Supplementary Material for HiFT: Hierarchical Feature Transformer for Aerial Tracking

Ziang Cao[†], Changhong Fu[†,*], Junjie Ye[†], Bowen Li[†], and Yiming Li[‡]
[†]Tongji University [‡] New York University
caoang233@gmail.com, changhongfu@tongji.edu.cn, yimingli@nyu.edu

1. Overview

To further demonstrate the validity of the novel transformer and the rationality of its structure, here presents supplemental experiments, more qualitative comparisons with other state-of-the-art (SOTA) trackers, supplemental implementation details, and comprehensive attribute-based experimental results of the proposed HiFT tracker.

2. Supplemental experiments

2.1. Effectiveness

To demonstrate the advantages of the hierarchical feature transformer, *i.e.*, comparable or better representative ability but faster-processing speed than the deep CNN, we also conducted experiments replacing AlexNet [2]+feature transformer structure with deep CNN, *i.e.*, ResNet [1]. Table 1 shows that HiFT provides an effective way to introduce global context without sacrificing efficiency.

2.2. Rationality

To validate the rationality of the transformation based on low-resolution features, we compare with other combination orders of the 3 level features in Table 2. From the first two rows, we can discover that when the features from the fifth layer $\mathbf{M}_5$ are treated as the input of the decoder, the result is better than those using other level features. Besides, albeit using $\mathbf{M_4}$ and $\mathbf{M_3}$ (second row) can obtain good performance, the $\mathbf{M_3}$ is more appropriate than $\mathbf{M_4}$ as the value of the encoder (first row) owing to more detail information. Comprehensive experiments demonstrate our superiority on average precision (**4**%) and success (**3.2**%) against the second-best structure, proving that it is appropriate to utilize $\mathbf{M_3}$ as the value of the encoder and use $\mathbf{M_5}$ as the bedrock of the feature transformation.

Table 1. Comparison between HiFT (AlexNet) and pure Siamese-based tracker without feature transformer (ResNet) on UAV123@10fps [4].

	Precision	Success	FPS
Baseline (ResNet)	0.722	0.528	57
HiFT (AlexNet)	**0.754**	**0.574**	**132**

3. More qualitative comparisons

The abundant superior qualitative results of the proposed HiFT tracker against other arts on four authoritative benchmarks are shown in Fig. 2.

4. Detailed calculation of classification & regression

For clarity, we denote the output of classification and regression network as $\mathbf{R}_{cls1} \in \mathbb{R}^{W\times H\times 1}$, $\mathbf{R}_{cls2} \in \mathbb{R}^{W\times H\times 1}$, and $\mathbf{R}_{loc} \in \mathbb{R}^{W\times H\times 4}$. The width and height of search patch is denoted as W_s and H_s. Besides, $N_i,\ i = 1, 2, 3, 4$ are the scaling proportions. In this paper, $(g_{x1}, g_{y1}), (g_{x2}, g_{y2})$ denote the coordinates of the top-left corner and bottom-right corner of ground truth box on the transformed feature maps while $(G_{x1}, G_{y1}), (G_{x2}, G_{y2})$ represent the coordinates of the top-left corner and bottom-right corner of ground truth on search image. There is a certain mapping $\mathcal{H}$ between them as (take g_{x1} for example):

$$\mathcal{H}(g_{x1}) = s \times (g_{x1} - \frac{W}{2}) + \frac{W_s}{2} = G_{x1} \quad , \qquad (1)$$

where s represents the total stride of the network.

Besides, the length and width of R1 and R2 are denoted

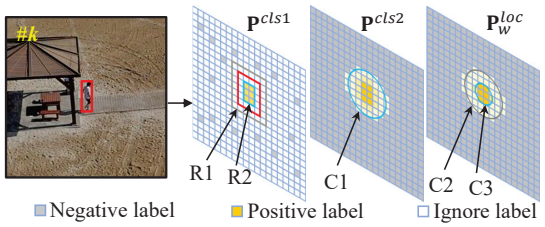

Figure 1. Visualization of classification labels $\mathbf{P}^{cls1}$, $\mathbf{P}^{cls2}$, and regression mask $\mathbf{P}_W^{loc}$. The C1, C2, C3, R1, and R2 represent the circular and rectangle areas with different scales whose centers are the same as ground truth. (Best viewed in color version)

Table 2. Validation study of the proposed structure. Note that Encoder_1 represents the feature map treated as the value of the feature encoder while Encoder_2 is the input of the modulation layer.

Encoder		Decoder	DTB70		UAV123		UAV123@10fps		UAV20L		Average	
Encoder_1	Encoder_2	Decoder	Prec.	Succ.	Prec.	Succ.	Prec.	Succ.	Prec.	Succ.	Prec.	Succ.
$\mathbf{M}_3$	$\mathbf{M}_4$	$\mathbf{M}_5$	**0.802**	**0.594**	**0.787**	**0.589**	0.754	0.574	**0.763**	**0.566**	**0.776**	**0.581**
$\mathbf{M}_4$	$\mathbf{M}_3$	$\mathbf{M}_5$	0.764	0.576	0.756	0.574	**0.759**	**0.577**	0.692	0.526	0.743	0.563
$\mathbf{M}_3$	$\mathbf{M}_5$	$\mathbf{M}_4$	0.686	0.518	0.729	0.555	0.743	0.565	0.642	0.485	0.700	0.531
$\mathbf{M}_4$	$\mathbf{M}_5$	$\mathbf{M}_3$	0.699	0.511	0.712	0.526	0.709	0.529	0.650	0.497	0.692	0.516
$\mathbf{M}_5$	$\mathbf{M}_4$	$\mathbf{M}_3$	0.686	0.510	0.700	0.517	0.719	0.533	0.648	0.498	0.688	0.514
$\mathbf{M}_5$	$\mathbf{M}_3$	$\mathbf{M}_4$	0.678	0.513	0.744	0.568	0.734	0.560	0.633	0.485	0.697	0.531

as l_{r1}, w_{r1}, l_{r2}, and w_{r2}, which can be calculated by:

$$\begin{aligned} l_{r1} &= (g_{x2} - g_{x1}) \times (1 + \frac{1}{N_1}) \\ w_{r1} &= (g_{y2} - g_{y1}) \times (1 + \frac{1}{N_1}) \\ l_{r2} &= (g_{x2} - g_{x1}) \times (1 + \frac{1}{N_2}) \\ w_{r2} &= (g_{y2} - g_{y1}) \times (1 + \frac{1}{N_2}) \end{aligned} . \tag{2}$$

As shown in Fig. 1, the first classification label $\mathbf{P}^{cls1} \in \mathbb{R}^{W \times H \times 1}$ can be expressed by:

$$\mathbf{P}^{cls1}(i,j) = \begin{cases} 1, & (i,j) \text{ in R2} \\ -2, & (i,j) \text{ in } \complement_{\text{R1}}\text{R2} \\ 0, & (i,j) \text{ in } \mathcal{T}(\text{R1}) \end{cases} , \tag{3}$$

where $\mathcal{T}(R1)$ aims to limit the number of negative labels randomly.

Besides, $\mathbf{P}^{cls2} \in \mathbb{R}^{W \times H \times 1}$ is implemented to distinguish the closer points to ground truth based on euclidean distance between the corresponding points and the center of ground truth. Considering the core area of classification is the region around ground truth. We set the points outside of the C1 as negative samples and the inside of the C2 as positive samples. Thus, the $\mathbf{P}^{cls2}$ can be written as:

$$\begin{aligned} &D(i,j) = (i - \frac{g_{x1} + g_{x2}}{2})^2 + (j - \frac{g_{y1} + g_{y2}}{2})^2 \\ &\text{C1} = \left\{ (i,j) |\ D(i,j) < (\frac{g_{x1} - g_{x2}}{N_3})^2 + (\frac{g_{y1} - g_{y2}}{N_3})^2 \right\} \\ &Dis(i,j) = 1 - \frac{D(i,j) - \min(\mathbf{D})}{\max(\mathbf{D}) - \min(\mathbf{D})} \\ &\mathbf{P}^{cls2}(i,j) = \begin{cases} Dis(i,j), & (i,j) \text{ in C1} \\ 0, & (i,j) \text{ out C1} \end{cases} \end{aligned} , \tag{4}$$

where $Dis(i,j)$ represents the score calculated by Euclidean distance between the point (i,j) and ground truth while $\mathbf{D}$ is the matrix containing all elements $D(i,j)$.

Similar to $\mathbf{P}^{cls2}$, the mask of regression, *i.e.*, $\mathbf{P}_W^{loc}$ can be calculated by:

$$\begin{aligned} &\text{C2} = \left\{ (i,j) |\ D(i,j) < (\frac{g_{x1} - g_{x2}}{N_4})^2 + (\frac{g_{y1} - g_{y2}}{N_4})^2 \right\} \\ &\text{C3} = \{(i,j) | Dis(i,j) > 0.8\} \\ &\mathbf{P}_W^{loc}(i,j) = \begin{cases} \mathcal{T}'(1.5, Dis(i,j)), & (i,j) \text{ in C3} \\ Dis(i,j), & (i,j) \text{ in } \complement_{\text{C2}}\text{C3} \\ 0, & (i,j) \text{ out C2} \end{cases} \end{aligned} , \tag{5}$$

where $\mathcal{T}'(a,b)$ it represents stochastic decision-maker between a and b, aiming to constrain the number of key points whose weight is 1.5 shown in Fig. 1.

Besides, we normalized the regression label using a function as follows:

$$\text{Transfer}(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} . \tag{6}$$

Thus, the regression label $\mathbf{P}^{loc} \in \mathbb{R}^{W \times H \times 4}$ can be obtained by:

$$\begin{aligned} \mathbf{P}^{loc}[i,j,0] &= \text{Transfer}(\frac{2(\mathcal{H}(i) - G_{x1})}{W_s}) \\ \mathbf{P}^{loc}[i,j,1] &= \text{Transfer}(\frac{2(G_{x2} - \mathcal{H}(i))}{W_s}) \\ \mathbf{P}^{loc}[i,j,2] &= \text{Transfer}(\frac{2(\mathcal{H}(j) - G_{y1})}{H_s}) \\ \mathbf{P}^{loc}[i,j,3] &= \text{Transfer}(\frac{2(G_{y2} - \mathcal{H}(j))}{H_s}) \end{aligned} . \tag{7}$$

Eventually, the overall loss function can be determined as:

$$\begin{aligned} L_{overall} &= \lambda_1 L_{cls1}(\mathbf{R}_{cls1}, \mathbf{P}^{cls1}) \\ &+ \lambda_2 L_{cls2}(\mathbf{R}_{cls2}, \mathbf{P}^{cls2}) + \lambda_3 \mathbf{P}_W^{loc} L_{loc}(\mathbf{R}_{loc}, \mathbf{P}^{loc}) \end{aligned} , \tag{8}$$

where λ_1, λ_2, and λ_3 are the coefficients to balance the contributions of each loss.

5. Exhaustive attribute-based evaluations

To better illustrate the effectiveness of HiFT under various aerial special challenges, attribute-based evaluations on UAV20L [4] are conducted. The comparisons among other SOTA trackers and HiFT with different components are shown in Table 3, which demonstrates the strength of the hierarchical feature transformer, modulation layer, and circular classification labels. More attribute-based evaluation results on DTB70 [3], UAV123 [4], and UAV123@10fps [4] are also presented in Fig. 3, Fig. 4, and Fig. 5, respectively.

6. Summary

During the whole experiment, we have the following discoveries:

- Applying the novel feature transformer on light-weight CNN can achieve comparable performance while keeping efficient.
- Object queries in the original transformer structure and direct positional encoding for low-resolution feature maps can hurt the original structure of semantic information, impeding the tracking performance.
- By fully utilizing the interdependencies between different level features created by the modulation layer, the discriminability of HiFT can be further raised. Besides, the circular label can keep the tracker focus on the areas closing to ground truth.
- Adopting high-resolution feature maps as the value of the encoder while low-resolution features as the bedrock of the decoder is the most appropriate for the feature transformer to discover the tracking-tailored feature space.

To our best knowledge, we are the first to attempt to deploy the Siamese-Transformer structure on the embedded platform for real-time aerial tracking. Besides, our hierarchical structure indeed achieves superior performance in the aerial tracking scenarios. Thus, we are convinced that our efficient and robust HiFT can promote the development of aerial tracking-related applications.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1097–1105, 2012.

[3] S. Li and D. Yeung. Visual Object Tracking for Unmanned Aerial Vehicles: A Benchmark and New Motion Models. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 1–7, 2017.

[4] M. Mueller, N. Smith, and B. Ghanem. A Benchmark and Simulator for UAV Tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 445–461, 2016.

Figure 2. Qualitative comparisons among HiFT and other top-performing trackers. Missing trackers are denoted as "x" using the corresponding color. Best viewed on screen with high-resolution.

Table 3. Attribute-based evaluations on UAV20L [4]. It shows the different components of HiFT under various aerial-specific challenges. ARC, CM, FOC, IV, LR, OV, POC, SV, SOB, VC, and Overall represents aspect ratio change, camera motion, full occlusion, low resolution, out of view, partial occlusion, scale variation, similar objects, viewpoint change, and overall performance. The best three performances are respectively highlighted with **red**, **green**, and **blue** color.

	ARC		CM		FOC		IV		LR		OV		POC		SV		SOB		VC		Overall	
	Prec.	Succ.	Prec.	Succ.	Prec.	Succ.	Prec.	Succ.	Prec.	Succ.	Prec.	Succ.	Prec.	Succ.	Prec.	Succ.	Prec.	Succ.	Prec.	Succ.	Prec.	Succ.
AutoTrack	0.418	0.277	0.487	0.329	0.403	0.198	0.443	0.321	0.425	0.238	0.506	0.325	0.490	0.319	0.487	0.330	0.449	0.353	0.420	0.303	0.512	0.349
ARCF	0.476	0.320	0.544	0.372	0.401	0.205	0.542	0.380	0.481	0.273	0.531	0.362	0.542	0.365	0.522	0.366	0.543	0.415	0.457	0.339	0.544	0.381
SiamRPN++	**0.620**	**0.466**	**0.680**	**0.512**	0.429	0.254	**0.655**	**0.507**	0.544	**0.368**	**0.689**	**0.521**	**0.666**	**0.499**	**0.680**	**0.520**	**0.718**	**0.575**	**0.626**	0.503	**0.696**	**0.528**
STRCF	0.472	0.331	0.553	0.393	0.406	0.217	0.429	0.346	0.513	0.293	0.525	0.376	0.564	0.402	0.553	0.394	0.547	0.440	0.441	0.333	0.575	0.411
fDSST	0.327	0.252	0.365	0.277	0.318	0.154	0.296	0.259	0.350	0.205	0.372	0.279	0.360	0.264	0.387	0.290	0.377	0.330	0.318	0.270	0.385	0.288
SRDCF	0.389	0.270	0.482	0.327	0.331	0.170	0.411	0.295	0.429	0.228	0.495	0.329	0.491	0.320	0.481	0.332	0.522	0.397	0.414	0.303	0.507	0.343
CoKCF	0.405	0.237	0.481	0.286	0.373	0.175	0.442	0.325	0.449	0.195	0.422	0.251	0.464	0.272	0.481	0.281	0.465	0.299	0.404	0.249	0.507	0.298
KCF	0.224	0.142	0.301	0.184	0.264	0.115	0.218	0.184	0.274	0.119	0.312	0.189	0.310	0.190	0.275	0.175	0.271	0.190	0.189	0.148	0.311	0.196
BACF	0.482	0.345	0.562	0.404	0.378	0.200	0.524	0.410	0.463	0.270	0.568	0.401	0.566	0.398	0.562	0.399	0.581	0.466	0.500	0.373	0.584	0.415
DaSiamRPN	0.589	0.411	0.648	0.455	**0.502**	0.269	0.545	0.401	**0.572**	0.338	0.662	0.485	0.633	0.443	0.648	0.454	0.620	0.476	0.584	0.447	0.665	0.465
C-COT	0.460	0.326	0.541	0.383	0.359	0.183	0.504	0.380	0.504	0.283	0.492	0.356	0.526	0.371	0.538	0.380	0.589	0.462	0.449	0.334	0.561	0.395
SiameseFC	0.501	0.335	0.578	0.389	0.465	0.252	0.492	0.367	0.425	0.233	0.611	0.400	0.556	0.370	0.578	0.388	0.581	0.432	0.524	0.362	0.599	0.402
UDT+	0.493	0.332	0.566	0.389	0.423	0.223	0.545	0.369	0.497	0.300	0.532	0.372	0.551	0.382	0.563	0.389	0.587	0.446	0.465	0.327	0.585	0.401
UDT	0.446	0.324	0.496	0.356	0.427	0.233	0.437	0.344	0.445	0.278	0.478	0.332	0.487	0.345	0.489	0.348	0.521	0.408	0.402	0.311	0.514	0.363
TADT	0.521	0.395	0.588	0.442	0.444	0.261	0.518	0.430	0.550	0.338	0.534	0.415	0.577	0.433	0.588	0.445	0.587	0.483	0.505	0.415	0.609	0.459
DeepSTRCF	0.488	0.370	0.566	0.425	0.429	0.236	0.523	0.416	0.512	0.317	0.549	0.420	0.556	0.417	0.566	0.428	0.563	0.465	0.503	0.412	0.588	0.443
MCCT	0.516	0.341	0.584	0.388	0.418	0.219	0.563	0.368	0.475	0.280	0.575	0.379	0.573	0.379	0.586	0.403	0.618	0.450	0.495	0.352	0.605	0.407
DSiam	0.547	0.342	0.582	0.371	**0.619**	**0.335**	0.519	0.357	**0.634**	0.354	0.538	0.336	0.578	0.366	0.585	0.379	0.513	0.362	0.526	0.348	0.603	0.391
ECO	0.489	0.359	0.567	0.412	0.409	0.223	0.551	0.428	0.486	0.283	0.546	0.400	0.554	0.401	0.567	0.415	0.559	0.454	0.507	0.391	0.589	0.427
Baseline	0.562	0.424	0.590	0.443	0.417	0.261	0.577	0.437	0.434	0.293	0.643	0.488	0.571	0.426	0.590	0.452	0.614	0.487	0.624	**0.508**	0.611	0.463
Baseline+OT	0.497	0.376	0.576	0.427	0.355	0.203	0.519	0.388	0.440	0.297	0.642	0.486	0.559	0.412	0.576	0.436	0.592	0.476	0.547	0.455	0.597	0.446
Baseline+FT	0.595	0.437	0.658	0.478	0.462	0.274	**0.581**	**0.448**	0.525	0.340	0.674	0.498	0.641	0.463	0.658	0.489	0.668	0.521	0.587	0.478	0.675	0.496
Baseline+HFT+PE	0.537	0.414	0.610	0.466	0.376	0.233	0.576	0.444	0.473	0.329	0.669	0.519	0.590	0.449	0.610	0.476	0.644	0.523	0.595	0.497	0.629	0.486
Baseline+HFT+RL	**0.613**	**0.460**	**0.673**	**0.506**	0.438	**0.276**	0.573	0.434	0.548	**0.378**	**0.724**	**0.555**	**0.660**	**0.493**	**0.673**	**0.515**	**0.720**	**0.574**	**0.645**	**0.525**	**0.689**	**0.523**
HiFT	**0.704**	**0.515**	**0.751**	**0.551**	**0.601**	**0.369**	**0.657**	**0.475**	**0.662**	**0.448**	**0.783**	**0.595**	**0.742**	**0.541**	**0.751**	**0.561**	**0.752**	**0.581**	**0.713**	**0.572**	**0.763**	**0.566**

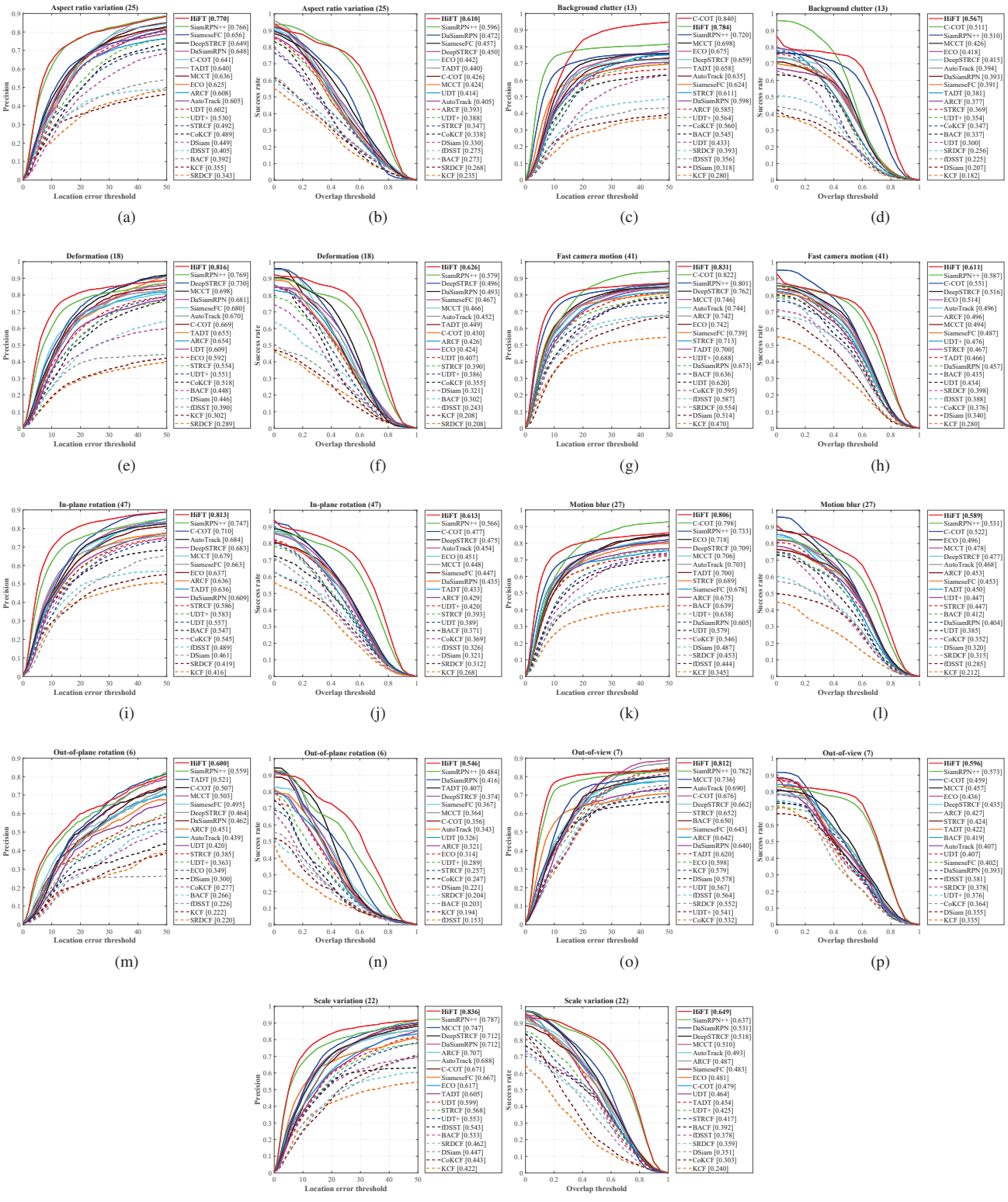

Figure 3. Attribute-based comparison among all trackers on DTB70 [3]. Owing to the transformed feature in the tracking-tailored feature space created by the hierarchical feature transformer, our HiFT outperforms other trackers, especially in motion and deformation conditions.

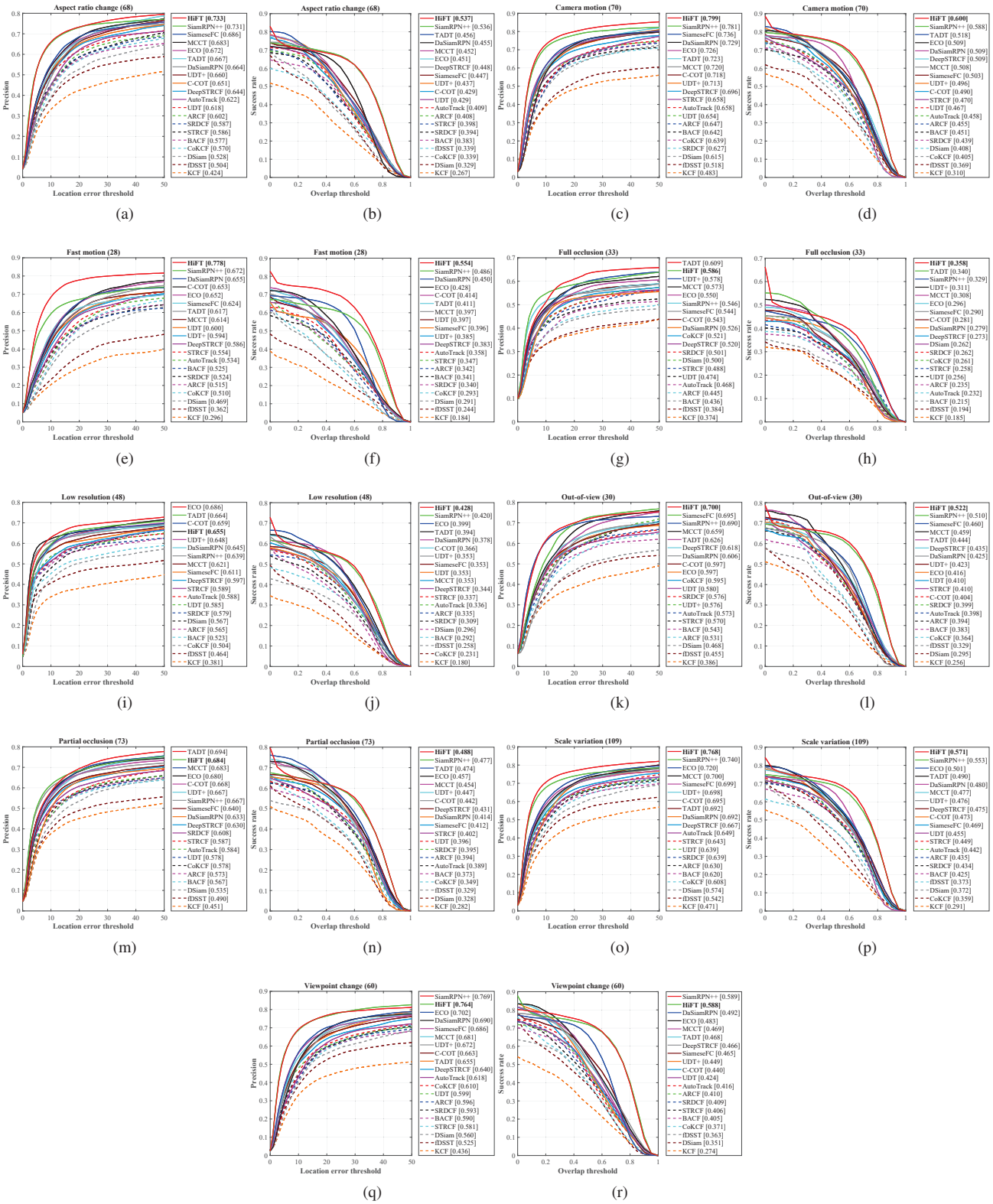

Figure 4. Attribute-based comparison among all trackers on UAV123 [4]. It demonstrates the effectiveness of the hierarchical structure in various UAV-specific challenges.

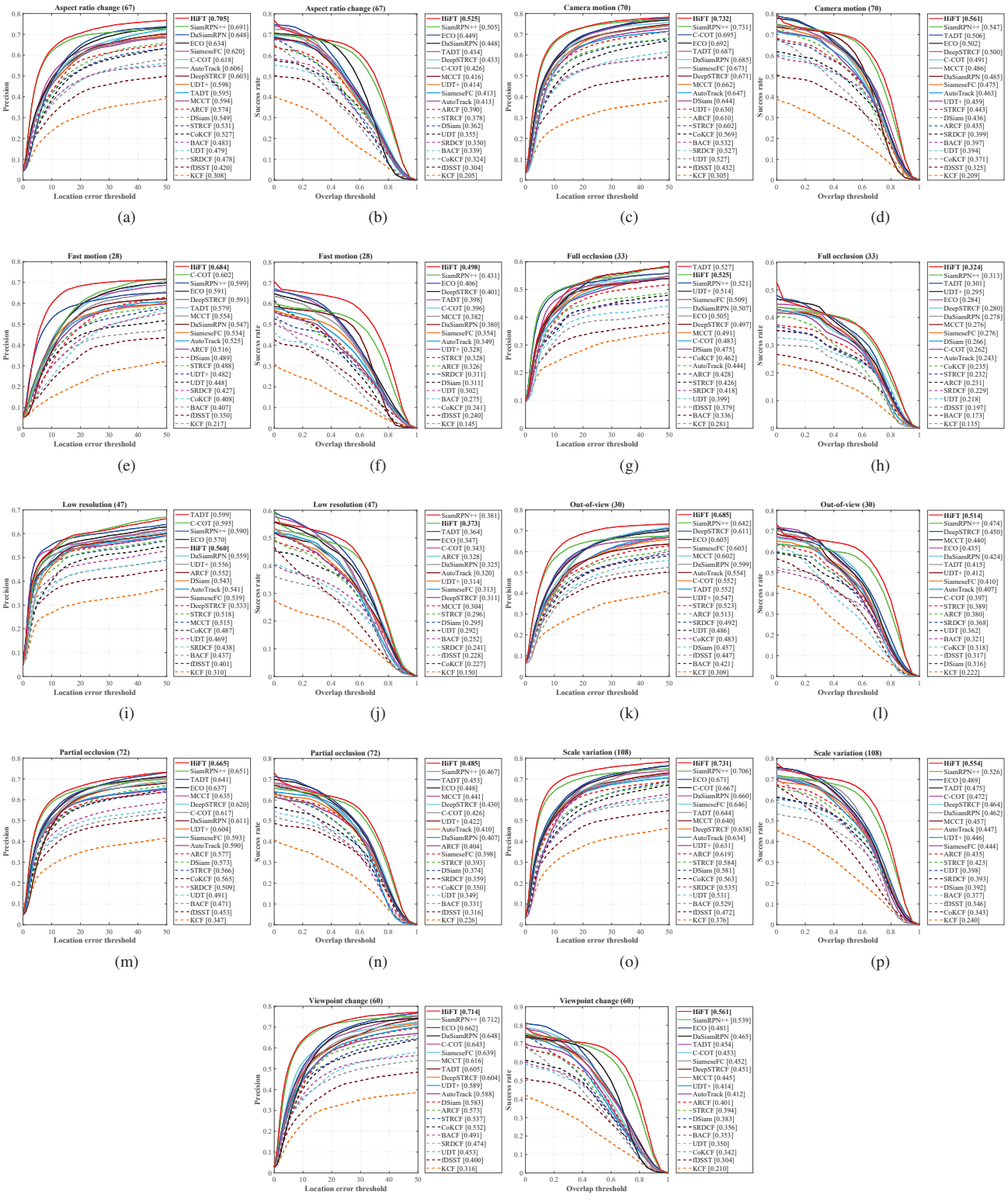

Figure 5. Attribute-based comparison among all trackers on UAV123@10fps [4]. Competitive performance on UAV123@10fps proves the promising robustness of our HiFT under severe motion and variation.