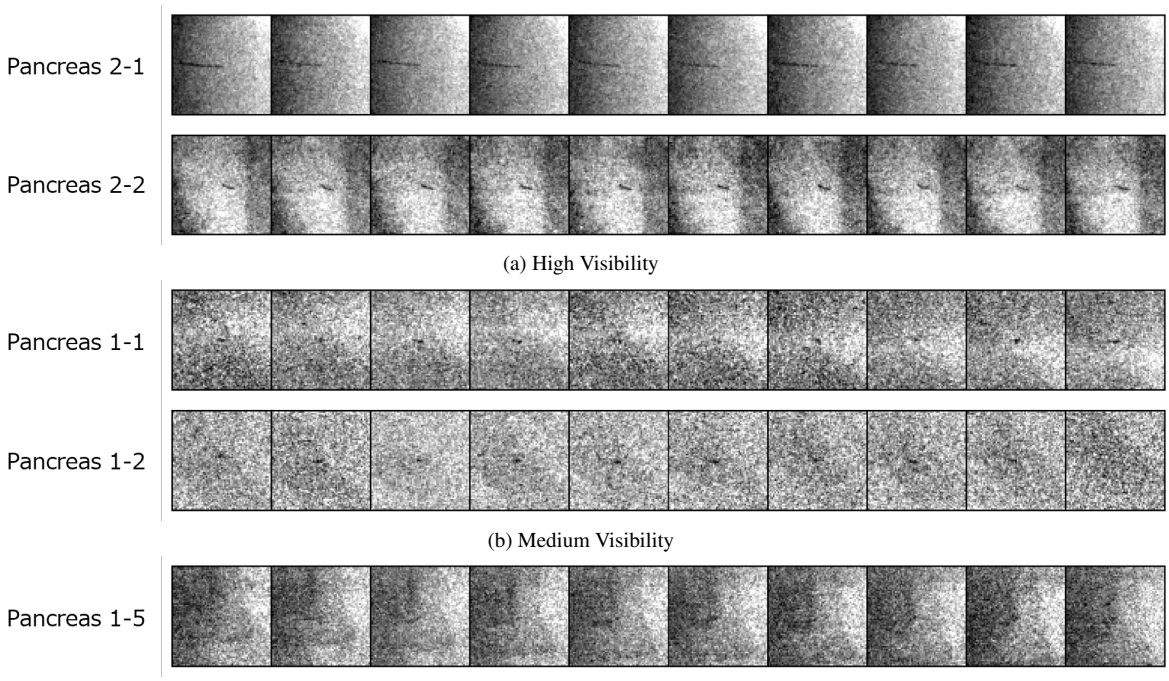
This section provides the implementation settings for both the network training phase and the tracking phase. Sec. D.1 describes the training configuration of our Siamese CNN on synthetic data. Sec. D.2 details the parameters used by the tracking algorithm on clinical images. Sec. D.3 provides the details of the YOLOX-tiny baseline used for comparison, including the construction of its synthetic training set and the training hyperparameters.

D.1. Network Training

The detailed settings for network training are shown in Tab. 4. The training data consisted of 500,000 synthetic images, with an additional 50,000 images prepared separately for validation. These images were generated to reproduce diverse visibility conditions, as described in the main paper. The training process also incorporated standard data augmentation techniques, such as random flips and adjustments to brightness and contrast.

D.2. Tracking Algorithm Parameters

The algorithm parameters used during tracking are shown in Tab. 5. Parameters such as the initial ROI size and the number of frames T_{init} for initial detection were adjusted based on the characteristics of each dataset. For the pancreatic cancer dataset, which has low visibility and larger motion, a larger ROI and a greater number of frames were used to ensure robust detection.



(a) High Visibility

(b) Medium Visibility

(c) Low Visibility

Figure 3. Clinical fluoroscopy images from pancreatic cancer cases, categorized by visibility level. Each image is cropped around the GT position. (a) **High Visibility:** The marker is clearly identifiable. (b) **Medium Visibility:** Marker presence can be confirmed, but it is difficult to identify in some frames. (c) **Low Visibility:** The marker is barely visible and requires careful observation, even by experts.

Table 5. Tracking algorithm parameters.

Parameter	Prostate	Pancreas
Initial Detection		
ROI Size [pixels]	20×20	100×100
Detection Frames T_{init}	5	50
Grid Spacing δ_{grid} [pixels]	2.0	2.0
DP Weight λ_{DP}	0.5	0.2
Detection Threshold τ	0.8	0.8
Particle Filter		
Particle Number N_p	2,000	2,000
Noise Std. Dev. σ_p [pixels]	3.0	3.0

D.3. Details of the YOLOX Baseline

We used YOLOX-tiny [1] as a learning-based baseline. To ensure a fair comparison, we trained it on a synthetic dataset created with the same pipeline as our proposed method. Specifically, we generated 100,000 images for training, where 0 to 4 markers were randomly placed on 160×160 pixel background patches, as shown in Fig. 4. These patches were then resized to 416×416 pixels to match the standard input size for YOLOX-tiny. We fine-tuned the learning rate from {1e-2, 1e-3, 1e-4} and the number of epochs from {30, 50, 100}, while other settings like the optimizer and data augmentation followed the standard YOLOX configuration.

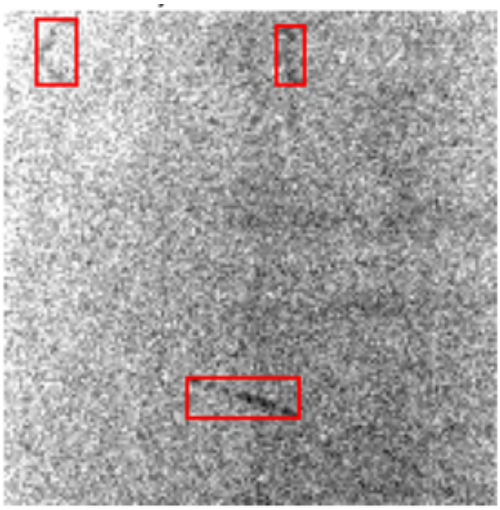


Figure 4. A synthetic image used for training the YOLOX baseline. In this example, three markers are randomly placed on a 160×160 pixel background patch.

E. Ablation Study of Siamese Network Components

To validate the design of the likelihood function within the particle filter, which leverages the outputs of our

Siamese network, we conducted an ablation study. The proposed method uses the product of the similarity score (S_{sim}) from the similarity head and the detection score (S_{det}) from the detection head. We compared this design with using only S_{sim} or only S_{det}.

The results on the pancreas dataset are summarized in Tab. 6. Using only S_{det} results in the largest error, as it does not utilize the template information from the initial frame. While using only S_{sim} significantly improves accuracy, it occasionally caused tracking to drift toward background textures resembling the template. The proposed method, which uses the product of both scores, achieves the best performance. This confirms that our combined likelihood design provides the most robust and stable performance by preventing drift.

Table 6. Comparison of likelihood designs for the particle filter on the pancreas dataset. We report the mean tracking error for the proposed method ($S_{\mathrm{sim}} \cdot S_{\mathrm{det}}$) and for variants using each score alone.

Likelihood Function	Mean Tracking Error [pixels]
S_{det} only	1.60 ± 1.10
S_{sim} only	1.05 ± 0.62
$S_{\mathrm{sim}} \cdot S_{\mathrm{det}}$ **(Proposed)**	**0.97 ± 0.53**

Ethical Statement

This study involved human participants. It was approved by our institution's Institutional Review Board and performed in accordance with the Declaration of Helsinki.

References

[1] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 3

[2] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015. 1

[3] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 2