Appendices

A Algorithm 1

Algorithm 1 CD-CNN online tracking algorithm

Input:

PVANET pool5 feature extractor;
pre-trained learning weights $W = \{W_1, ..., W_5\}$;
initial target patch $P_*^{(1)}$ with its bounding box $B_*^{(1)}$;

Output:

Predicted target patches $\{\hat{P}_*^{(t)}\}_{t=2}^{T}$ with bounding boxes $\{\hat{B}_*^{(t)}\}_{t=2}^{T}$;

Draw positive samples $\{P_j^{(1)}\}$, $\{P_k^{(1)}\}$ and negative samples $\{N_l^{(1)}\}$;
Update $\{W_1, ..., W_5\}$ using $\{P_j^{(1)}\}$, $\{P_k^{(1)}\}$, $\{N_l^{(1)}\}$;
$t = 2$;
repeat
  Draw m candidates $\{P_j^{(t)}\}_{j=1}^{m}$, with corresponding bounding boxes $\{B_j^{(t)}\}$;
  Forward m candidates through CD-CNN;
  Select the top five candidates $\{P_{(1)}^{(t)}, ..., P_{(5)}^{(t)}\}$ with highest object-centroid scores $p(\cdot)$;
  Determine $\hat{P}_*^{(t)}$ by averaging bounding boxes $\{B_{(1)}^{(t)}, ..., B_{(5)}^{(t)}\}$;
  if t mod $5 == 0$ and $p(\hat{P}_*^{(t)}) > 0.95$ **then**
    Draw positive samples $\{P_j^{(t)}\}$, $\{P_k^{(t)}\}$ and negative samples $\{N_l^{(t)}\}$;
    Update $\{W_1, ..., W_5\}$ using $\{P_j^{(t)}\}$, $\{P_k^{(t)}\}$, $\{N_l^{(t)}\}$;
  end if
  $t = t + 1$;
until end of sequence

B Proof of Theorem 1

The target appearance representation error incurred by our tracking method can be defined as the Euclidean distance $\mathcal{E} := \|\Phi(\hat{P}_*^{(t+1)}) - \Phi(P_*^{(t+1)})\|_2$, where $P_*^{(t)}$ is the ground-truth and $\hat{P}_*^{(t)}$ is the predicted target patch in the tth frame.

Theorem 1 (Upper Bound of Target Appearance Representation Error)**.** *With probability no less than* $1 - \rho$*, the target prediction error* $\mathcal{E}$ *is upper-bounded by* $\sum_j \sqrt{\hat{\mathcal{L}}_j^C}/m + n(\delta + K\Delta t)$*, for any* $\delta > \sqrt{\frac{n}{m} max_i \mathrm{Var}\,(\Phi_i)}$*, where* $\hat{\mathcal{L}}_j^C = \|\Phi(\hat{P}_*^{(t+1)}) - \Phi(P_j^{(t)})\|_2^2$ *is the estimated temporal appearance continuity loss for the predicted target in the* $(t+1)$*th frame with respect to* $P_j^{(t)}$ *and* $\rho = n \max_i \mathrm{Var}(\Phi_i)/m\delta^2$.

Proof. The target appearance representation error incurred at each time slot is given by

$$\begin{aligned}
\mathcal{E} &= \left\| \Phi\left(\hat{P}_*^{(t+1)}\right) - \Phi\left(P_*^{(t+1)}\right) \right\|_2 \\
&= \left\| \Phi\left(\hat{P}_*^{(t+1)}\right) - \Phi\left(P_*^{(t)}\right) + \Phi\left(P_*^{(t)}\right) - \Phi\left(P_*^{(t+1)}\right) \right\|_2 \\
&\leq \left\| \Phi\left(\hat{P}_*^{(t+1)}\right) - \Phi\left(P_*^{(t)}\right) \right\|_2 + \epsilon \\
&= \left\| \Phi\left(\hat{P}_*^{(t+1)}\right) - \frac{1}{m}\sum_{j=1}^{m} \Phi\left(P_j^{(t)}\right) + \frac{1}{m}\sum_{j=1}^{m} \Phi\left(P_j^{(t)}\right) - \Phi\left(P_*^{(t)}\right) \right\|_2 + \epsilon \\
&\leq \left\| \Phi\left(\hat{P}_*^{(t+1)}\right) - \frac{1}{m}\sum_{j=1}^{m} \Phi\left(P_j^{(t)}\right) \right\|_2 + \left\| \frac{1}{m}\sum_{j=1}^{m} \Phi\left(P_j^{(t)}\right) - \Phi\left(P_*^{(t)}\right) \right\|_2 + \epsilon \\
&= \left\| \Phi\left(\hat{P}_*^{(t+1)}\right) - \frac{1}{m}\sum_{j=1}^{m} \Phi\left(P_j^{(t)}\right) \right\|_2 + \left\| \overline{\Phi\left(P^{(t)}\right)} - \Phi\left(P_*^{(t)}\right) \right\|_2 + \epsilon \\
&\leq \left\| \Phi\left(\hat{P}_*^{(t+1)}\right) - \frac{1}{m}\sum_{j=1}^{m} \Phi\left(P_j^{(t)}\right) \right\|_2 + \left\| \overline{\Phi\left(P^{(t)}\right)} - \Phi\left(P_*^{(t)}\right) \right\|_1 + \epsilon \\
&= \left\| \Phi\left(\hat{P}_*^{(t+1)}\right) - \frac{1}{m}\sum_{j=1}^{m} \Phi\left(P_j^{(t)}\right) \right\|_2 + \sum_{i=1}^{n} \left| \overline{\Phi_i\left(P^{(t)}\right)} - \Phi_i\left(P_*^{(t)}\right) \right| + \epsilon
\end{aligned} \tag{1}$$

where $P_j^{(t)}$ is the jth positive samples drawn around $P_*^{(t)}$ and $\overline{\Phi_i\left(P^{(t)}\right)}$ denotes the arithmetic mean. Mathematically, it is assumed that, in the feature space χ, the random vector Φ, obeys some unknown distribution $\mathbb{P}(\varphi)$, whose expectation is given by

$$\mathbb{E}\left[\Phi\left(P_j^{(t)}\right)\right] = \int_{\chi} \varphi \, d\mathbb{P}(\varphi) = \Phi\left(P_*^{(t)}\right) \tag{2}$$

By Chebyshev inequality and Inclusion-exclusion Principle,

$$\begin{aligned}
\mathbb{P}\left(\bigcap_{i=1}^{n} \left| \overline{\Phi_i\left(P^{(t)}\right)} - \Phi_i\left(P_*^{(t)}\right) \right| < \delta \right) &\geq 1 - \sum_{i=1}^{n} \mathbb{P}\left(\left| \overline{\Phi_i\left(P^{(t)}\right)} - \Phi_i\left(P_*^{(t)}\right) \right| \geq \delta \right) \\
&\geq 1 - \frac{1}{m\delta^2} \sum_{i=1}^{n} \mathrm{Var}(\Phi_i) \\
&\geq 1 - \frac{n}{m\delta^2} \max_i \mathrm{Var}(\Phi_i)
\end{aligned} \tag{3}$$

Therefore, with the lower-bounded probability above, the target appearance representation error incurred at each time slot is upper-bounded by

$$\begin{aligned}
\mathcal{E} &\leq \left\| \Phi\left(\hat{P}_*^{(t+1)}\right) - \frac{1}{m}\sum_{j=1}^{m} \Phi\left(P_j^{(t)}\right) \right\|_2 + \sum_{i=1}^{n} \left| \overline{\Phi_i\left(P^{(t)}\right)} - \Phi_i\left(P_*^{(t)}\right) \right| + \epsilon \\
&\leq \frac{1}{m} \sum_{j} \left\| \Phi\left(\hat{P}_*^{(t+1)}\right) - \Phi\left(P_j^{(t)}\right) \right\|_2 + n\delta + \epsilon \\
&= \frac{1}{m} \sum_{j} \sqrt{\left\| \Phi\left(\hat{P}_*^{(t+1)}\right) - \Phi\left(P_j^{(t)}\right) \right\|_2^2} + n\delta + \epsilon \\
&\leq \frac{1}{m} \sum_{j} \left(\hat{\mathcal{L}}_j^C\right)^{\frac{1}{2}} + n\left(\delta + K\Delta t\right)
\end{aligned} \tag{4}$$

□

C Quantitive Comparison

C.1 OTB2015 Comparison

Figure 1 and Figure 2 show the overall performance comparison with state-of-the-art trackers and the internal comparison, respectively, on the OTB2015 dataset.

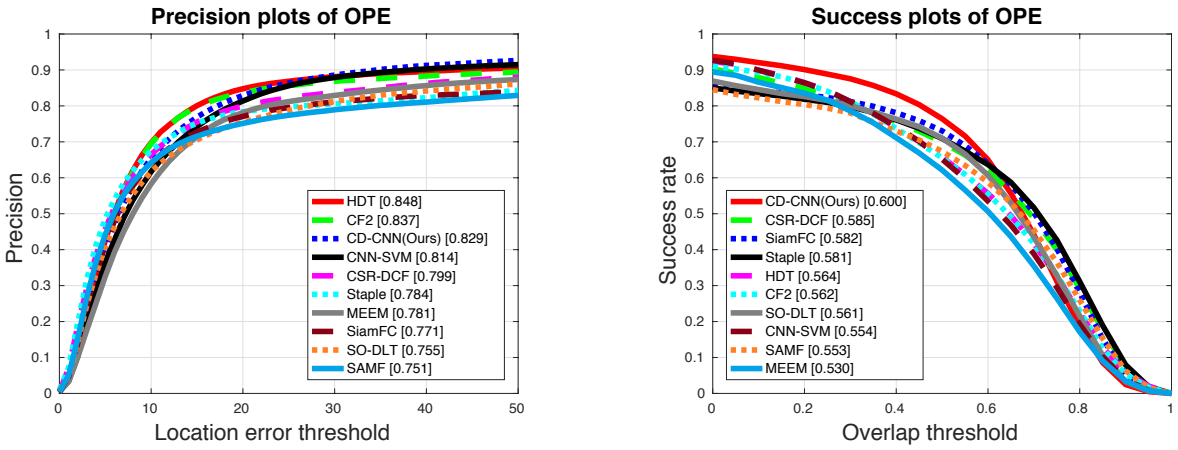

Figure 1: Precision plots and success plots for the overall performance comparison on OTB2015.

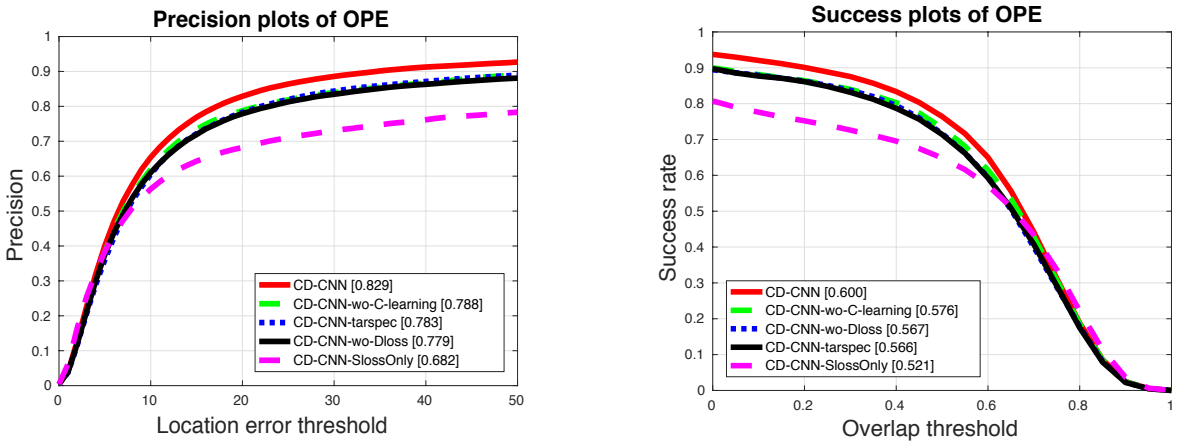

Figure 2: Precision plots and success plots of OPE for the internal comparisons on OTB2015.

C.2 OTB2013 Comparison

Figure 3 and Figure 4 show the overall performance comparison with state-of-the-art trackers and the internal comparison, respectively, on the OTB2013 dataset.

D Qualitative Comparison

Figure 5 illustrates the qualitative performance of our tracker on some challenging sequences, compared with state-of-the-art trackers including SINT, SiamFC, CF2, HDT and SO-DLT.

In the CarScale sequence, when the occlusion occurs between #157 and #180, our tracker can still track the car. In #239 of CarScale, only CD-CNN can successfully track the target with the largest IoU.

In #2940 of the Doll sequence, SO-DLT yields drifting to the man's face. This might be due to its inaccurate inverse mapping. In #3725, only CD-CNN can successfully handle the rapid scale variation, which benefits from its sensitivity to objectness and the relative position of the target in a patch, that is, object-centroid. In #3769, CD-CNN outperforms others by successfully tracking the blurred target, as a result of its temporal appearance continuity transferring.

In #562 and #574 of the Woman sequence, when the camera zooms in, CD-CNN is the only tracker that can successfully handle the rapid scale variation.

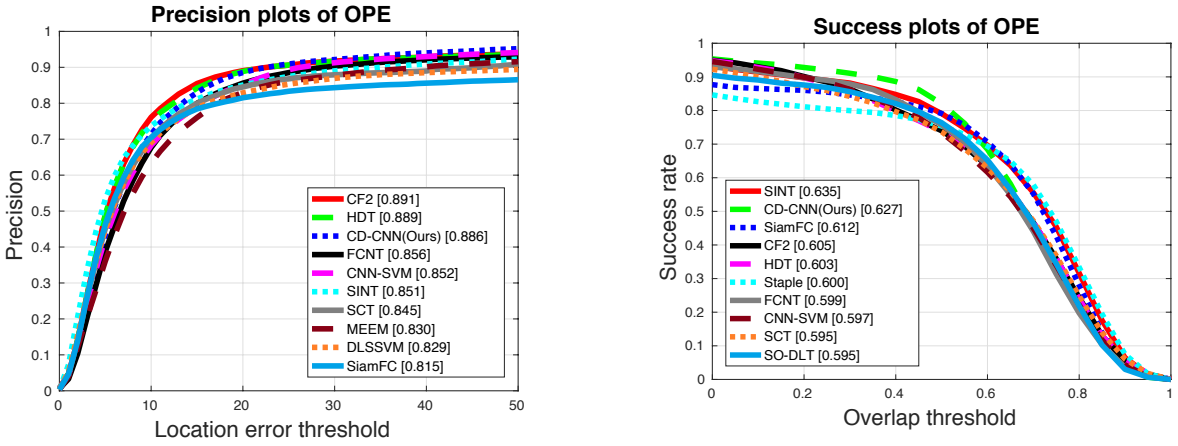

Figure 3: Precision plots and success plots for the overall performance comparison on OTB2013.

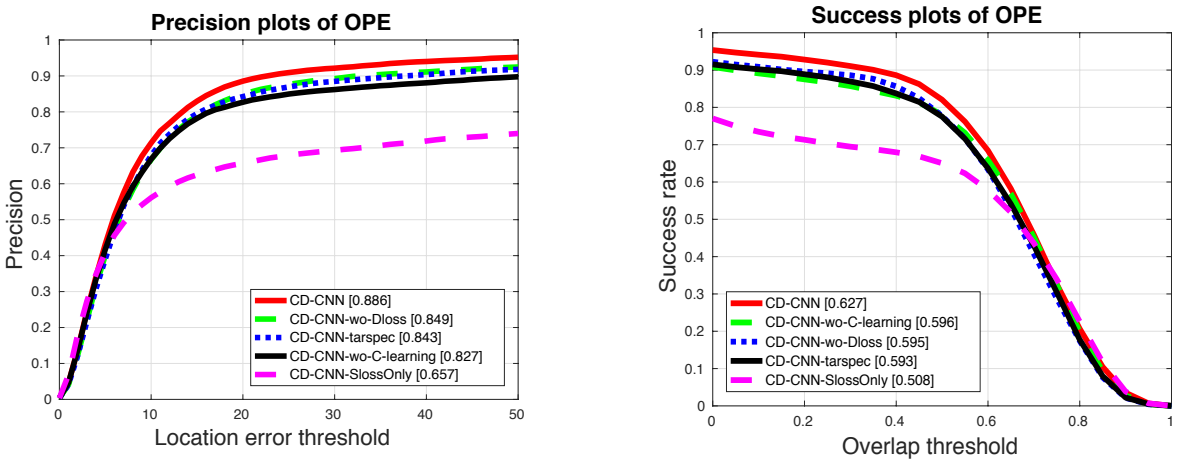

Figure 4: Precision plots and success plots of OPE for the internal comparisons on OTB2013.

In the Football sequence, SINT, which focuses on learning an implicit matching function, fails to track the target when a similar object appears nearby. Apparently, its lack of discriminability causes the drifting.

In #112 of the Ironman sequence, when the head gets out of the view, only CD-CNN can successfully estimate the target position. This benefits from the temporal appearance continuity learning. Again, in #117 and #141, our CD-CNN apparently outperforms other trackers while the similarity matching based tracker including SINT, SiamFC and CF2 drifts into similar background.

In the end of the Skiing sequence, only CD-CNN yields a tight bounding box for the target. This is again attributed to the object-centroid discrimination of our model.

Figure 5: Qualitative comparisons on some challenging sequences including CarScale, Doll, Dudek, Football, Ironman, Skiing and Woman.