

# Enabling High-Quality In-the-Wild Imaging from Severely Aberrated Metalens Bursts

## Supplementary Material

We organize the supplementary document into 3 sections namely, the camera design, additional implementation details and additional results and figures.

### 1. Camera design

#### 1.1. Metalens Optimization

Metalenses are flat optical devices that focus light using arrays of nanoscale antennas rather than curved surfaces. Each antenna imparts a precise phase delay to create the desired optical function through controlled interference. Nanostructures provide discrete phase values, performance varies with wavelength and incident angle, and manufacturing imposes geometric limitations. These factors prevent direct implementation of ideal phase functions which ordinary refractive lenses such as convex lenses follow. Modern approaches use differentiable wave propagation models for end-to-end optimization. Light propagation from metalens to the image plane can be simulated using differentiable operations such as Fast Fourier Transforms. Phase parameters ( $\phi(r)$ ) are iteratively optimized using gradient descent to minimize optical performance metrics such as focal spot size or aberration levels. This computational approach helps us design complex optical functions that exceed traditional lens capabilities while addressing practical manufacturing and performance constraints.

#### 1.2. Metalens Fabrication and Assembly

We fabricated a 1 cm large aperture metaoptic using a nanofabrication approach, facilitated in an ISO Class 5-7 clean room environment. First, quartz wafer (purchases from University Wafer), were cleaned in subsequent ultrasonication baths of Acetone, IPA, and DI water, then exposed to a short oxygen descum in a Barrel Etcher. We then deposited an ~800 nm thick Silicon Nitride film using plasma enhanced chemical vapor deposition in an SPTS chamber. Afterwards the wafer was diced into 1.5 cm square pieces and again cleaned using an ultrasonication bath and barrel etch steps (as before). We then applied a positive resist (ZEP 520A) with a thickness of 400nm. To mitigate charging during the patterning, we also applied a conductive polymer layer (Dis-Charge H2O). We then patterned the resist using a 8 nA, 100 keV electron beam (JEOL JBX6300FS) at a dose of about  $300 \mu C/cm^2$ . After electron beam lithography, we removed the conductive polymer layer using a short IPA bath and developed the resist at room temperature in Amyl Acetate for 2 min. Subsequently, the sample was again descummed in a short barrel etch step and a layer of about 75 nm alumina was evaporated onto the sample. The resist was then lifted off overnight in an NMP bath at on a hot plate. Subsequently, the SiN layer was etched using a fluorine based etch mixture

in an inductively coupled reactive ion etcher (Oxford PlasmaLab System 100). Finally, the chip was integrated in a 3D printed holder and mounted with the sensor. Scanning Electron Microscope Images after fabrication can be seen in Fig. 1. The resulting transmissive metalens is integrated with an Allied Vision 1800 U-510 CMOS sensor and paired with a Jetson Nano Orin board for handheld operation.

### 2. Additional implementation details

**Reference frame selector** Please refer to the reference selection algorithm pseudo-code in Algorithm 1.

---

#### Algorithm 1: Reference Frame Selection Algorithm

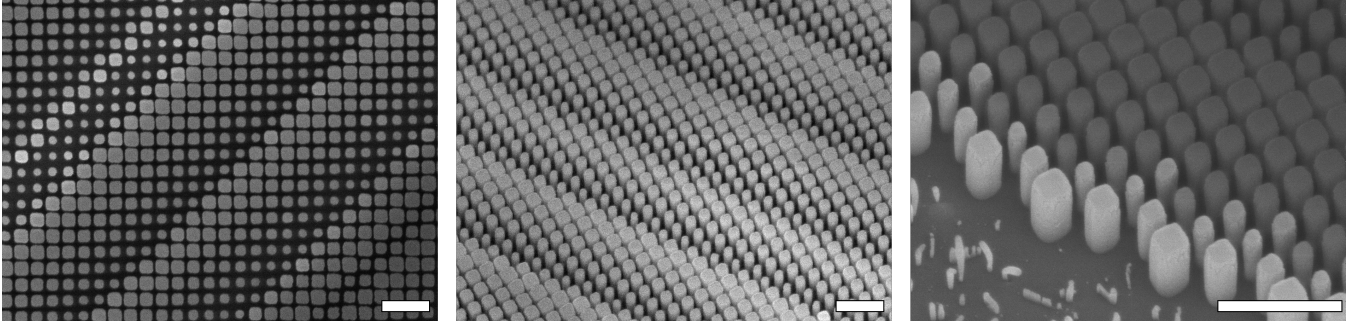
---

**Input:** Reference exposure time  $\leftarrow$  ISP, burst size  $N$   
**Output:** Reference frame  $I_{ref}$

```
1  $t_{ref} \leftarrow \text{AutoExposure}();$ 
2 Initialize burst set  $\mathcal{B} \leftarrow \{\}$ ;
3 for  $i = 1$  to  $N$  do
4    $g_i \leftarrow \text{SampleDigitalGain}();$ 
5    $I_i \leftarrow \text{Capture}(t_{ref}, g_i);$ 
6    $\mathcal{B} \leftarrow \mathcal{B} \cup \{I_i\};$ 
7 end
8  $s_{max} \leftarrow 0, I_{ref} \leftarrow \text{null};$ 
9 for each  $I_i \in \mathcal{B}$  do
10   $G_i \leftarrow \text{ExtractGreenChannel}(I_i);$ 
11   $s_i \leftarrow \text{GradientSharpness}(G_i);$ 
12  if  $s_i > s_{max}$  then
13     $s_{max} \leftarrow s_i, I_{ref} \leftarrow I_i;$ 
14  end
15 end
16 return  $I_{ref};$ 
```

---

**Multi-scale burst alignment algorithm** Similar to HDR+, our method also adopts a pyramid-based iterative refinement strategy for displacement estimation. As shown in Fig. 2, however, under the same framework our multi-scale displacement fusion produces motion fields that are notably more coherent and stable across pyramid levels. The displacement map indicates a pronounced shift in the upper-left and only subtle motion in the bottom-right (with the red object nearly static), all of which are faithfully captured by our flow visualization. In contrast, HDR+ [4] provides reasonable estimates in the upper-right and red object regions but substantially overestimates motion in the bottom-right, yielding incorrect magnitudes and directions. Our method, on the other hand, preserves smooth global motion patterns and structural consistency, leading to more reliable alignment. Moreover, in the bottom-left region our flow reveals a clear downward displacement, not evident in the displacement map due to the repetitive vertical texture of the floor



**Figure 1.** Meta-atoms visible under an electron microscope

---

**Algorithm 2:** Multi-scale Feature Extraction and Homography Estimation

---

**Input:** Burst images  $\{I^i\}$ , deconvolution parameter  $\rho$ , weights  $\{\lambda_k\}$ , pyramid levels  $K$

**Output:** Homographies  $\{H^k\}$  across scales

```

1 for  $k = 2$  to  $K$  do
2   Compute  $\tilde{I}_k^b$  via Tikhonov deconvolution:
3   
$$\tilde{I}_k^b = \mathcal{F}^{-1} \left\{ \frac{\bar{\rho}^*}{|\bar{\rho}^2| + \lambda} \mathcal{F}(\tilde{I}_{k-1}^b \downarrow_2) \right\}$$

   where,  $\bar{\rho}$  is the optical transfer function or the Discrete
   Fourier Transform of the PSF ( $\rho$ ) and  $\mathcal{F}$  represents the
   Fast Fourier operation
4 end
5 for  $k = K$  to 1 do
6   Extract feature correspondences and estimate
   homography:
   
$$H^k = \text{get\_pts}(\tilde{I}_k^i, \tilde{I}_k^{i+1})$$

7 end

```

---

tiles, demonstrating the robustness of our approach for burst imaging with complex textures and weak motions. The complete pseudo-code is detailed in Algorithms 2 and 3.

**Adaptive pixel correction unit (APCU)** The adaptive pixel correction unit is designed similar to the weighted burst fusion block using a series of residual blocks followed by a sigmoid weighting layer,

$$I_{\text{init}} = I_{\text{fused}} \cdot \text{sigmoid}(\{\text{ResBlocks}^n(I_{\text{fused}})\}) \quad (1)$$

**Attention Fusion Block** The skip connections from the SFT layer, containing features for all burst frames, are fused into the attention fusion block (AFB). Fusion between the reference and other burst frames is modeled using a series of channel-wise cross-attentions:

$$\begin{aligned}
\text{AFB}(\{SFT\}_i) &= \text{crossatten}(CA^n, \{SFT\}_i - SFT_{\text{ref}}) \\
CA^n &= \text{crossatten}(CA^{n-1}, \{SFT\}_i - SFT_{\text{ref}}) \\
CA^0 &= SFT_{\text{ref}}
\end{aligned}$$

---

**Algorithm 3:** Multi-scale displacement fusion

---

**Input:** Homographies  $\{H^k\}$ , pyramid levels  $K$

**Output:** Displacement maps  $\{\vec{V}_p\}_i^b$  (final aligned flow fields)

1 Initialize global displacement from coarsest level:

$$\{\vec{V}_p\}_K^i \quad \text{from } H^K$$

```

2 for  $k = K - 1$  to 1 do
3   for each patch  $p$  do
4     Compute local homography  $H_p^k$  and displacement
      $\{\vec{V}_p\}_k^i$ 
5     Fix the weight to a constant value:
     
$$\omega_p = 0.5 \quad (\text{empirically chosen constant})$$

6     Update displacement via weighted aggregation:
     
$$\vec{V}_{pk}^i = (1 - \omega_p) \text{best}(\vec{V}_{pk}^i, \{\vec{V}_{p'k}^i\}_{p' \in \mathcal{N}(p)}) + \omega_p \vec{V}_{pk+1}^i$$

7   end
8 end

```

---

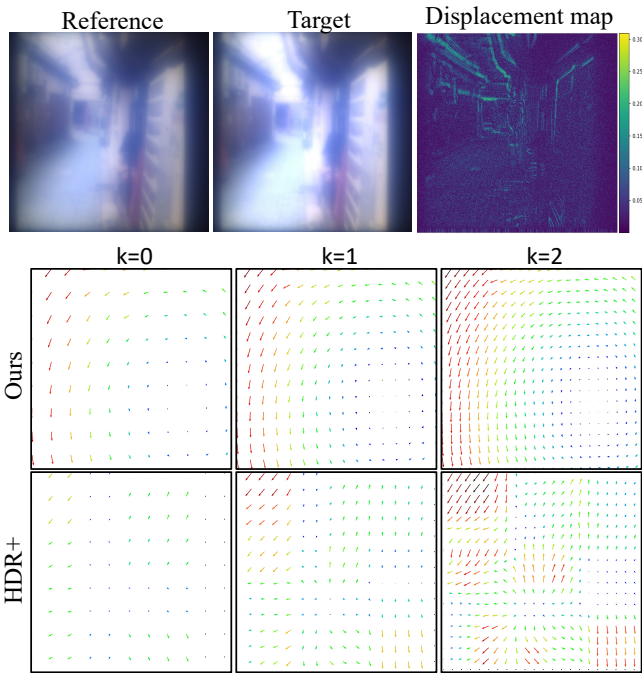
where  $SFT_{\text{ref}}$  is the scale-shift features coming from the skip connections for the reference frame index and cross-attention is performed via multi-head attention (MHA) between the query ( $CA^n$ ) and all the burst frames as the context:

$$\begin{aligned}
\text{crossatten}(Q, \{SFT\}_i) &= \{\text{MHA}(Q^h, K^h, V^h)\}_b \\
Q^h &= f_q^h(Q), K^h = f_k^h(SFT_i), V^h = f_v^h(SFT_i)
\end{aligned}$$

where MHA denotes multi-head attention over heads  $h$ , and  $f_{(\cdot)}^h$  are depthwise separable convolutions. The resulting value  $D$  is then added to  $SFT_{\text{ref}}$  and forwarded to the U-net decoder,

$$D = \text{AFB}(\{SFT_i\}) + SFT_{\text{ref}}.$$

**Unsupervised finetuning** For real-world adaptation, we perform fine-tuning using the saturation loss  $L_{\text{sat}}$  computed on saturation masks. We reduce the learning rate by a factor of 10 from the base model's training rate and freeze all model layers except the fusion weight prediction network. The weight prediction network parameters are updated using an exponentially moving average with decay factor 0.995



**Figure 2.** *Multi-scale displacement fusion.* The first row shows the reference frame, target frame, and their displacement map. The subsequent rows visualize flow fields across pyramid levels ( $k = 0, 1, 2$ ), where arrow direction encodes motion direction and arrow length/color encode displacement magnitude.

across iterations. We also include the original training losses computed over a small subset of synthetic training data during fine-tuning. The saturation loss on real images and original training losses on synthetic images are balanced by a weighting factor determined through cross-validation. Fine-tuning continues until the combined loss converges and no longer decreases.

### 3. Additional results and ablations

#### 3.1. Additional results

We provide evaluation results of our burst restoration framework on OLED dataset in Fig. 5, showing superior image quality and restoration of finer details. In Tab. 2, we show further evaluations of our burst alignment algorithm at different exposure levels and more visual comparisons in Fig. 6. In Fig. 7 we provide visual results for in-the-wild restoration without any manual intervention during the capture process. As shown in the figure, our method can generalize under arbitrary lighting conditions due to our robust training strategy. As an added bonus our in-the-wild adaptation in Sec. 4.4 helps generalize to outdoor HDR scenes as shown in Fig. 8. We analyze exposure fusion maps from the burst fusion module in Fig. 9 and compare against those of Mertens et al. [11]. We also separately show in Fig. 10 the contribution of the restoration module ( $I_{\text{res}}$ ) from the modules before it ( $I_{\text{init}}$ ).

**Table 1.** MetaHDR Dataset Composition

Dataset	Size	#Burst	Scenes	Train %
<i>Burst HDR Datasets</i>				
Burst-HDR+ [4]	3750	2-10	Indoor/outdoor, day/night	80%
Kalantari et al. [7]	222	3	Indoor/outdoor, person, motion	90%
<i>Non-Burst HDR Datasets</i>				
HDM-HDR [3]	423	N/A	Outdoor, Night, extreme bright/dark	90%
Zurich Raw [6]	504	N/A	Outdoor driving	85%
<i>Non-Burst Image Datasets</i>				
Flickr2K	2650	N/A	General scenes	70%
Div2K [1]	800	N/A	General scenes	70%
Liu4K [9]	1600	N/A	Ultra high-res scenes	70%
<i>Real metalens images (for unsupervised finetuning)</i>				
MetaHDR	200	5-20	Indoor/outdoor, dark/bright, direct illumination/shade	100%

#### 3.2. Additional dataset details

**Synthetic OLED dataset** Our synthetic indoor dataset includes aligned ground truth and low-quality images for burst and non-burst datasets (Tab. 1). We simulate artificial hand shake during capture using the pipeline from [2], adding minor misalignment to training bursts relative to reference frames. The restoration network trains on this synthetic burst data.

**Real dataset** Our real outdoor dataset for self-supervised finetuning contains unaligned image pairs captured with a compound optic camera. We include diverse scenes spanning daylight, nighttime, and directly illuminated objects (Tab. 1). Burst images are first aligned using our training-free alignment module, then fused in the restoration network.

#### 3.3. Extended ablations analysis

We provide detailed analysis of our ablation studies to justify key design decisions in our framework.

**Burst Size Analysis.** Ablation results in Table 4 (main paper) shows that increasing burst frames from 3 to 5 provides consistent improvements ( $26.5 \rightarrow 26.9$  dB PSNR,  $0.25 \rightarrow 0.24$  LPIPS). The performance gain plateaus at 5 frames, suggesting this represents the optimal trade-off between information aggregation and motion-induced artifacts.

**Model Architecture Choices.** Channel depth analysis reveals that 20 channels provide the optimal capacity, with fur-

ther increases to 24 channels showing minimal gains (26.5 vs 26.3 dB). Similarly, increasing transformer blocks per Attention-based Fusion Block (AFB) beyond 2 units yields diminishing returns, validating our architectural efficiency. These results demonstrate that our model achieves strong performance without excessive parameterization.

**Efficacy of  $L_{sat}$  in fine-tuning:** We analyze the importance of real-world adaptation in Fig. 3, which improves burst fusion map quality. Initially, due to dynamic range differences between OLED and real-world illumination, the model incorrectly weights contributions from low and high exposed frames, producing poor fusion maps with halo artifacts around bright saturated regions. Our unsupervised fine-tuning with saturation masks corrects these fusion maps, yielding both qualitative and quantitative improvements (NIQE: 7.78  $\rightarrow$  5.43, BRISQUE: 31.04  $\rightarrow$  23.75).

**Impact of alignment errors on restoration quality:** As described in Sec. 3.2 our synthetic dataset involves bursts with small artificially introduced shifts. This enables our feature alignment module to spatially align the bursts using the scale-shift feature transform (SFT) layer. In Fig. 4 we visualize the quality of restoration with burst alignment methods which perform inferior to ours and observe that the image quality degrades only slightly when the input burst is not precisely aligned. Specifically, we pair our restoration network with other methods such as LoFTR [14] and HDR+ [4] and observe the metrics drop slightly when evaluated over our MetaHDR benchmark:

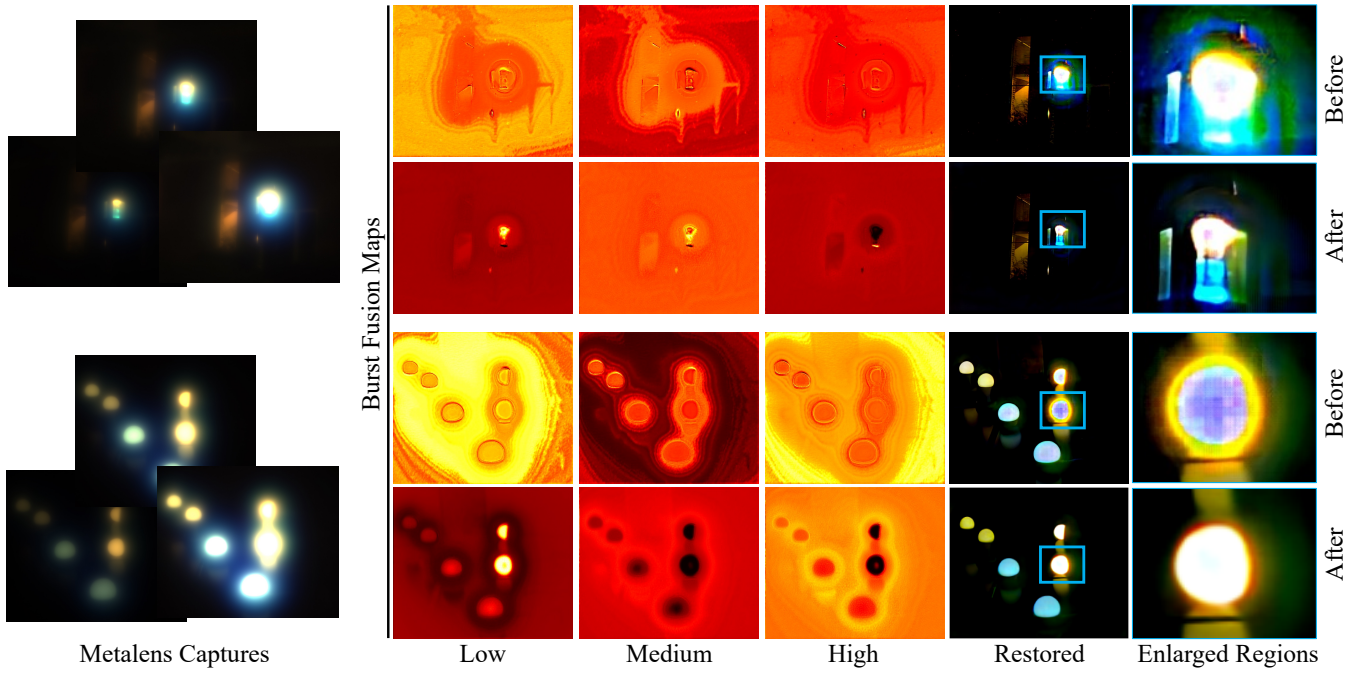
Method	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	NIQE $\downarrow$
LoFTR	26.2	0.74	0.29	5.7
HDR+	25.8	0.69	0.35	6.1
Ours	27.5	0.81	0.23	5.4

**Component Importance Analysis.** The restoration module ( $I_{res}$ ) emerges as the most critical component, with its removal causing the largest performance drop (26.5  $\rightarrow$  19.9 dB). This substantial degradation highlights the importance of dedicated restoration processing beyond simple feature aggregation. The APCU and AFB modules show comparable importance, with individual removal reducing performance to 24.7 and 24.4 dB respectively. When both are disabled simultaneously, performance drops further to 24.2 dB, indicating some complementary effects between these components.

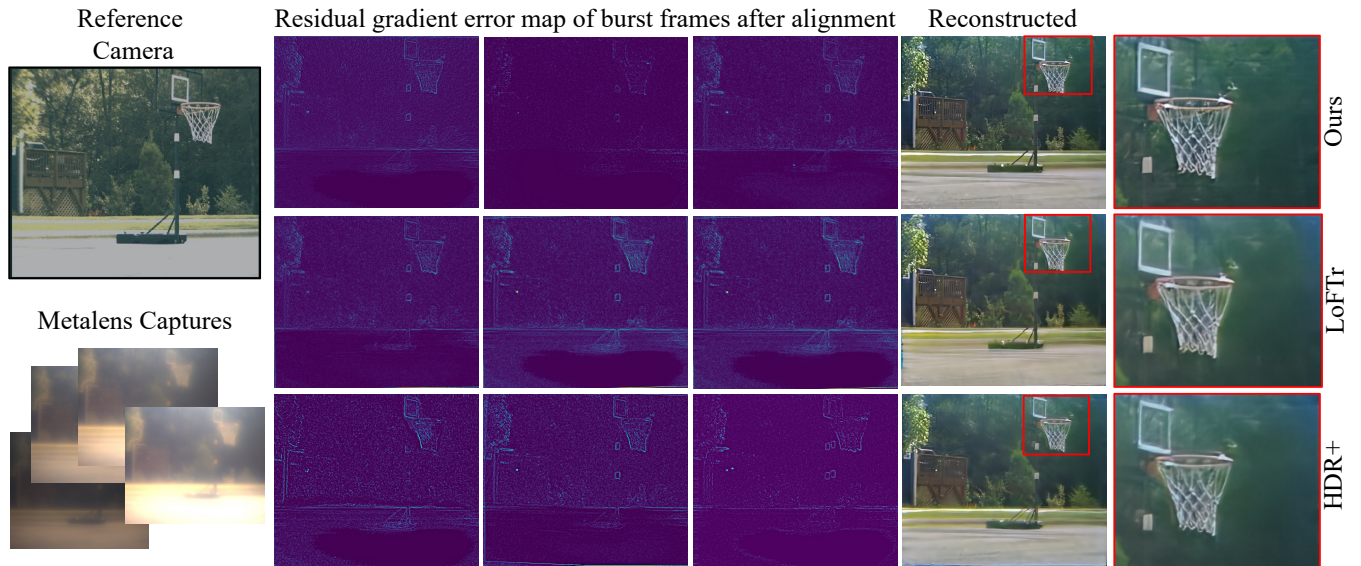
**Failure Case Analysis.** Extended burst sequences can suffer from accumulated alignment errors, particularly in scenes with complex motion patterns or significant camera shake. Our analysis in Figure 13.a (main paper) shows that performance degradation typically occurs when motion exceeds the effective receptive field of our alignment module, suggesting potential areas for future improvement through more robust motion modeling. Secondly, although our average capture time per frame is relatively short (20 - 50 ms), fast moving regions such as hands or fingers may appear blurred due to incorrect fusion as visible in Figure 10 (main paper). On the other hand, such blurs might also act as useful indicators of scene motion presence up to a certain threshold.

## References

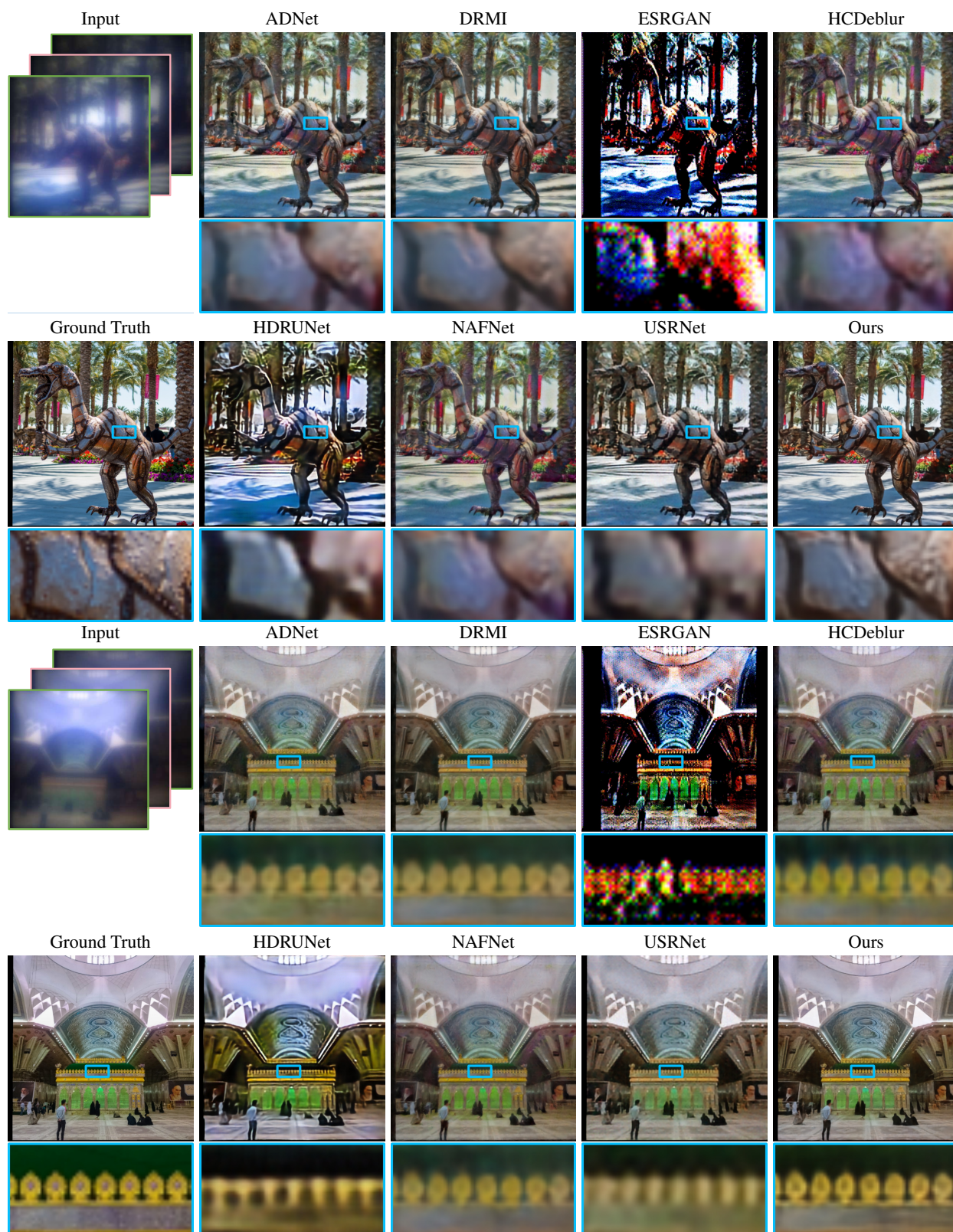
- [1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017. 3
- [2] Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, and Jonathan T Barron. Unprocessing images for learned raw denoising. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11036–11045, 2019. 3
- [3] Jan Froehlich, Stefan Grandinetti, Bernd Eberhardt, Simon Walter, Andreas Schilling, and Harald Brendel. Creating cinematic wide gamut hdr-video for the evaluation of tone mapping operators and hdr-displays. In *Digital photography X*, volume 9023, pages 279–288. SPIE, 2014. 3
- [4] Samuel W Hasinoff, Dillon Sharlet, Ryan Geiss, Andrew Adams, Jonathan T Barron, Florian Kainz, Jiawen Chen, and Marc Levoy. Burst photography for high dynamic range and low-light imaging on mobile cameras. *ACM Transactions on Graphics (ToG)*, 35(6):1–12, 2016. 1, 3, 4, 5
- [5] Samuel W. Hasinoff, Dillon Sharlet, Ryan Geiss, Andrew Adams, Jonathan T. Barron, Florian Kainz, Jiawen Chen, and Marc Levoy. Burst photography for high dynamic range and low-light imaging on mobile cameras. *ACM Trans. Graph.*, 35(6), Dec. 2016. 7
- [6] Andrey Ignatov, Luc Van Gool, and Radu Timofte. Replacing mobile camera isp with a single deep learning model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 536–537, 2020. 3
- [7] Nima Khademi Kalantari, Ravi Ramamoorthi, et al. Deep high dynamic range imaging of dynamic scenes. *ACM Trans. Graph.*, 36(4):144–1, 2017. 3
- [8] Jamy Lafenetre, Gabriele Facciolo, and Thomas Eboli. Implementing Handheld Burst Super-Resolution. *Image Processing On Line*, 13:227–257, 2023. <https://doi.org/10.5201/ipol.2023.460>. 7
- [9] J. Liu, D. Liu, W. Yang, S. Xia, X. Zhang, and Y. Dai. A comprehensive benchmark for single image compression artifact reduction. *IEEE Transactions on Image Processing*, 29:7845–7860, 2020. 3
- [10] Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2, IJCAI’81*, page 674–679, San Francisco, CA, USA, 1981. Morgan Kaufmann Publishers Inc. 7
- [11] Tom Mertens, Jan Kautz, and Frank Van Reeth. Exposure fusion. In *15th Pacific Conference on Computer Graphics and Applications (PG’07)*, pages 382–390. IEEE, 2007. 3, 10
- [12] Anurag Ranjan and Michael J. Black. Optical flow estimation using a spatial pyramid network, 2016. 7
- [13] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume, 2018. 7
- [14] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers, 2021. 4, 5, 7
- [15] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow, 2020. 7



**Figure 3.** *Significance of real-world adaptation.* Our model trained just on the synthetic dataset performs poorly in HDR scenes in the real world visualized by the incorrect burst fusion maps. We refine them through unsupervised finetuning using saturation maps computed directly from the real burst frames leading to higher quality fusion maps.



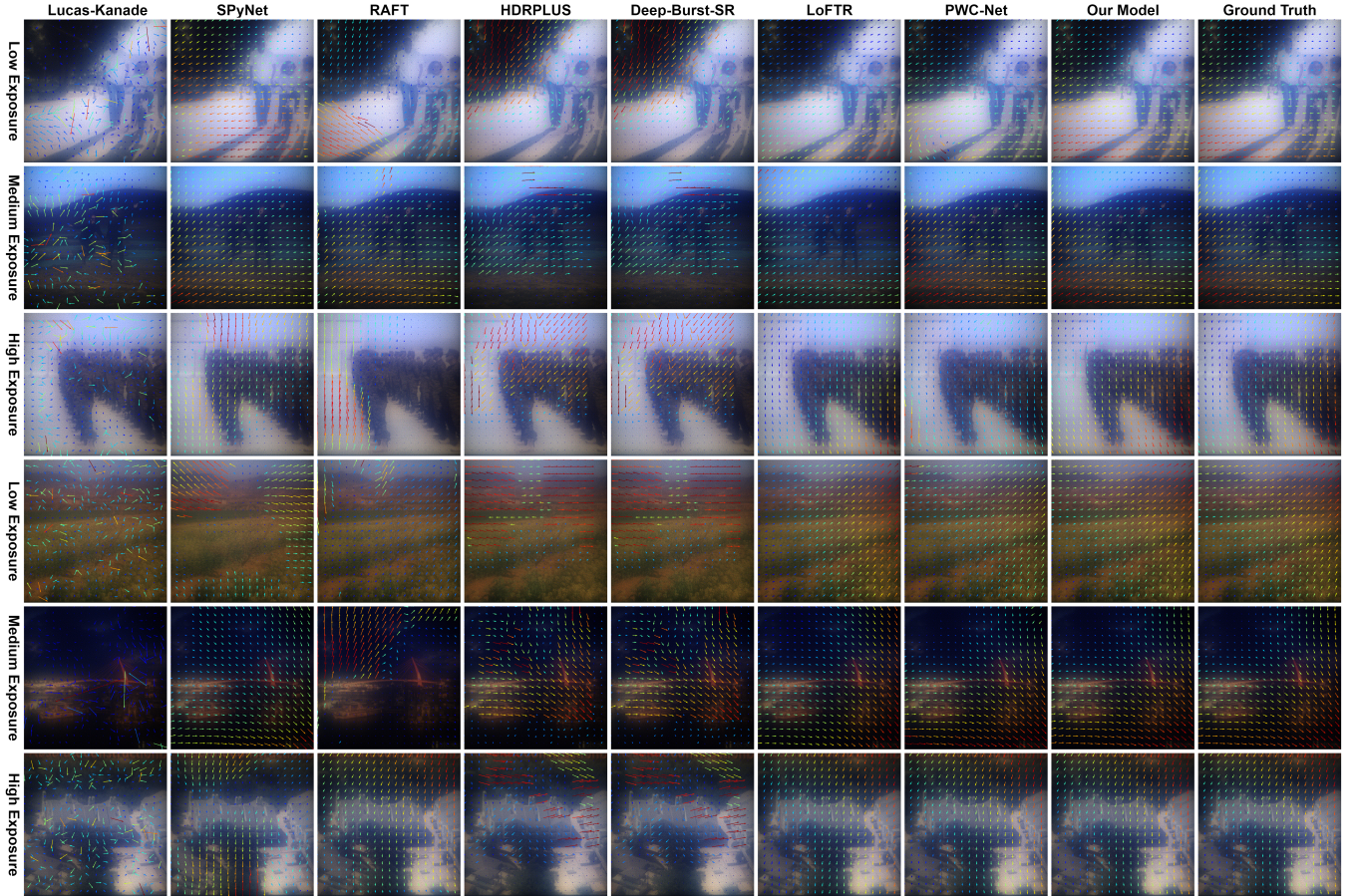
**Figure 4.** *Restoration with mis-alignment.* Our restoration network performs well in the case of moderately misaligned bursts (LoFTR [14], second row) while leads to slight blurring in case of higher misalignment (HDR+ [4], third row). On the other hand, precise alignment using our proposed alignment module (first row) leads to sharper image features.



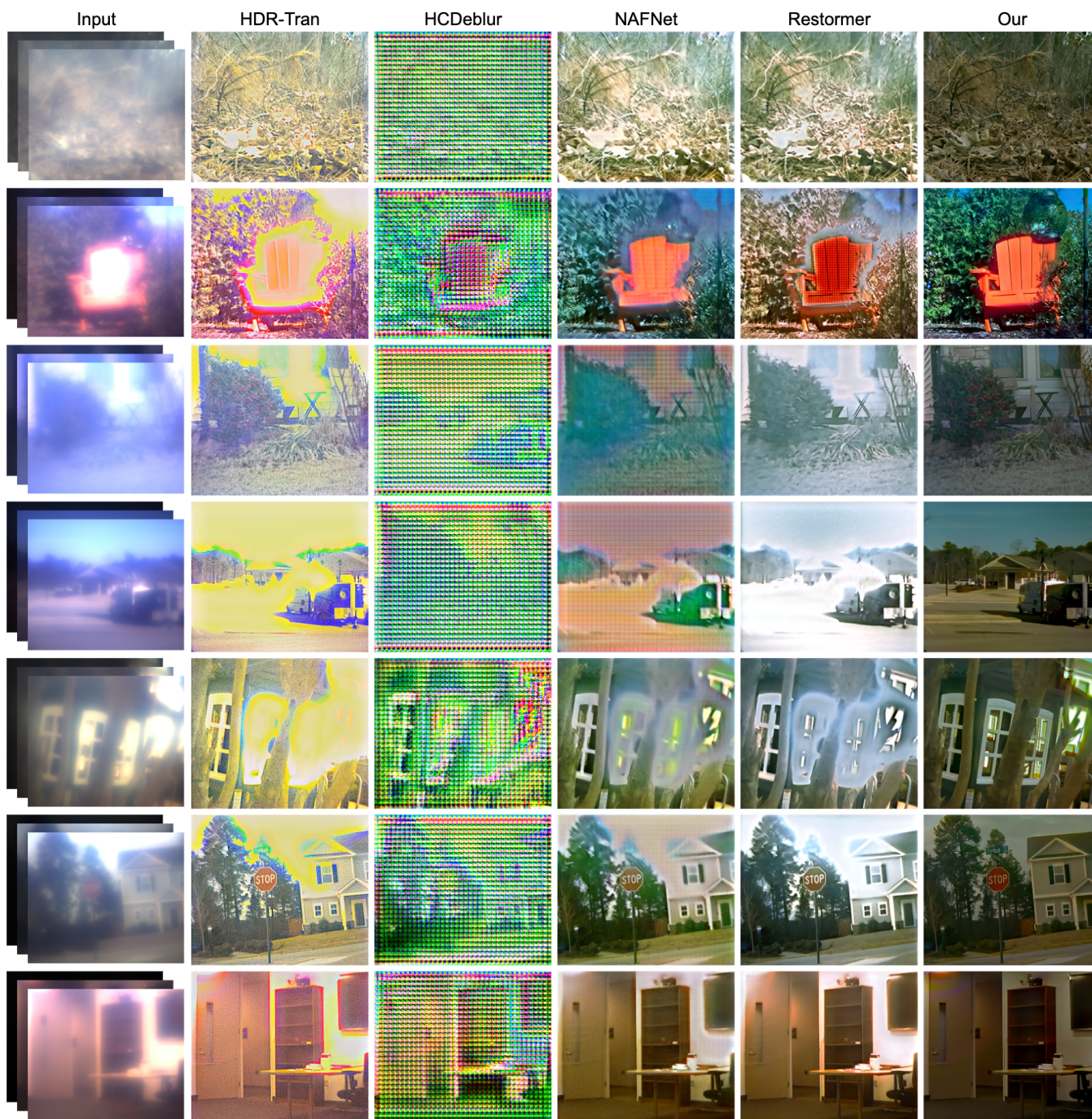
**Figure 5.** Visual comparisons for different restoration models evaluated on OLED data

**Table 2.** Comparison of alignment methods across different exposure conditions. Each metric is shown as mean  $\pm$  standard deviation where applicable.

Method	Low exposure			High exposure		
	Mean $\downarrow$	Cosine $\uparrow$	Median $\downarrow$	Mean $\downarrow$	Cosine $\uparrow$	Median $\downarrow$
Lucas-Kanade [10]	109.35 $\pm$ 28.02	0.02 $\pm$ 0.71	135.91 $\pm$ 41.43	50.25 $\pm$ 11.20	0.02 $\pm$ 0.70	69.22 $\pm$ 15.20
SPyNet [12]	75.82 $\pm$ 27.48	0.05 $\pm$ 0.70	110.72 $\pm$ 45.02	59.64 $\pm$ 26.94	0.15 $\pm$ 0.69	92.74 $\pm$ 45.66
RAFT [15]	63.70 $\pm$ 61.19	0.54 $\pm$ 0.68	75.85 $\pm$ 103.93	49.11 $\pm$ 56.97	0.63 $\pm$ 0.61	54.69 $\pm$ 97.80
HDRPLUS [5]	44.92 $\pm$ 13.10	0.10 $\pm$ 0.70	59.63 $\pm$ 21.73	34.76 $\pm$ 11.40	0.06 $\pm$ 0.70	47.09 $\pm$ 13.63
Deep-Burst-SR [8]	45.00 $\pm$ 13.09	0.11 $\pm$ 0.70	59.78 $\pm$ 21.68	34.79 $\pm$ 11.39	0.06 $\pm$ 0.70	47.15 $\pm$ 13.61
LoFTR [14]	3.94 $\pm$ 3.04	0.94 $\pm$ 0.19	5.95 $\pm$ 4.76	3.04 $\pm$ 2.59	0.96 $\pm$ 0.16	4.62 $\pm$ 4.01
PWC-Net [13]	2.17 $\pm$ 3.47	0.97 $\pm$ 0.18	<b>1.25<math>\pm</math>1.20</b>	1.50 $\pm$ 2.05	0.98 $\pm$ 0.15	<b>0.92<math>\pm</math>0.47</b>
<b>Ours</b>	<b>1.40<math>\pm</math>1.15</b>	<b>0.99<math>\pm</math>0.06</b>	2.02 $\pm$ 1.54	<b>1.32<math>\pm</math>0.75</b>	<b>0.99<math>\pm</math>0.05</b>	1.89 $\pm$ 0.99



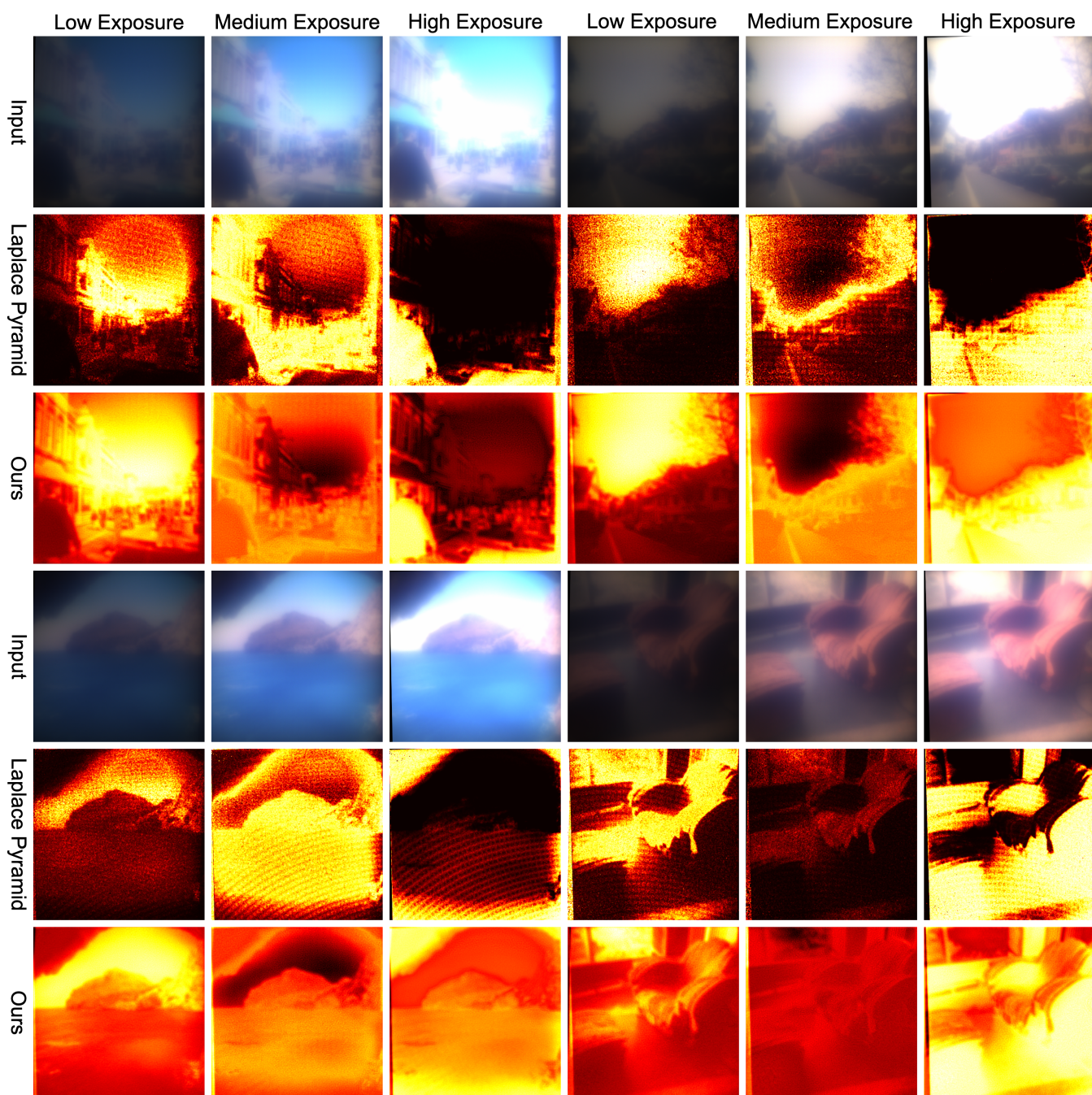
**Figure 6.** Visualizing predicted displacement fields across various burst alignment algorithms



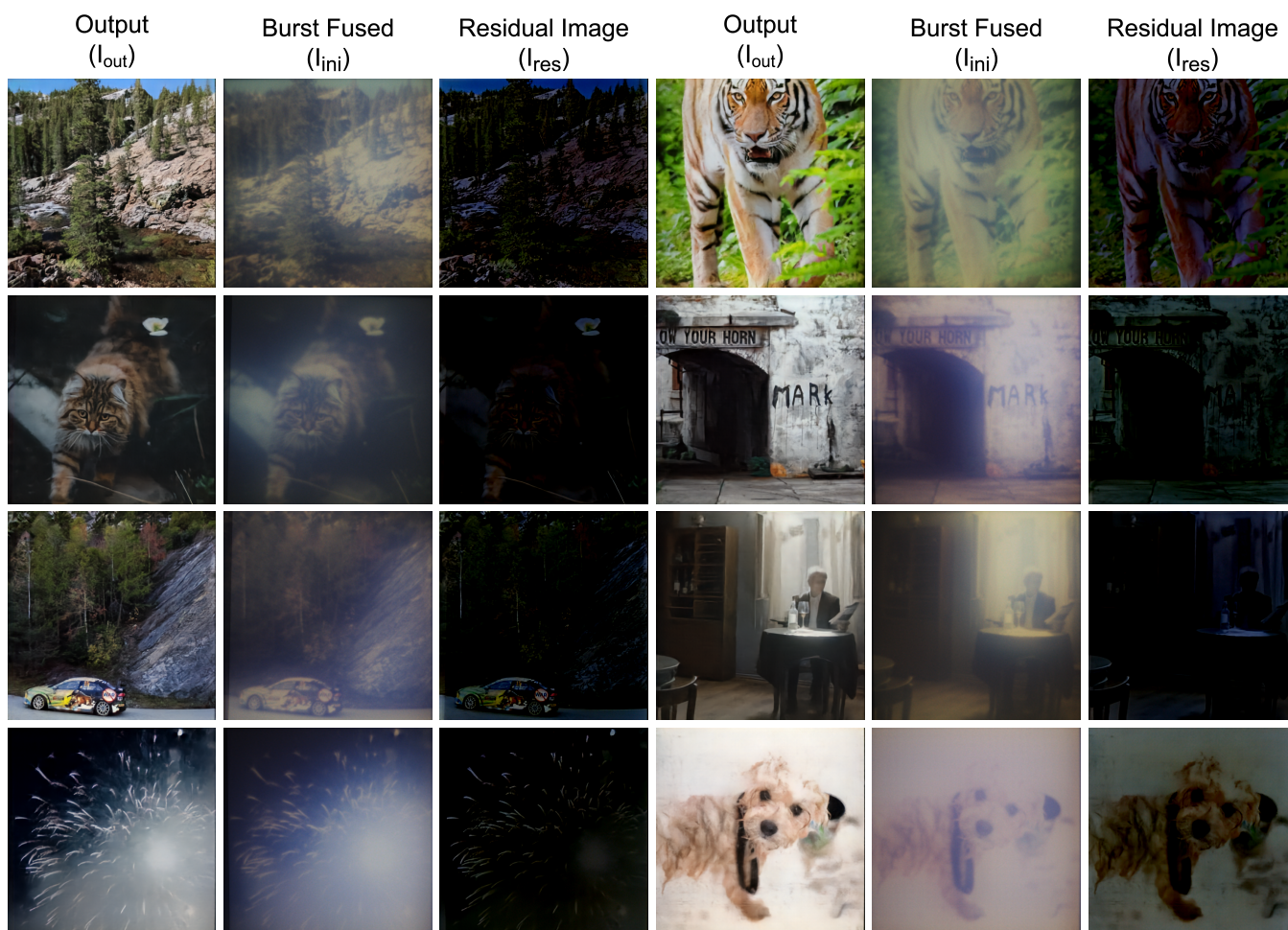
**Figure 7.** Non-cherry picked examples showcasing performance of the best methods against ours.



**Figure 8.** More HDR scene reconstructions with metalens bursts



**Figure 9.** Laplace Pyramid fusion maps from Mertens et al. [11] vs our learned fusion maps.



**Figure 10.** Visualizing the contribution of the weighted fusion module ( $I_{ini}$ ) and the restoration module ( $I_{res}$ ) separately.