Stable Autofocus with Focal Consistency Loss
– *Supplementary Document* –

In this supplementary document, we report the additional experimental results and analyses that are not included in the main paper due to page limit. In Sec. A, we describe the implementation details, including the precise architectural modifications for multi-frame settings. In Sec. B, we study an alternative symmetric form of KL divergence. In Sec. C, we show the full quantitative results for various types of focal consistency loss, and demonstrate the effectiveness and generalizability of our method. In Sec. D, we provide an additional ablation study on the FCL weight hyperparameters. In Sec. E, we show the ablation study on different mini-batch configurations and justify our chosen hyperparameters. In Sec. F, we show the examples of the difficult corner cases and analyze the results. Lastly, In Sec. G, we demonstrate additional qualitative results.

A. Implementation Details

Multi-frame Architecture Details. For both baseline AF models, we minimize the architectural modifications for the model to be able to receive multi-frame inputs. For the AFPE baseline [4], the multi-frame input is constructed by concatenating the input data channel-wise. Let us denote the single-frame dual-pixel input data as $D \in \mathbb{R}^{5\times H\times W}$, where $H = W = 128$ since we use 128×128 patches for training and validation. The 5 channels consist of L/R/f/x/y, where the meaning of each channel is as follows:

- L: left pixels of the dual-pixel data
- R: right pixels of the dual-pixel data
- f: Lens-PE (position encoding) of the current focal index; the value of f is a single scalar of $0 \sim 1$, but spatially broadcasted as in [4]
- x: relative coordinates on the x-axis w.r.t. the full image (RoI-PE)
- y: relative coordinates on the y-axis w.r.t. the full image (RoI-PE)

Note that, out of 5 channels, only 3 (L/R/f) are modified w.r.t. lens movements, and the relative coordinates of the RoI (x/y) do not change. Therefore, for a multi-frame setting with m frames, we can build $(3m + 2)$-channel input tensor (instead of $5m$ channels) to fully represent the multi-frame data, resulting in $D \in \mathbb{R}^{(3m+2)\times 128\times 128}$. For the model architecture, only the first convolution layer is modified accordingly to match the different multi-frame input shape; for instance, `Conv2d(in_channel=5, out_channel=32)` for D1 would become `Conv2d(8, 32)` for D2). Likewise, for I2 $\sim$ I5 settings, we use conventional single-channel raw pixel data instead of L/R dual-pixel, resulting in $(2m + 2) \times 128 \times 128$ input shape and $2m + 2$ input channels for the first convolution layer.

For the L2A baseline [9], the input data shape for the D1 setting is already $98 \times 128 \times 128$, where the 98 channels come from 49 discrete lens positions for the full focal stack, for dual-pixel L/R data (49×2). Since L2A already has placeholders for all lens positions and only the observed frames are activated (others filled with zeros), it is straightforward to apply multi-frame inputs to the L2A baseline without any architecture modifications.

Training Details. Our model is implemented using PyTorch [45] and trained for 60k iterations using a mini-batch size of 128, where each mini-batch consists of 32 scenes each with 4 different focal indices. We used the Adam optimizer [40] with an initial learning rate of 0.001, momentum hyperparameters $(\beta_1, \beta_2) = (0.5, 0.999)$, and weight decay 0.0001. We used the cosine learning rate scheduler [42] for decaying the learning rate.

Following the previous works, all baseline models are trained with the ordinal regression loss [8]. The proposed models are trained using Eq. (9), where the consistency weights λ_{KLD} and λ_{MSD} tuned for each model (see Sec. D). In addition, we employed FCL after 30k training iterations (half of the full training schedule) to improve training convergence, since enforcing the intra-scene prediction consistency is more meaningful if the prediction model is accurate enough.

For the multi-frame models, we use the same training scheme as described above. We always obtained the consecutive input frames in increasing order (*e.g.* f_i, f_{i+1} for D2, if we start at f_i), and channel-wise concatenate the frames to

build the multi-frame input data. We acknowledge that using some simple additional cue for deciding the multi-frame *search* direction—whether to obtain the consecutive frames in increasing order or decreasing order—may reduce the overall lens movements in cases when the GT focal index is at the opposite side of the initial search direction. However, we decided that designing such cues is out of scope of the main idea in this paper and choose the increasing order to simplify our presentation. Also, when we first open the camera app in an Android smartphone (*e.g.* Google Pixel 3, Samsung Galaxy A54, Samsung Galaxy S22 Ultra), we found that the lens position always started from f_0, which makes the initial increasing search direction correct.

B. Alternative Symmetrization of KLD

In this section, we explore an alternative symmetrization of the Kullback-Leibler divergence, known as the Jensen-Shannon Divergence (JSD), and present the corresponding results and analysis.

Previously, in Sec. 3.1 of the main paper, we introduced a symmetric version of the KL divergence Eq. (3), which is referred to as the Jeffreys divergence [43]. Now, we turn our attention to the JSD [41], another symmetrization of the KLD. The JSD is computed as follows:

$$\mathcal{L}_{\text{JSD}} = \frac{1}{2}\left(D_{\text{KL}}\left(p(f_i) \parallel m\right) + D_{\text{KL}}\left(p(f_j) \parallel m\right)\right), \quad \text{where } m = \frac{p(f_i) + p(f_j)}{2}. \tag{1}$$

Note that unlike the Jeffreys divergence, the JSD is bounded, a property derived from [44]:

$$\begin{aligned} D_{\text{KL}}\left(p(f_i) \parallel m\right) &= \sum_{k=1}^{n} p_k(f_i) \log \frac{2p_k(f_i)}{p_k(f_i) + p_k(f_j)} \\ &\leq \sum_{k=1}^{n} p_k(f_i) \log \frac{2p_k(f_i)}{p_k(f_i)} = \log 2. \end{aligned}$$

Experiments were conducted using the JSD on AFPE [4]. The results are presented in Tab. A for dual-pixel and conventional images. For comparison, the results using the MSD, KLD, and FCL are also shown. While JSD yields comparable results, we can observe that the Jeffreys divergence (denoted as +KLD) performs better most of the time, which is why we chose the Jeffreys divergence as our final symmetric form of KL-Divergence.

We believe this slight performance gap can be attributed to the fact that while the Jeffreys divergence calculates the KL divergence between the original distributions $p(f_i)$ and $p(f_j)$, the JSD computes the KL divergence with respect to the average distribution $m = \frac{p(f_i)+p(f_j)}{2}$. As the average distribution tends to be smoother, the JSD may result in a smoother distribution than the Jeffreys divergence. On the other hand, the Jeffreys divergence has the potential to produce sharper distributions, which could be more effective in representing diverse scenes.

Alg.	Type	MAE	RMSE	MSD*	TV	Type	MAE	RMSE	MSD*	TV
AFPE		1.760	2.855	1.338	0.895		3.629	6.083	3.947	1.534
+MSD		1.739	2.782	1.175	0.780		3.570	6.019	3.631	1.407
+JSD	D1	1.724	2.692	1.220	0.825	I1	3.461	5.830	3.695	1.424
+KLD		1.736	2.787	1.130	0.737		3.533	5.898	3.471	1.285
+FCL		1.735	2.744	1.070	0.691		3.506	5.902	3.356	1.246
AFPE		1.656	2.593	1.161	0.606		2.547	4.358	2.647	1.401
+MSD		1.608	2.468	1.049	0.538		2.460	4.134	2.296	1.249
+JSD	D2	1.573	2.484	1.061	0.552	I2	2.476	4.206	2.340	1.324
+KLD		1.582	2.539	1.001	0.511		2.461	4.078	2.265	1.240
+FCL		1.577	2.466	0.968	0.488		2.491	4.210	2.204	1.188
AFPE		1.542	2.421	1.073	0.496		2.222	3.705	2.097	1.151
+MSD		1.495	2.346	0.951	0.428		2.189	3.657	1.907	1.006
+JSD	D3	1.516	2.352	1.005	0.450	I3	2.097	3.530	1.914	1.025
+KLD		1.488	2.382	0.923	0.398		2.156	3.572	1.864	1.010
+FCL		1.522	2.350	0.884	0.376		2.110	3.519	1.779	0.928
AFPE		1.516	2.351	1.003	0.420		1.990	3.299	1.814	0.935
+MSD		1.481	2.316	0.929	0.358		1.987	3.254	1.749	0.871
+JSD	D4	1.485	2.320	0.967	0.392	I4	1.917	3.163	1.718	0.881
+KLD		1.456	2.296	0.916	0.356		1.933	3.160	1.604	0.794
+FCL		1.434	2.279	0.868	0.319		1.958	3.202	1.557	0.790
AFPE		1.456	2.236	0.954	0.368		1.883	3.081	1.681	0.825
+MSD		1.455	2.271	0.911	0.332		1.878	3.032	1.538	0.723
+JSD	D5	1.468	2.239	0.930	0.350	I5	1.798	2.976	1.545	0.752
+KLD		1.439	2.259	0.852	0.306		1.805	3.001	1.499	0.705
+FCL		1.431	2.226	0.833	0.290		1.801	2.959	1.484	0.703
AFPE†	D*	1.356	2.128	-	-	I*	1.550	2.399	-	-

Table A. Quantitative results for multi-frame settings with dual-pixel (D1 $\sim$ D5) and conventional-image (I1 $\sim$ I5) using AFPE [4] baseline and the proposed MSD, JSD, KLD, and FCL. The top three methods for each metric are highlighted in red, orange, and yellow, respectively. We can observe that our FCL notably improves the consistency metrics MSD* and TV for all settings while preserving the accuracy. A † indicates that values are from the reference article.

C. Additional Quantitative Results

In Tab. A, we present the complete quantitative results for both dual-pixel and conventional image inputs for the AFPE baseline, including various types of the proposed FCL and the multi-frame setting. It is evident that all types of FCL (MSD, JSD, KLD, and the mixed FCL) significantly surpass the baseline AFPE performance on the consistency metrics (MSD*, TV), while maintaining or even improving the accuracy metrics (MAE, RMSE) in most cases.

Figure A illustrates the improvements in AF accuracy (MAE, RMSE) and consistency (MSD*, TV) for the dual-pixel settings D1 $\sim$ D5 (first row) and the conventional-image settings I1 $\sim$ I5 (second row) w.r.t. the number of input frames for AFPE [4]. In general, all metrics improved as we use more frames, and the proposed FCL, which is a weighted combination of MSD and KLD, shows the best performance. For I1 $\sim$ I5 settings, we can observe that using multiple frames particularly improves the overall accuracy (MAE) by a large margin. This is because a conventional single-channel RAW image (I1) has very little information about focus, while we can capture the pixel patterns

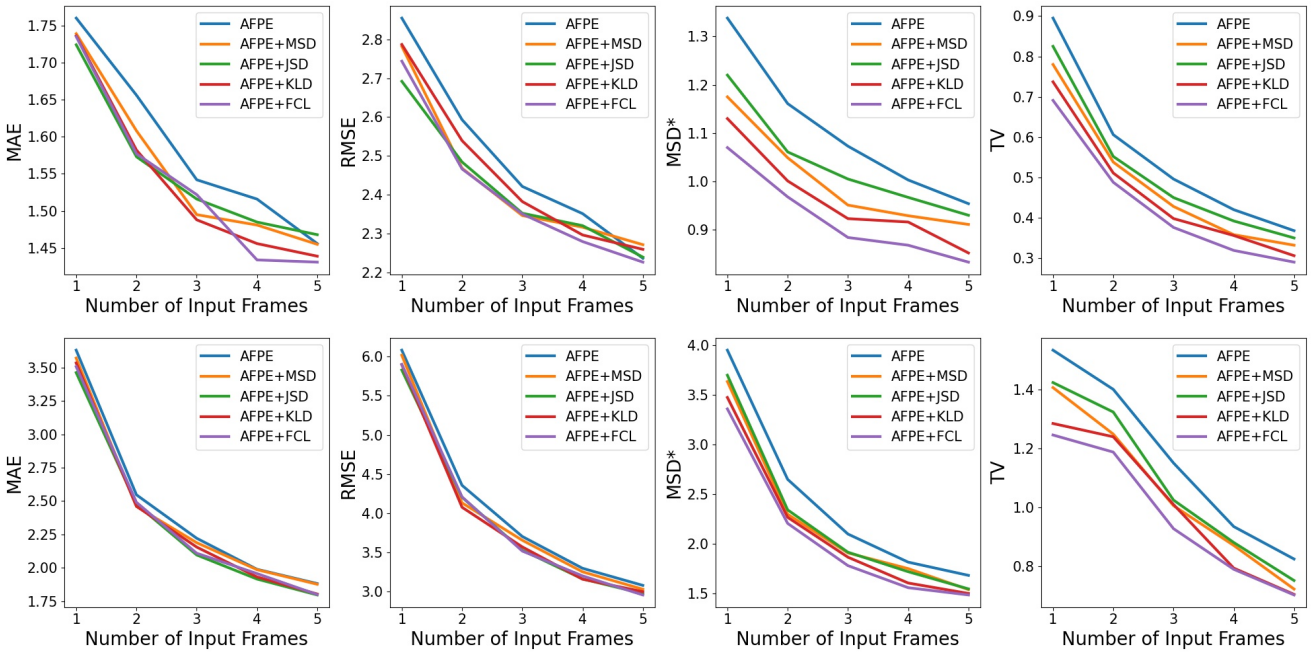

Figure A. AF accuracy and consistency comparisons w.r.t. the number of input frames for the baseline model AFPE [4] for different types of FCL (1st row: D1 $\sim$ D5, 2nd row: I1 $\sim$ I5). The results prove the effectiveness of all types of FCL (MSD, JSD, KLD, and the combined FCL).

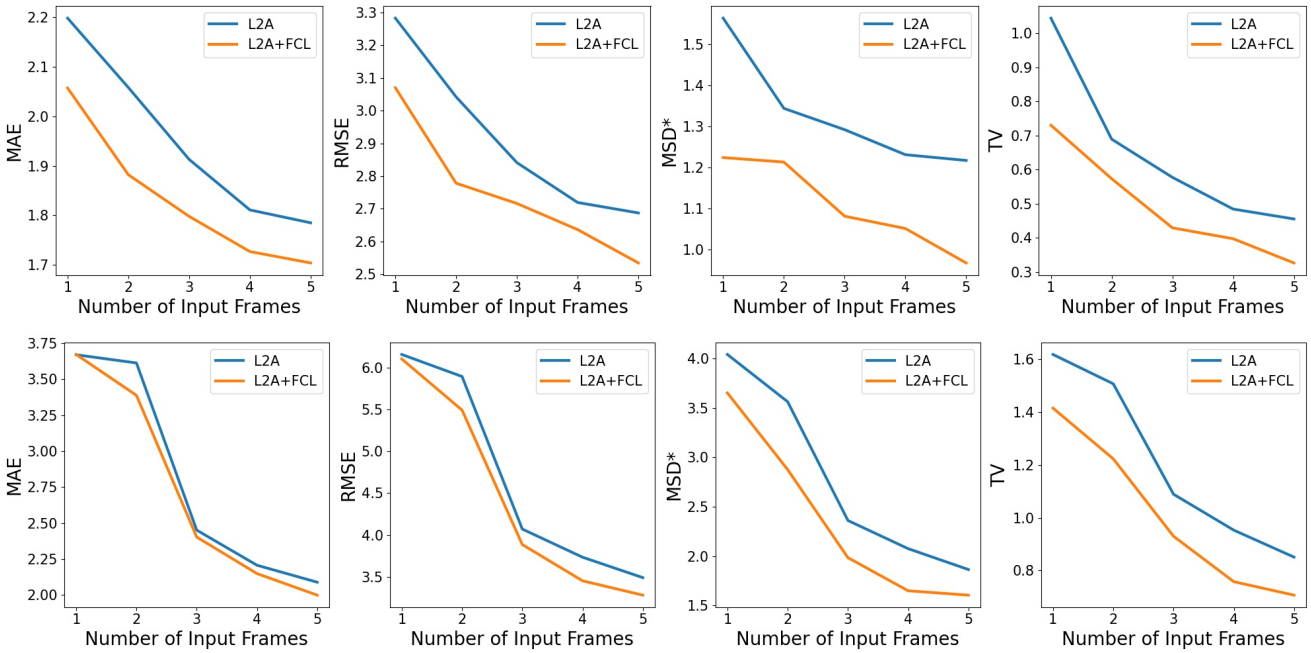

Figure B. AF accuracy and consistency comparisons w.r.t. the number of input frames for the baseline model L2A [9] with the proposed FCL (1st row: D1 $\sim$ D5, 2nd row: I1 $\sim$ I5). The results prove the effectiveness of FCL.

regarding depth if we use 2 or more frames. Also, all types of focal consistency loss (MSD, JSD, KLD, and FCL) are able to enhance the consistency metrics (MSD* and TV) compared with the baseline AFPE, which proves the effectiveness of the proposed FCL. Given that FCL does not introduce any additional computation at the inference stage, we can conclude that using FCL is always a considerable choice when training a learning-based AF model.

Similar patterns can be observed in Fig. B if we change the baseline model to L2A [9] (first row: D1 $\sim$ D5, second row: I1 $\sim$ I5). We can observe that the overall accuracy (MAE, RMSE) is consistently improved across all problems when using the proposed FCL. Moreover, the improvement in the consistency metric (MSD*, TV) is substantial, highlighting the effectiveness of FCL in stabilizing the AF model. This demonstrates that the proposed FCL and multi-frame model are robust and effective, regardless of the architectural modifications.

D. Quantitative Results for FCL weight

In Tab. B, we provide the full quantitative results of D1 problem for different values of weight parameters λ_{MSD} and λ_{KLD}. This corresponds to the graphs in Fig. 6 of the main paper, for $\lambda_{\text{MSD}} = 4$ (3rd column) and $\lambda_{\text{KLD}} = 4$ (5th row). We can observe that the overall accuracy metrics (MAE, RMSE) maintain a satisfactory level for a certain range of λ_{MSD} and λ_{KLD}. We selected $\lambda_{\text{MSD}} = 4$ and $\lambda_{\text{KLD}} = 4$ for AFPE+FCL, which enhances the MSD* without sacrificing the accuracy metric MAE. For all other settings (D2 $\sim$ D5, I1 $\sim$ I5), we selected different λ_{MSD} and λ_{KLD} for each setting following the same process.

E. Effects of Mini-batch Configuration

During the training phase, we randomly selected 4 distinct focal indices for each of the 32 scenes, creating a mini-batch of size 128. While keeping the batch size fixed, the number of different focal indices, denoted as F, can potentially influence the overall performance; when a smaller F is employed, the model can capture more diverse scenes within each batch. This could potentially enhance the generalization performance and enable robust learning of the features. Conversely, using a larger F enables the model to grasp more diverse views of the same scene. This could lead to a more comprehensive understanding of intra-scene geometric information. To investigate this, we conducted a series of experiments aimed at evaluating the effect of the parameter F, and the results are presented in Tab. C.

Note that the baseline method AFPE [4] in the reference article used a batch size of 128 with all different scenes ($F = 1$). The values from the original reference is marked with a †, and we can observe that our reproduced model for D1 shows similar results. We first conducted experiments for the reproduced version of AFPE with $F = 1, 2, 4, 8$ (rows 2 $\sim$ 5) while keeping the batch size equal to 128, hence resulting in 128, 64, 32, 16 different scenes in a mini-batch, respectively. The results indicate that $F = 4$ provides the highest accuracy metric (MAE), suggesting that $F = 4$ is the optimal choice for capturing diverse scene and intra-scene information. Note that the consistency metric MSD* is not affected by F, since the FCL is not utilized yet.

We then experimented with the AFPE model utilizing FCL, adjusting $F = 2, 4, 8$ (rows 6 $\sim$ 8). For $F = 4$, we used parameters $\lambda_{\text{MSD}} = 4$ and $\lambda_{\text{KLD}} = 4$. When F was changed, we adjusted the parameters to maintain similar weights. This is because KLD is computed by summing over all possible pairs for each scene. Specifically, $F = 2$ has $\binom{2}{2} = 1$ pair, $F = 4$ has $\binom{4}{2} = 6$ pairs, and $F = 8$ has $\binom{8}{2} = 28$ pairs. Therefore, the corresponding λ_{KLD} values are $\lambda_{\text{KLD}} = 4 * 6$ for $F = 2$ and $\lambda_{\text{KLD}} = 4 * 6/28 \fallingdotseq 0.857$ for $F = 8$. Note that λ_{MSD} is computed as the standard deviation of sets, hence the same weight $\lambda_{\text{MSD}} = 4$ is used for all F.

The results show that $F = 4$ is also the optimal choice for both AF accuracy and focal consistency, as it demonstrates the best performance in terms of all metrics (MAE, RMSE, MSD*, and TV). Note that AFPE+FCL outperforms the corresponding baseline AFPE, especially in terms of consistency. This further proves the effectiveness of our FCL, regardless of changes in hyperparameter F.

F. Corner Case Analysis

We provide qualitative examples for the challenging corner cases that the existing baseline typically fails to handle. We show the input image, baseline AFPE output, AFPE+FCL (D1), AFPE+FCL (D5), and the ground truth image, in order.

Figure C shows the example scenes with multiple depths. We can observe that these scenes either have two (or more) objects or a distinct foreground and background. Although the baseline AFPE usually finds the correct focus, such distinct regions can make the model confused and lead to the predictions oscillating between the foreground focus and the background focus. On the other hand, our FCL and the multi-frame results can clearly give more stable outputs.

Figure D shows the dark and noisy scenes. These challenging examples usually occur in low-light environments, and the low signal-to-noise ratio of the input image is one of the most difficult corner case for an AF model to tackle. AFPE+FCL models, while not perfect, demonstrate robust predictions for these dark and noisy scenes, compared with the original baseline.

Figure E shows the examples with nearby objects. As these objects are close to the camera, taking an exact macro shot can be challenging. The baseline AFPE model tends to produce wide-angle shots. On the other hand, by leveraging intra-scene geometric information, our FCL methods can precisely capture the nearby objects.

Figure F depicts the saturated or textureless scenes, which can occur in bright conditions or when photographing smooth objects. Measuring sharpness in these cases is difficult, making it hard for the baseline model to predict focus. However, our FCL methods can effectively capture the subtle cues in these challenging scenes.

λ	$\lambda_{MSD}=0$		$\lambda_{MSD}=2$		$\lambda_{MSD}=4$		$\lambda_{MSD}=6$		$\lambda_{MSD}=8$		$\lambda_{MSD}=10$	
	MAE	MSD*	MAE	MSD*	MAE	MSD*	MAE	MSD*	MAE	MSD*	MAE	MSD*
$\lambda_{KLD}=0$	1.760	1.338	1.764	1.279	1.762	1.250	1.739	1.175	1.776	1.197	1.761	1.147
$\lambda_{KLD}=1$	1.695	1.220	1.740	1.188	1.756	1.162	1.771	1.143	1.791	1.126	1.748	1.105
$\lambda_{KLD}=2$	1.712	1.171	1.730	1.148	1.752	1.129	1.774	1.109	1.773	1.089	1.763	1.063
$\lambda_{KLD}=3$	1.722	1.143	1.736	1.124	1.729	1.091	1.764	1.082	1.787	1.064	1.762	1.044
$\lambda_{KLD}=4$	1.736	1.130	1.746	1.107	1.735	1.070	1.750	1.058	1.810	1.054	1.786	1.028
$\lambda_{KLD}=5$	1.729	1.102	1.743	1.078	1.760	1.058	1.774	1.043	1.789	1.029	1.801	1.020

Table B. Quantitative results of D1 setting for different values of λ_{MSD} and λ_{KLD} using AFPE [3] baseline and proposed FCL. The top two methods for each metric are highlighted in red and orange. We can observe that FCL notably improves the consistency metric MSD* without sacrificing the accuracy metric.

Algorithm	Type	F	λ_{MSD}	λ_{KLD}	MAE	RMSE	MSD*	TV
AFPE† [3]		1	-	-	1.803	2.826	-	-
AFPE		1	0	0	1.824	2.821	1.347	0.859
AFPE	D1	2	0	0	1.833	2.804	1.362	0.880
AFPE		4	0	0	1.760	2.855	1.338	0.895
AFPE		8	0	0	1.851	2.919	1.342	0.886
AFPE+FCL		2	4	24	1.782	2.799	1.121	0.723
AFPE+FCL	D1	4	4	4	1.735	2.744	1.070	0.691
AFPE+FCL		8	4	0.857	1.908	2.983	1.138	0.741

Table C. Quantitative comparison for different number of focal indices F using AFPE [4] baseline, and the proposed FCL. The top method for each metric is highlighted in red. For both the AFPE and AFPE+FCL models, $F=4$ is the optimal choice. For each F, the model with FCL outperforms the corresponding baseline AFPE. A † indicates that values are from the reference article.

G. Additional Qualitative Results

In this section, we provide additional qualitative results demonstrating the effectiveness and reliability of our proposed FCL and multi-frame settings.

Figures G and H show qualitative comparisons of our proposed FCL with the baseline AFPE [4], for the D1 (single-slice, dual-pixel) and I1 (single-slice, single-channel) settings, respectively. Figures I and J show qualitative comparisons of our proposed FCL with the baseline L2A [9], for the D1 and I1 settings, respectively. As the left-most graphs illustrate, the output focal indices of the baseline are not particularly consistent for the given input focal slices. We can also observe that if the prediction of baseline at the GT focal index is wrong, the baseline model suffers from focus hunting problems. On the other hand, by using our proposed FCL, the output focal indices become significantly more consistent, almost perfectly solving the focus hunting problems. The visualized patches also demonstrate the effectiveness of FCL for AF.

Figures K and L show additional qualitative comparisons for AFPE+FCL with different multi-frame settings (D1, D3, D5) and (I1, I3, I5), respectively. Figures M and N show additional qualitative comparisons for L2A+FCL with different multi-frame settings (D1, D3, D5) and (I1, I3, I5), respectively. Notably, we demonstrate more challenging scenarios when our D1 model fails to predict consistent predictions even if it is trained with the FCL. For such difficult cases, including scenes with multiple depths or severe noise, using our multi-frame settings could substantially enhance the prediction consistency, resulting in stable performance.

References

[40] Diederick P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 1

[41] Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 1991. 2

[42] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *ICLR*, 2017. 1

[43] Frank Nielsen. Jeffreys centroids: A closed-form expression for positive histograms and a guaranteed tight approximation for frequency histograms. *IEEE Signal Processing Letters*, 2013. 2

[44] Frank Nielsen. On the jensen–shannon symmetrization of distances relying on abstract means. *Entropy*, 2019. 2

[45] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 1

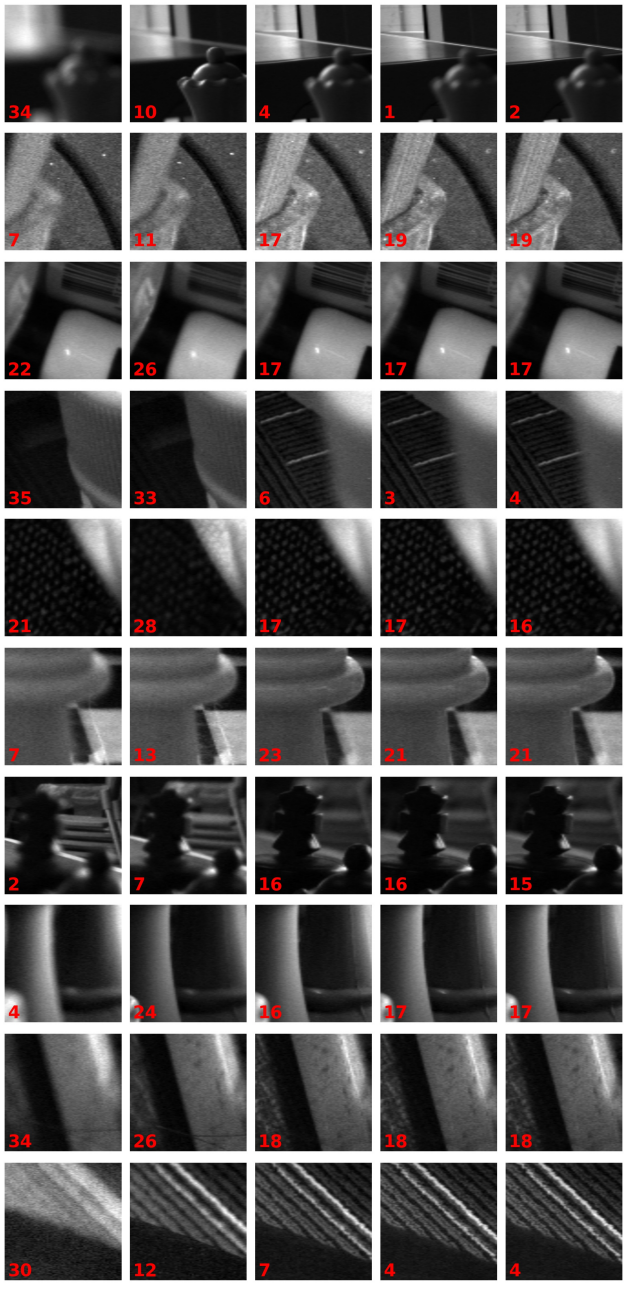

Figure C. Scenes with multiple depths

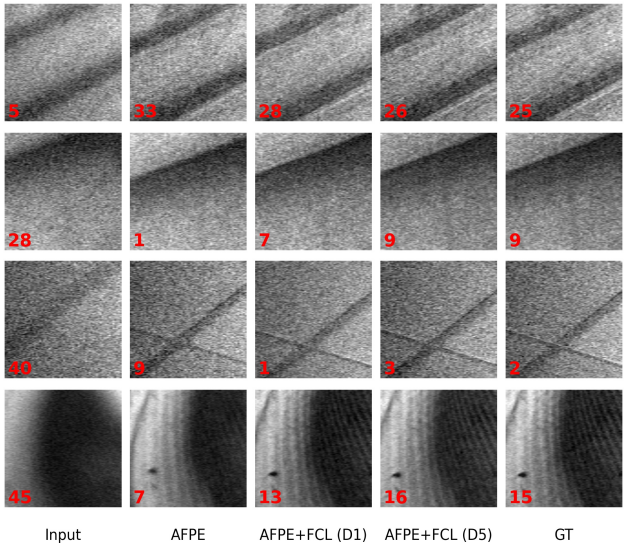

Figure D. Dark and Noisy scenes

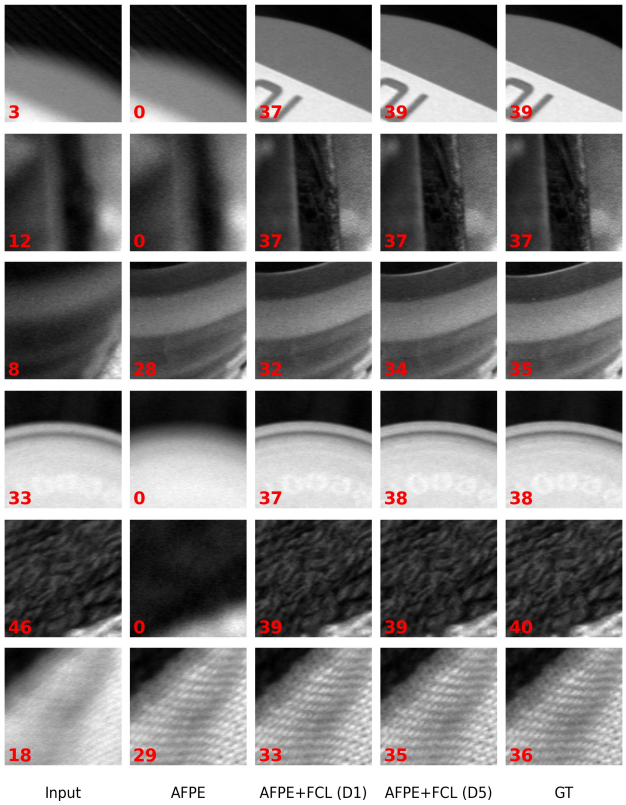

Figure E. Near scenes

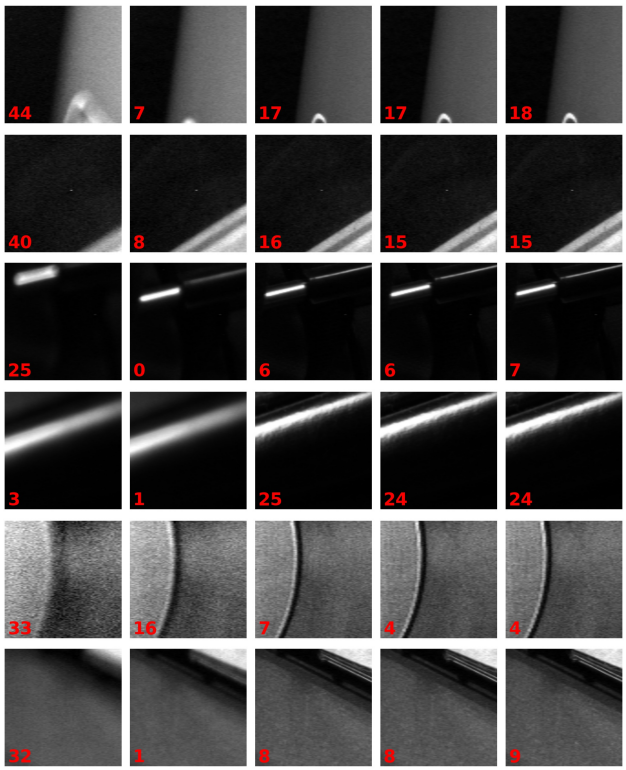

Figure F. Saturated or Textureless scenes

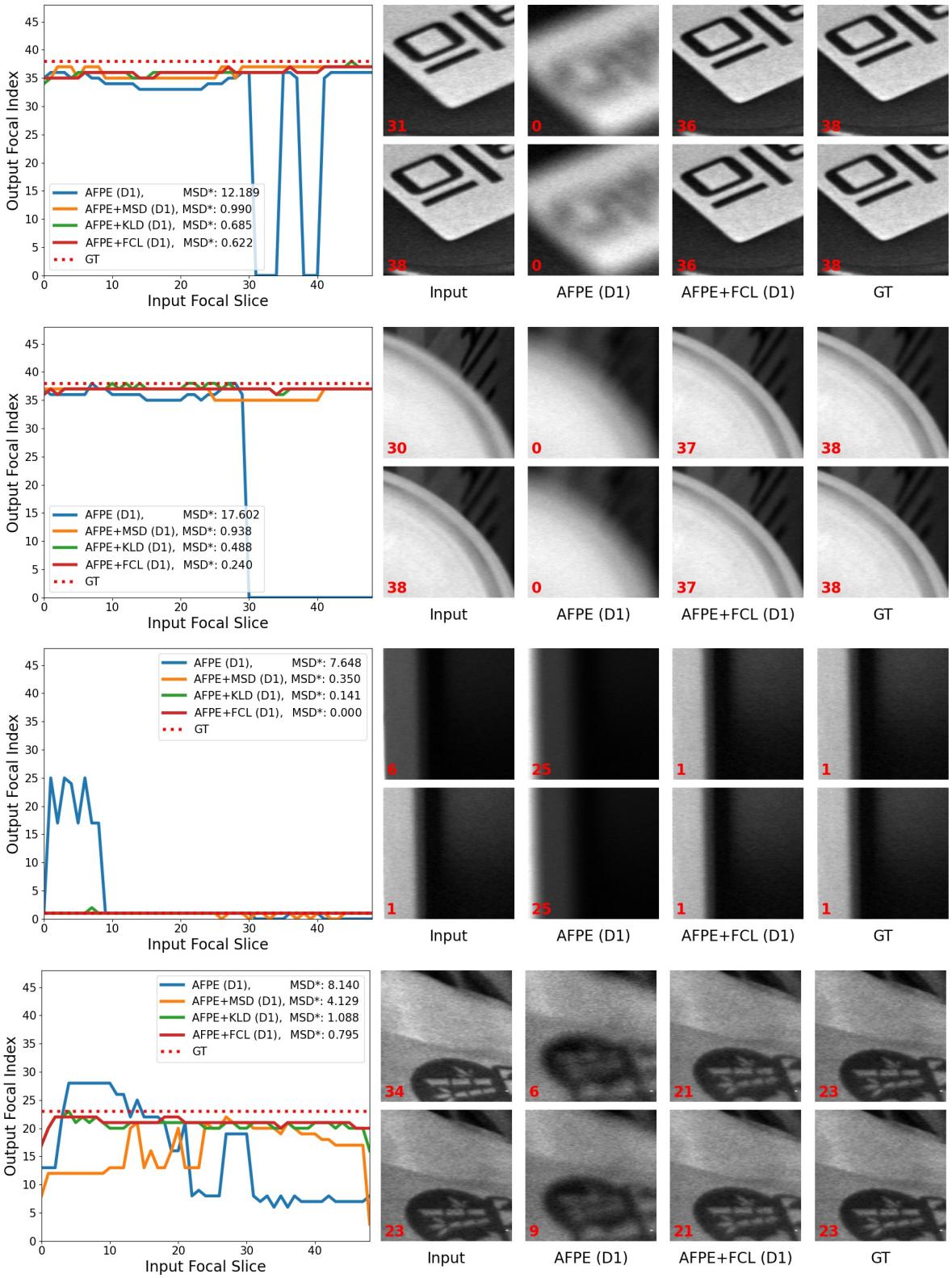

Figure G. Qualitative comparison our proposed FCL methods with the AFPE [4] baseline for D1 (single-slice, dual-pixel) setting. The leftmost graph illustrates the output focal index predictions for each input focal slice. AFPE in these cases encounter focus hunting, whereas FCL methods show more consistent predictions. The red numbers in each patch indicate the focal index.

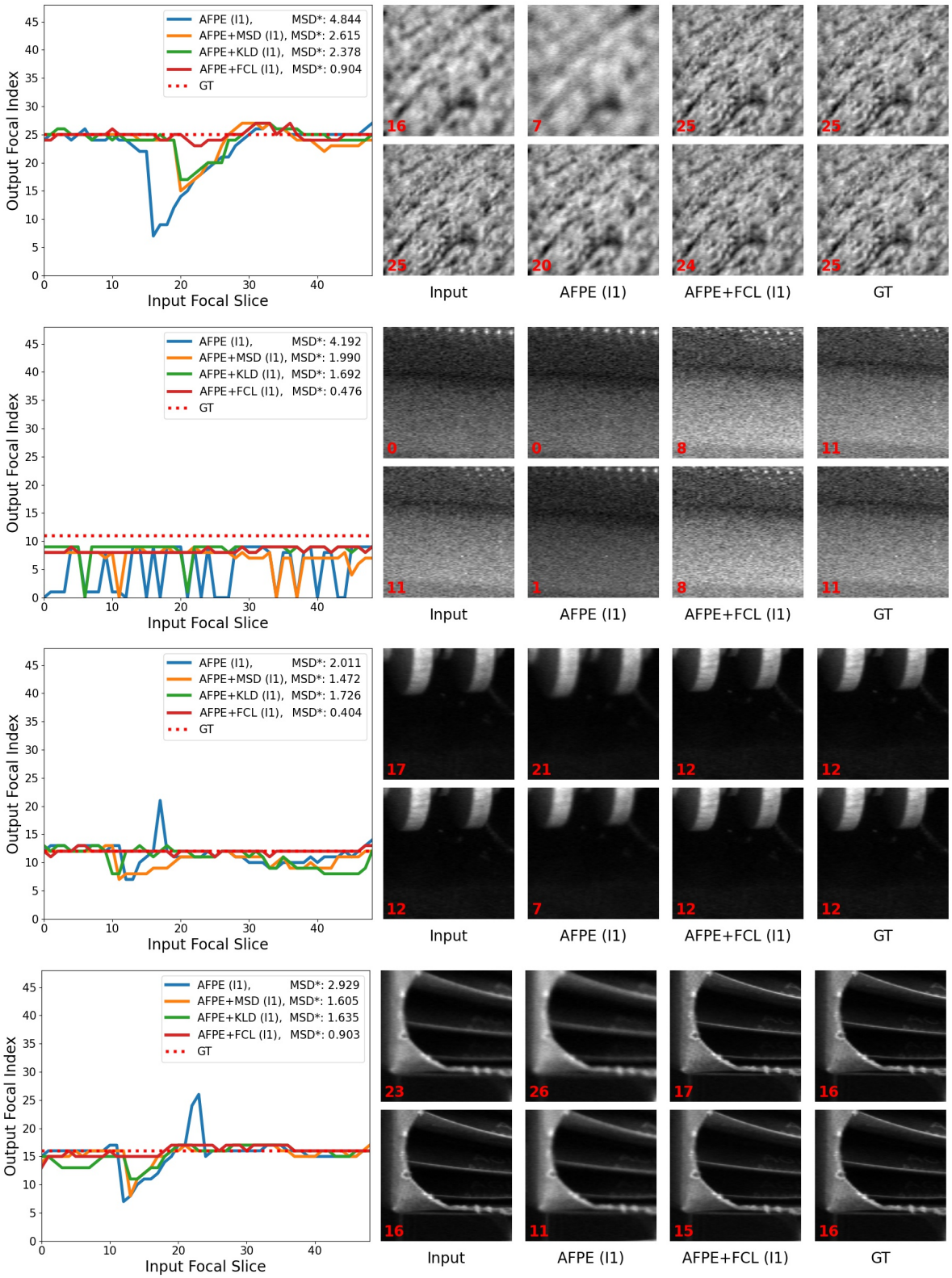

Figure H. Qualitative comparison our proposed FCL methods with the AFPE [4] baseline for I1 (single-slice, single-channel) setting. The leftmost graph illustrates the output focal index predictions for each input focal slice. AFPE in these cases encounter focus hunting, whereas FCL methods show more consistent predictions. The red numbers in each patch indicate the focal index.

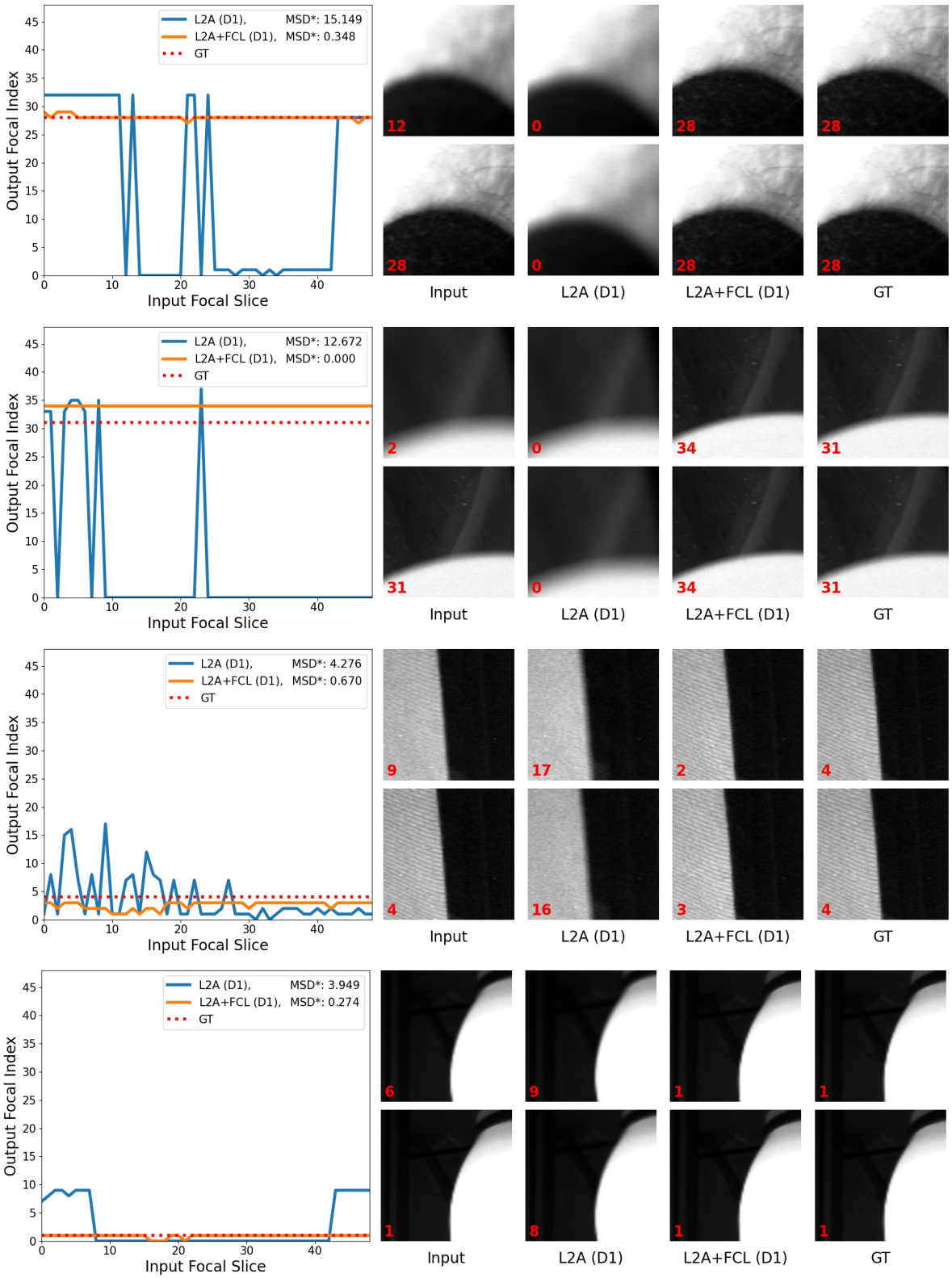

Figure I. Qualitative comparison our proposed FCL methods with the L2A [9] baseline for D1 (single-slice, dual-pixel) setting. The leftmost graph illustrates the output focal index predictions for each input focal slice. L2A in these cases encounter focus hunting, whereas FCL methods show more consistent predictions. The red numbers in each patch indicate the focal index.

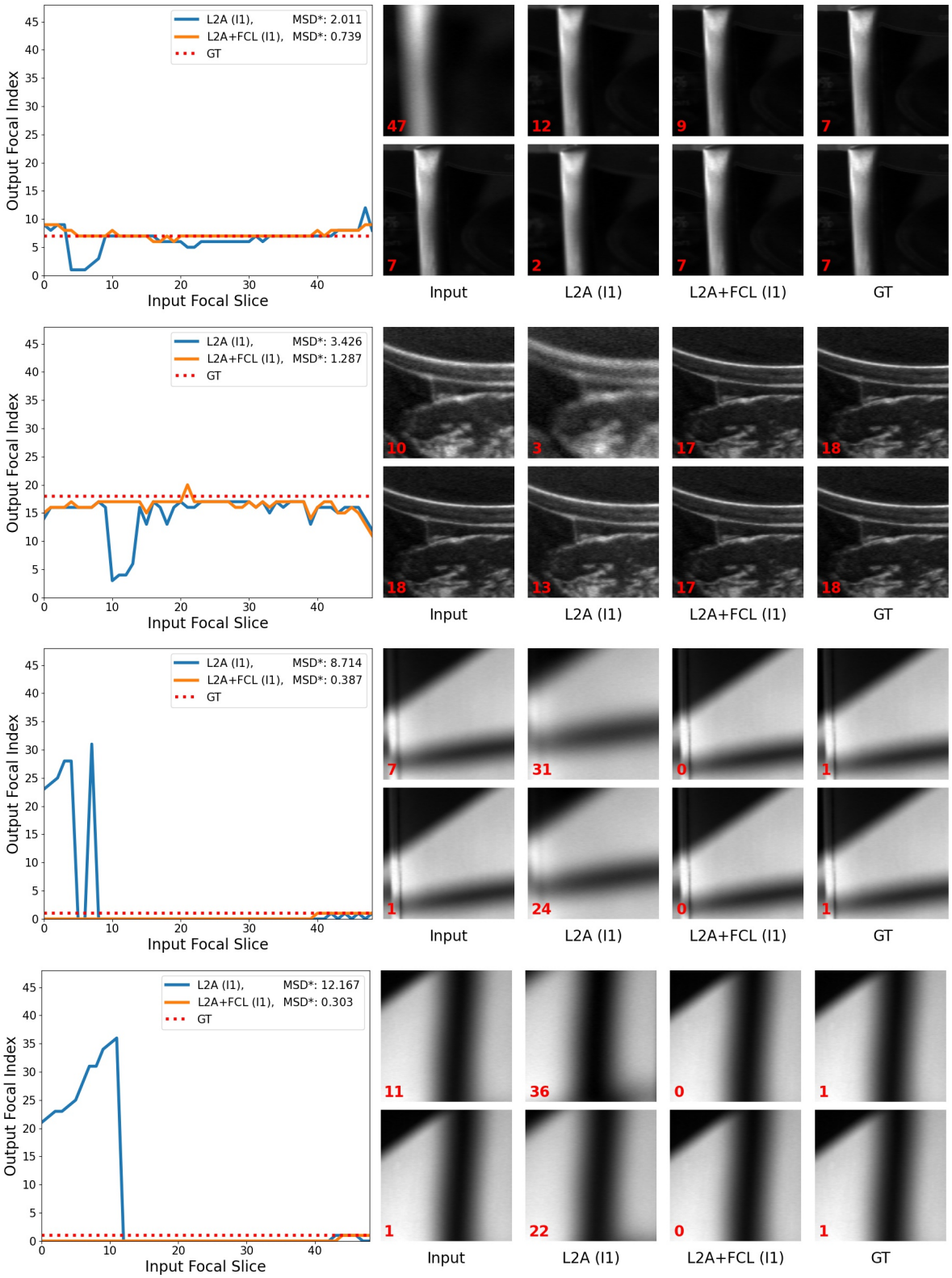

Figure J. Qualitative comparison our proposed FCL methods with the L2A [9] baseline for I1 (single-slice, single-channel) setting. The leftmost graph illustrates the output focal index predictions for each input focal slice. L2A in these cases encounter focus hunting, whereas FCL methods show more consistent predictions. The red numbers in each patch indicate the focal index.

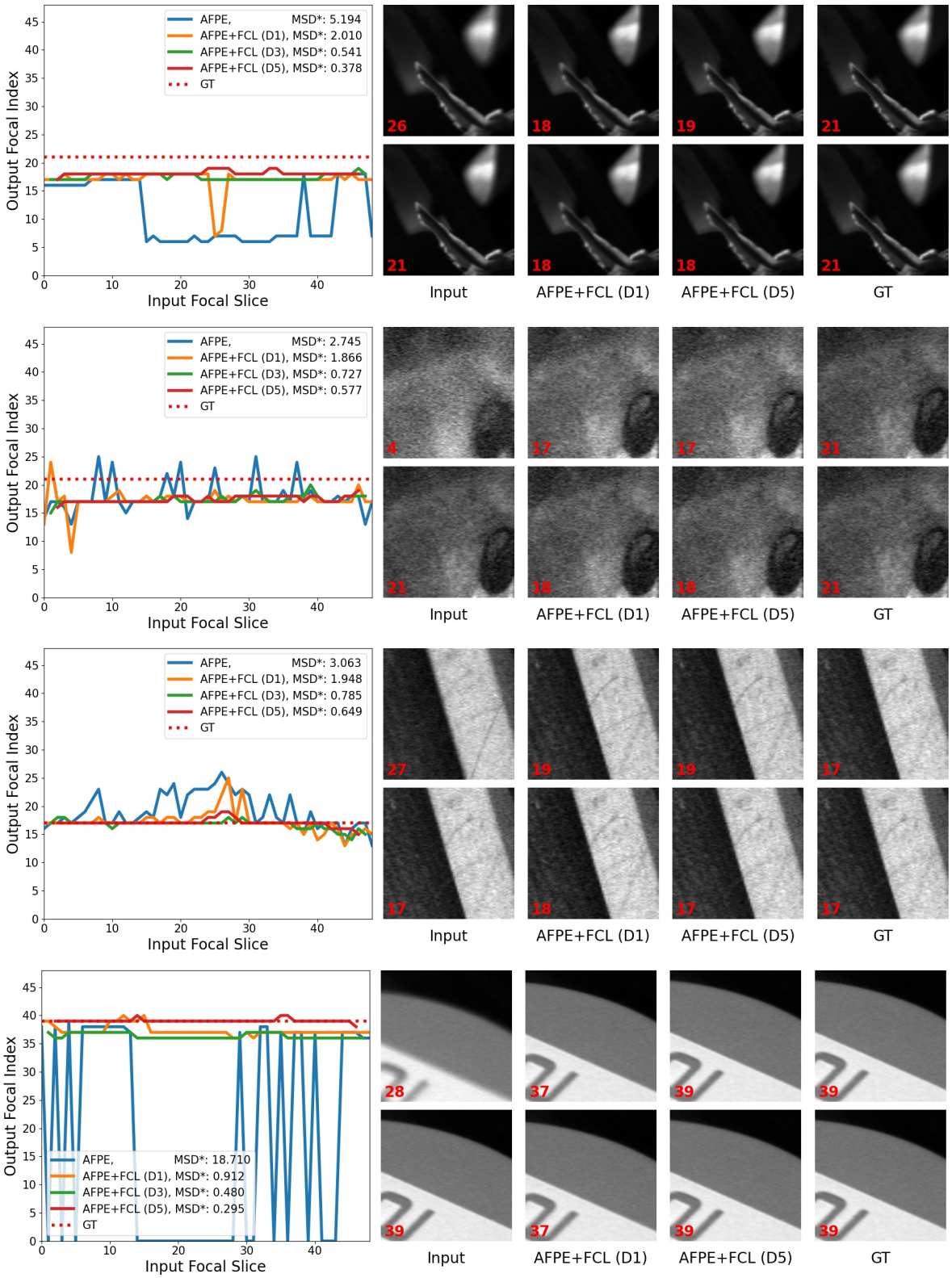

Figure K. Qualitative comparison of results from AFPE+FCL with different multi-frame settings: D1, D3, and D5. The left graph shows that, while our FCL enhances prediction consistency, using only a single slice input (D1) may sometimes produce false answers for challenging scenes. Our multi-frame models (D3, D5) can alleviate this issue and stably converge near the GT, as shown in the right.

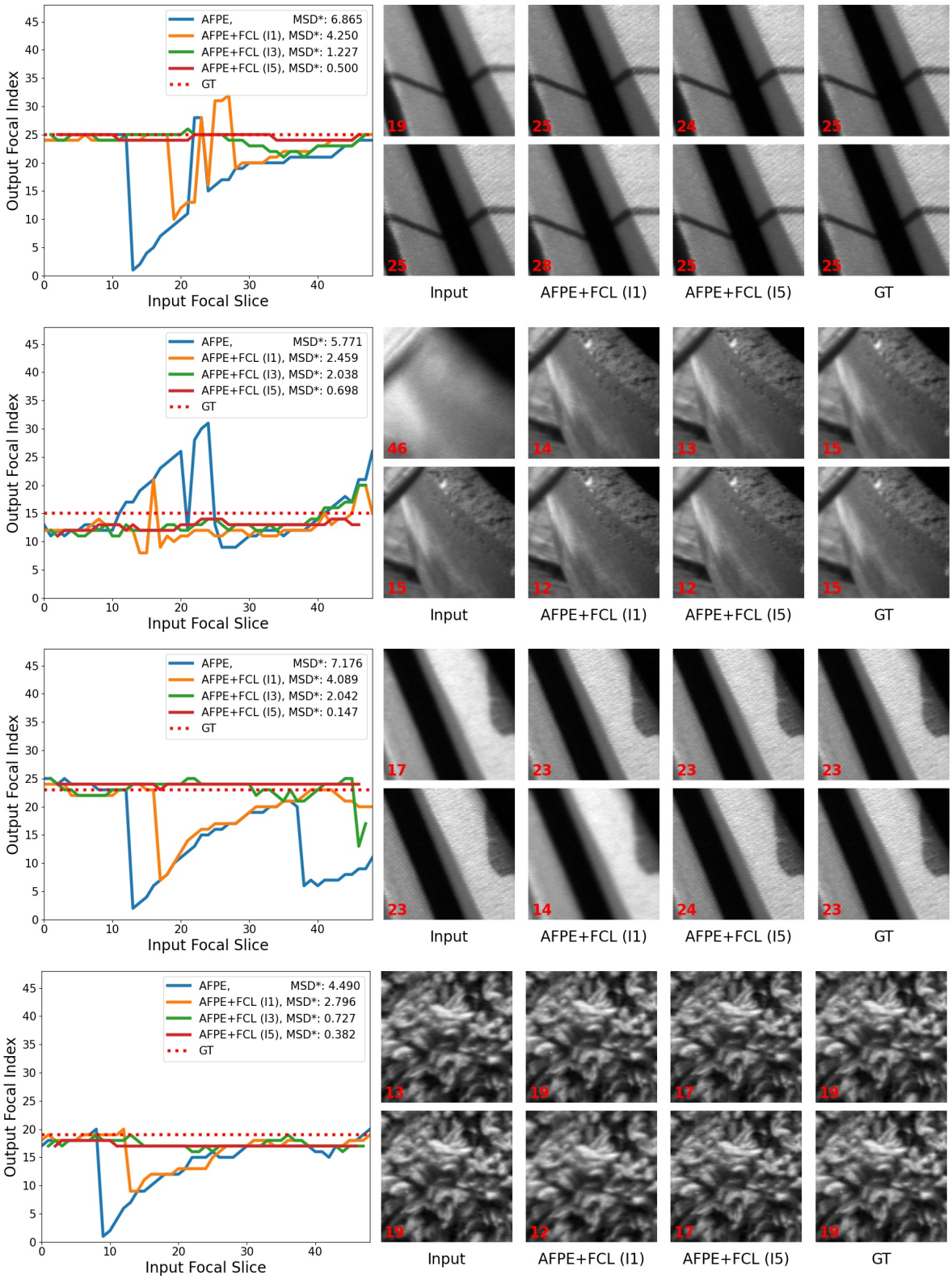

Figure L. Qualitative comparison of results from AFPE+FCL with different multi-frame settings: I1, I3, and I5. The left graph shows that, while our FCL enhances prediction consistency, using only a single slice input (I1) may sometimes produce false answers for challenging scenes. Our multi-frame models (I3, I5) can alleviate this issue and stably converge near the GT, as shown in the right.

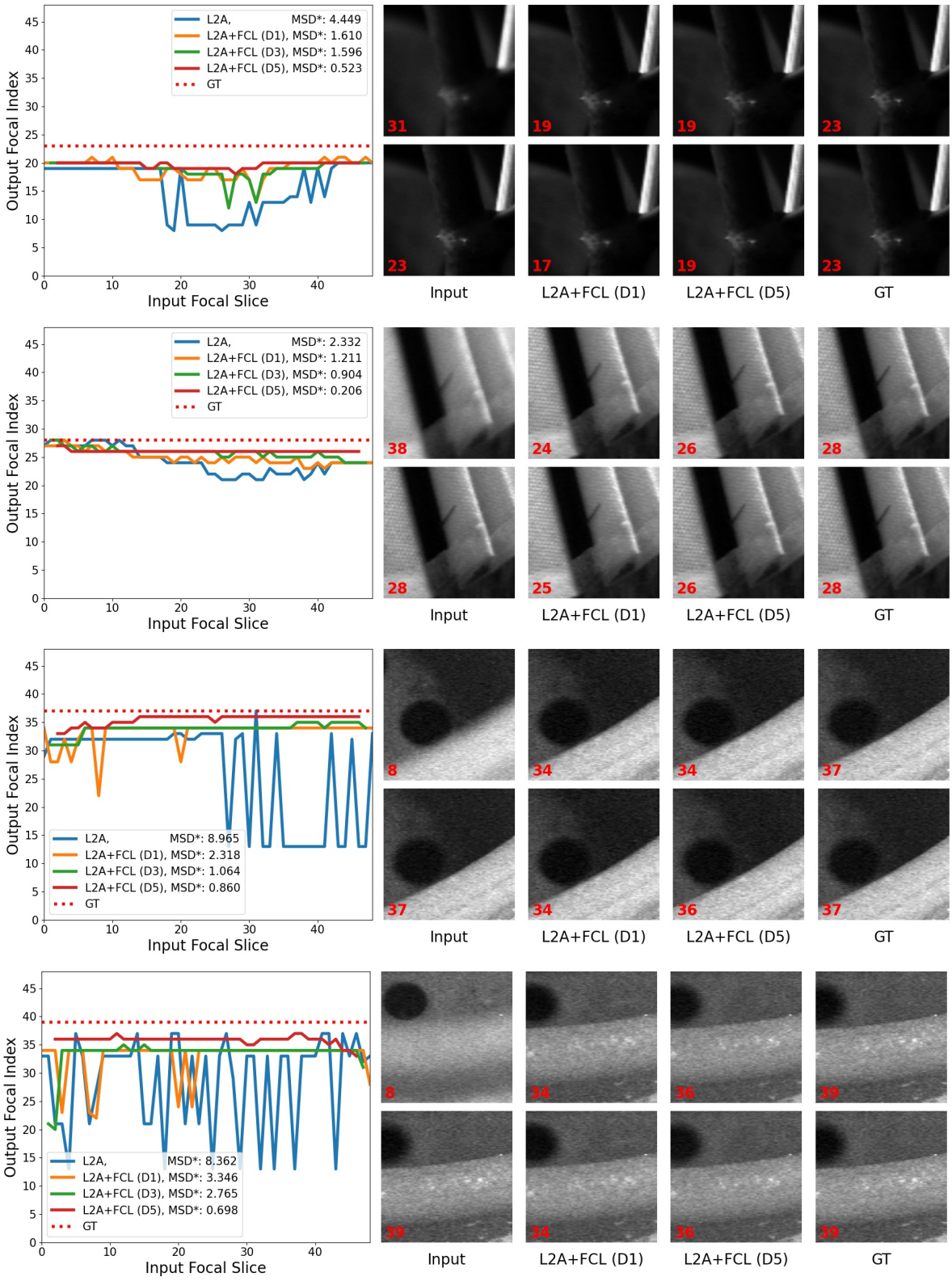

Figure M. Qualitative comparison of results from L2A+FCL with different multi-frame settings: D1, D3, and D5. The left graph shows that, while our FCL enhances prediction consistency, using only a single slice input (D1) may sometimes produce false answers for challenging scenes. Our multi-frame models (D3, D5) can alleviate this issue and stably converge near the GT, as shown in the right.

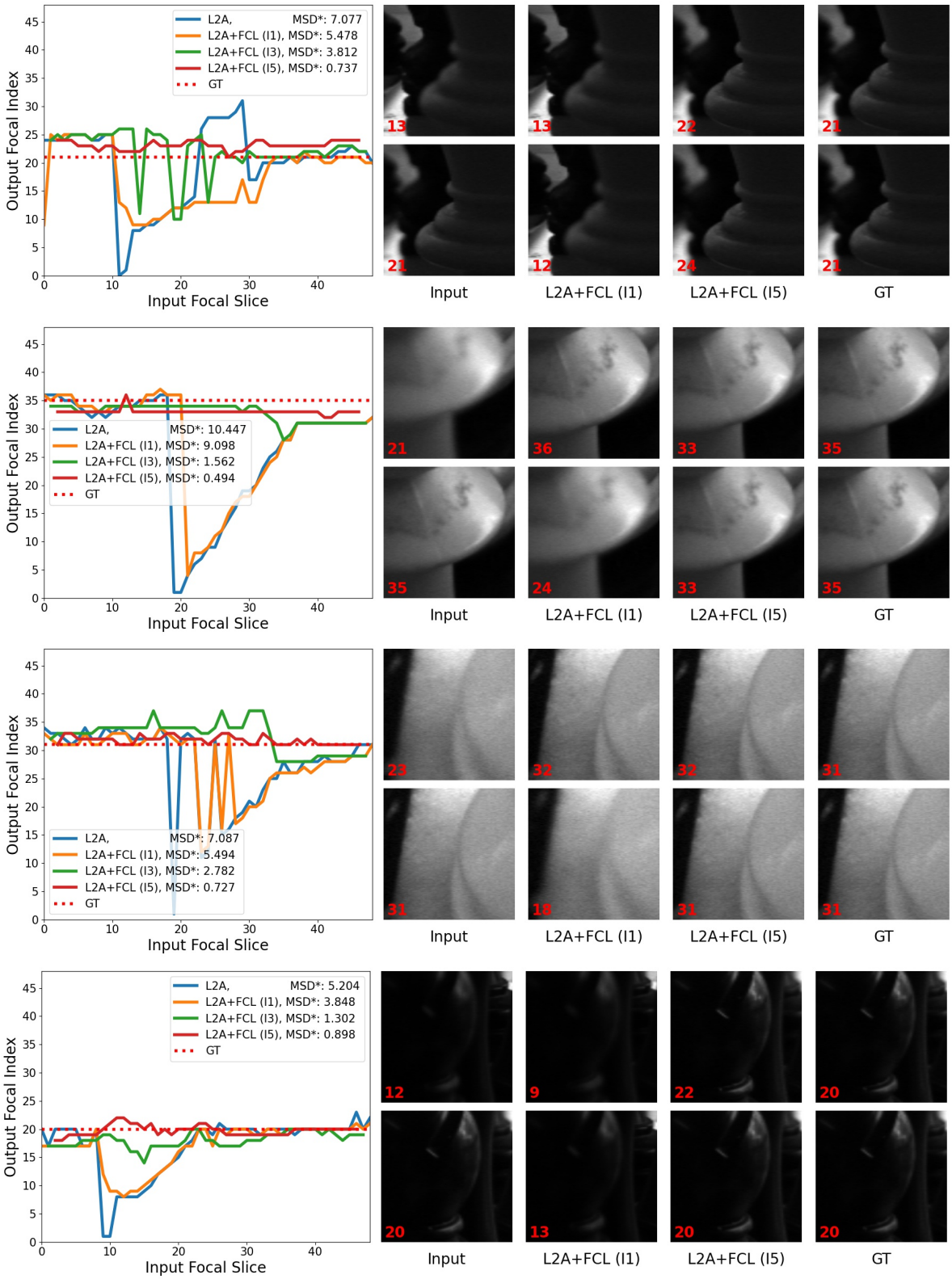

Figure N. Qualitative comparison of results from L2A+FCL with different multi-frame settings: I1, I3, and I5. The left graph shows that, while our FCL enhances prediction consistency, using only a single slice input (I1) may sometimes produce false answers for challenging scenes. Our multi-frame models (I3, I5) can alleviate this issue and stably converge near the GT, as shown in the right.