Supplementary Material
Learning the Model Update for Siamese Trackers

Lichao Zhang[1], Abel Gonzalez-Garcia[1], Joost van de Weijer[1], Martin Danelljan[2], Fahad Shahbaz Khan[3,4]
[1] Computer Vision Center, Universitat Autonoma de Barcelona, Spain
[2] Computer Vision Laboratory, ETH Zürich, Switzerland
[3] Inception Institute of Artificial Intelligence, UAE
[4] Computer Vision Laboratory, Linköping University, Sweden
{lichao,agonzalez,joost}@cvc.uab.es, martin.danelljan@vision.ee.ethz.ch, fahad.khan@liu.se

S1. Difference with offline weighted fusion

Another possible update mechanism that improves on the linear update could be learning an offline weighted fusion of three templates (the input of UpdateNet). In order to demonstrate the benefits of more sophisticated updating mechanisms, we propose the following experiment.

UpdateNet uses a convolutional neural network that leverages previous templates to predict an accumulated template that is similar to the real one. Therefore, the fusion mechanism implemented by UpdateNet is more sophisticated than a simple offline weighted fusion, which depends on the actual input features and can be adapted accordingly. In order to confirm this, we propose here the following experiment to compare UpdateNet to an offline weighted fusion. We express the template update as a weighted linear combination $\widetilde{T_i} = \alpha_{init} T_0^{GT} + \alpha_{accu} \widetilde{T}_{i-1} + \alpha_{curr} T_i$. We initialize the three weights to 0, 0.9898, and 0.0102 respectively following the default settings for the linear update. Then, we train these weights with the same training process as UpdateNet until convergence, as shown in Figure S1.

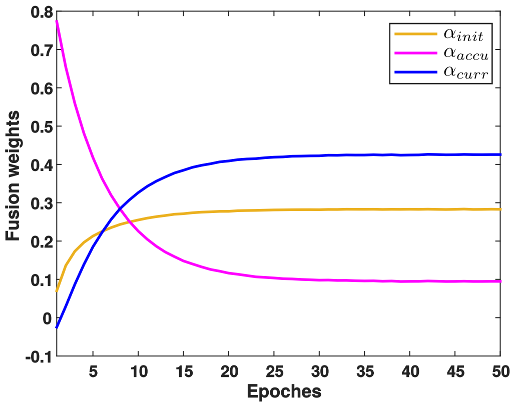

Figure S1. Training of the learned fusion weights offline.

Next we compare the EAO results on VOT2018. SiamFC with an offline weighted fusion achieves *0.198*, which is a little higher than the baseline linear update (*0.188*) but much lower than *0.262* achieved with UpdateNet. These results show that our UpdateNet is significantly better than an offline weighted fusion. We mainly attribute the superior results of UpdateNet to the following reasons. The offline weighted fusion learns a high value for α_{curr}. In this case, the tracker is likely to succumb to drift when the current template is not reliable. Instead of excessively relying on the current template, UpdateNet can benefit from all the input templates due to the representation bottleneck in the channel dimension. Furthermore, UpdateNet includes a non-linearity which allows it to better adjust to the non-linear variations, such as rotation and object motion. Thus, yielding more expressive and reliable representations for the predicted template.

S2. Visualization of updating templates

We provide additional accumulated templates of SiamFC for both linear update and UpdateNet in Figure S2 (similar to Figure 3 in the paper). By visualizing more exemplar videos, we can see that UpdateNet learns templates which are more similar to the ground-truth and predicts more accurate response maps for cross-correlation. For visualization ease, we add a red cross ‘+’ to split the four channels of the template feature. The four channels are the most dynamic channels in the ground-truth template for the corresponding video. We select them as follows. For each $j \in \{1, ..., C\}$ we compute the average difference in the template as $\Delta_j = \frac{1}{|N|}\sum_N \frac{1}{|A|}\sum_A |T_i^{GT} - T_{i-1}^{GT}|$, where N is the number of frames in a video and the sum runs over the spatial area of each channel of the feature maps (e.g. $A = 6 \times 6$). We select largest 4 channels in terms of Δ_j.

We can observe multiple interesting behaviors in Figure S2. Firstly, the accumulated templates using UpdateNet resemble the ground-truth more closely than those with lin-

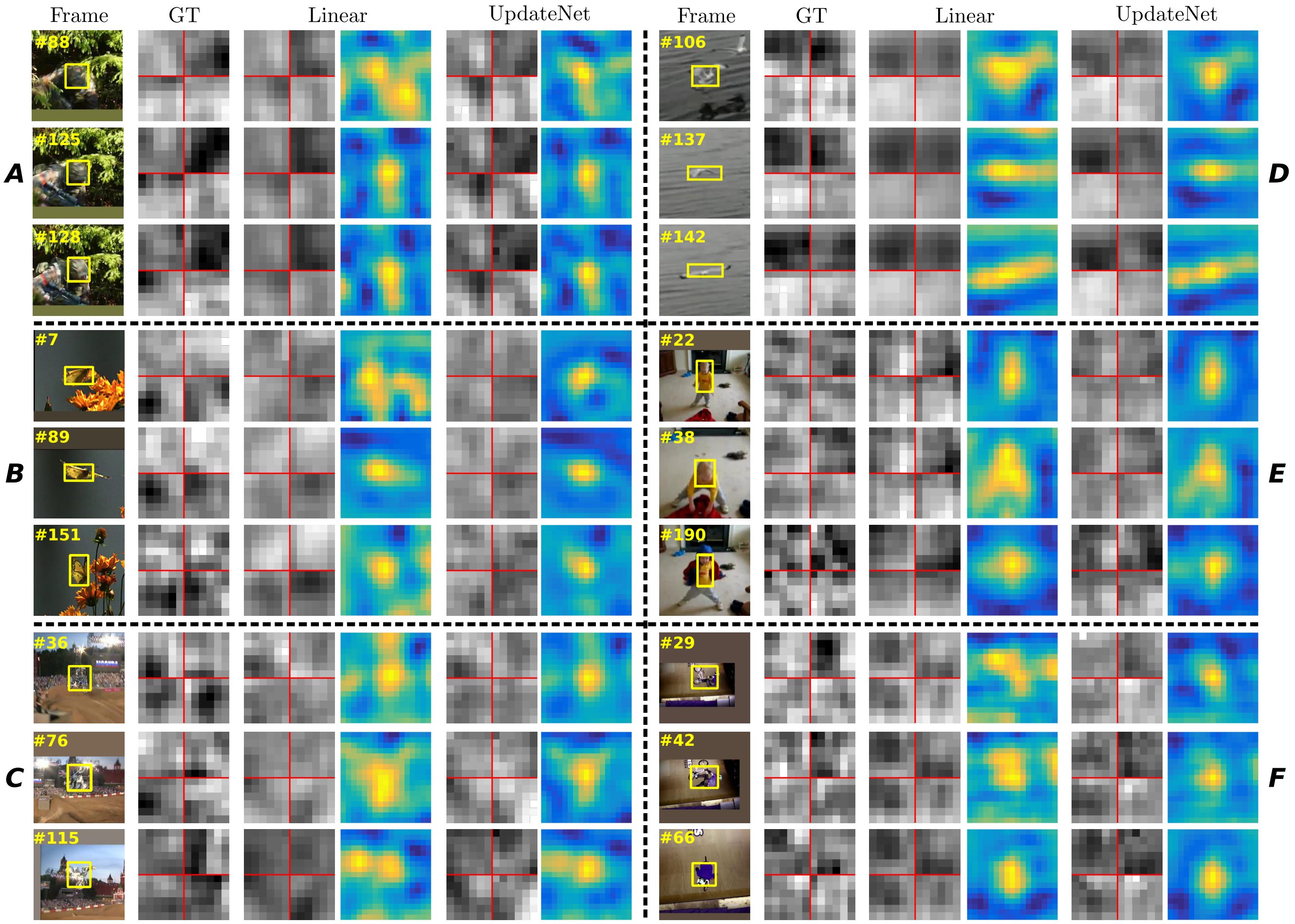

Figure S2. **Visualization accumulated and ground-truth templates for SiamFC.** The first column shows the search region and the ground-truth box. 'GT' shows top four channels of the real template extracted from the ground-truth box. For each update strategy ('Linear' and 'UpdateNet') we show the accumulated templates and the resulting response map when applied to the search region, respectively.

ear update, see e.g. the bottom-right channel (106) in frame 125 of ***A***, and the bottom-right channel (121) in frame 42 of ***F***, where the same highlighted region appears for both UpdateNet and the ground-truth. The template of the linear update, instead, does not resemble the ground truth for either of these two sequences. In general, the accumulated templates from UpdateNet are almost as dynamic as the ground-truth templates, meaning that our UpdateNet can adapt to the template change in a video much better than linear update, which changes very slowly. Secondly, we can see in the cross-correlation response map how UpdateNet better predicts the object location, while linear update predicts many spurious peaks on the response map and the true peak in the center is less sharp, see e.g. frame 115 in example ***C*** with an additional peak, frame 76 in example ***C*** and frame 106 in example ***D*** with blurred peaks, frame 88 in example ***A*** and frame 7 in example ***B*** with multiple peaks, among others. To summarize, Figure S2 clearly shows that our strategy does not negatively interfere with the desired correlation properties of the learned features, on the contrary, it helps by adaptingly updating the templates. On the other hand, the accumulated templates of the linear update change at a very slow rate and are inefficient in keeping up with the appearance variation exhibited in videos.

S3. Change rate for update

In addition to Figure 4 in the paper, we here provide similar results for the sequences shown in Figure S2 of this supplementary material. We calculate the change rate δ between templates of contiguous frames and show the results in Figure S3. Our UpdateNet provides an adaptive update strategy that is close to the change rate of the real template, while linear update can only offer a constant change rate. The change rate for UpdateNet follows the same trends as the ground-truth, see for example the high correlation with the high peaks in e.g. frame 50 in 'soldier', frame 60 in 'butterfly' and frame 61 in 'blanket'. This leads to predicting better response maps as shown in Figure S2.

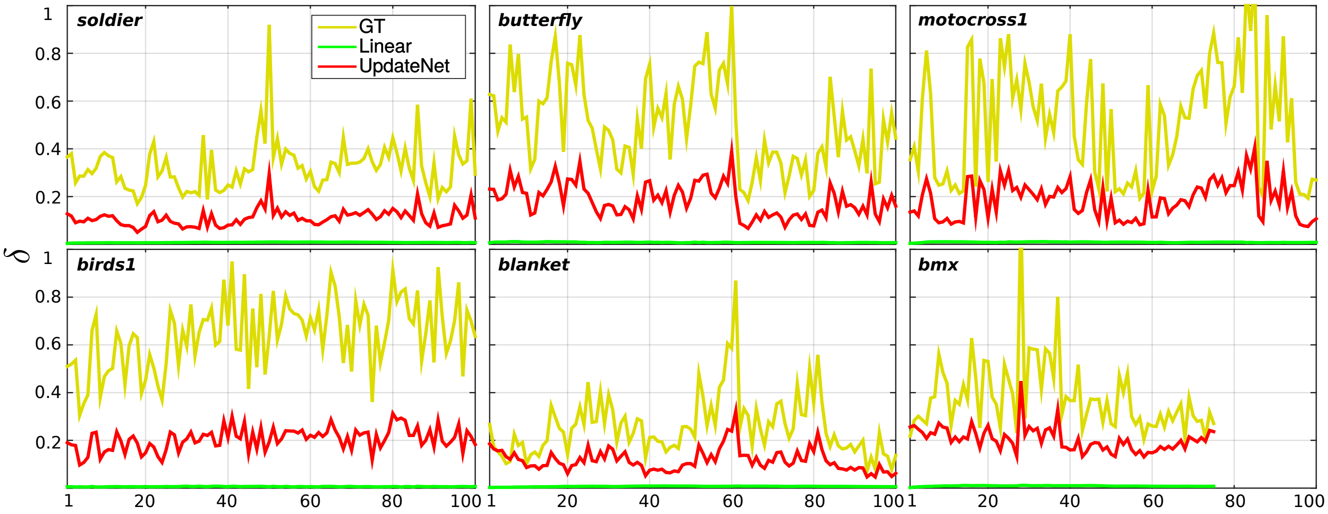

Figure S3. **Change rate between contiguous frames.** We present additional results for six example videos in VOT2018.