MIORe & VAR-MIORe: Benchmarks to Push the Boundaries of Restoration

Supplementary Material

Lens	CANON24	SIGMA85	TAMRON15-30	LAOWA100
# Sequences	172	62	78	21

Table 7. MIORe Statistics: lenses

Translation	Oxyz	Oxz	Oxy	Oyz	Ox	Oy	Oz	stable
# Sequences	9	27	6	11	56	5	27	192

Table 8. MIORe Statistics: Ego-camera translation

Rotation	YPR	YP	YR	PR	Y	P	R	stable
# Sequences	28	16	113	11	29	51	3	82

Table 9. MIORe Statistics: Ego-camera rotations

Weather	clouds	fog	rain	snow	clear
# Sequences	76	4	21	13	219

Table 10. MIORe Statistics: Weather conditions

Exposure intensity	Slight	Medium	Severe
Overexposure	47	20	9
Underexposure	18	12	6

Table 11. MIORe Statistics: Exposure

7. Appendix

7.1. Additional Metadata Information

We emphasize the importance of comprehensive metadata annotation to enhance both the utility and reliability of our dataset. For each sequence, detailed information is recorded, thereby highlighting the dataset's diversity and real-world relevance. Such rich annotations provide comprehensive insights into each sequence, ensuring transparency and reproducibility in future research, while also facilitating the creation of new splits for upcoming challenges. This Section is linked to the main paper Section 4.

Overall, our per-sequence annotations include the following details:

- General sequence information: lens type, sequence number, number of samples generated for the dataset, initial number of raw frames in the sequence, blur intensity used, extreme motion label.
- Types of ego-camera motion: State: static (standing), dynamic (walking, riding, driving); Translations: Ox, Oy, Oz; Rotations: Shake, Pitch, Yaw, Roll; Tracking.
- Scene-specific details: Background: depth-variable or flat; Dynamic entities throughout the scene: vehicles, humans, animals, liquid, fire, foliage; Conditions: Occlusions, Defocus; Weather: clouds, fog, rain, snow; Light-based effects: sun-flare, reflections, overexposure, underexposure.
- Artifacts or other potential problems: sensor retention, aliasing, bayering, LED flickering.

Table 8 depicts the number of sequences affected by translations on the Oxyz axis. While the Ox and Oy movements influence the blur in a linear fashion, namely the relationship between the translated points remains geometrically mostly constant, the Oz translation creates a radial effect - as we zoom in towards the vanishing point, or zoom out of it, the projected speed of the objects on the frame increases with their distance to the zooming center.

Similarly, Table 9 depicts the number of sequences affected by Yaw, Pitch, and Roll rotations. On the one hand, Y and P rotations have a similar valence as the Oy and Ox translations, respectively. Nevertheless, rotations and translations are different, especially in highly depth-variable scenes. However, when objects are at a considerable distance from the ego-camera, the movement types converge in behavior. On the other hand, the R rotation is different, as it creates variable blur proportional to the distance of the points to the center of rotation. Again, we may observe some similarities to the Oz translation, because of the motion magnitude dependence to a center point. Nonetheless, the interaction between points observed in the blurry images are different, and only the observation regarding linear versus non-linear patterns that emerge holds.

In another line of thought, by analyzing the moving entities of all the sequences, we may tell none have all 6 types of moving entities present all at once, and not even a combination of all but one. Moreover, only 3 out of the 333 sequences have 4 entities at once. This further underlines the difficulty of recording data, and the rarity of real-life scenarios in which these types of objects occur. Obviously, we selected few static scenes, namely 37, since they are of lower relevance to our set of selected tasks. Mostly, the sequences contain only one (158) or two (104) entities in motion. It is important to mention that the occurrence of fewer objects at the same time does not decrease the difficulty of the dataset.

7.2. Data Acquisition Procedure

During the data collection phase, we adhered to best practices to ensure high-quality recordings. Black calibration was performed prior to filming and repeated as necessary when sensor temperatures increased, maintaining optimal image quality. All calibration processes were conducted in a light-controlled environment to eliminate light leaks and other potential interferences.

Exposure and Lighting: Exposure settings were customized for each scene:

- Adjustments to shutter angle or lens aperture were used to control light input, avoiding overexposure.
- No artificial exposure gain (+EV, ISO) was applied for underexposed scenes, limiting outdoor captures to day-

Model	X	F	MF	M	MS	S
FFTformer	0.216	0.117	0.035	0.075	0.095	0.043
[14]	176.2	88.7	14.3	29.5	30.2	7.2
NAFNet	0.305	0.254	0.025	0.083	0.159	0.046
[4]	141.1	113.6	10.6	21.6	21.0	8.9

Table 12. LPIPS and FID of Motion Deblurring on *MIORe*

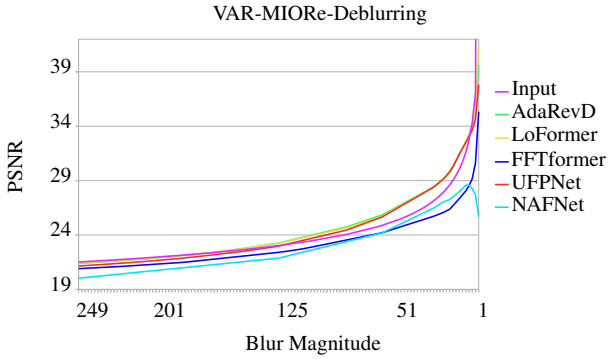

Figure 10. Trend of PSNR performance for Motion Deblurring on *VAR-MIORe*.

light hours and indoor recordings to Amraam 300c 300W lights (6400K temperature).

Custom color correction matrices (CCM) were applied based on light intensity, season, and geographic location. Additionally, white balancing (W/B) was dynamically adjusted to reflect environmental conditions and lighting variations. This Section is related to the main paper Subsection 4.2.

7.3. Further Benchmarking Results

In this Section we are presenting some visualizations of the tables from main paper, Section 5.

7.3.1. Additional Evaluation Metrics

We employ LPIPS to measure perceptual similarity and FID to assess distributional alignment with ground truth. As shown in Table 12, these metrics reveal complementary insights across motion categories of varying difficulty, and highlighting, for example, that lower LPIPS does not always correlate with better FID. This reinforces the need for a multi-metric evaluation protocol. Table 12 includes this subset as a representative case for frequency-based versus regular self-attention models' behavior.

7.3.2. Quantitative Results

Single Image Motion Deblurring:

Figure 14 depicts the radar chart PSNR and SSIM performances visualization of the single image non-uniform motion deblurring task on *MIORe*, corresponding to the main paper Table 2. Figures 10 and 11 present the PSNR and

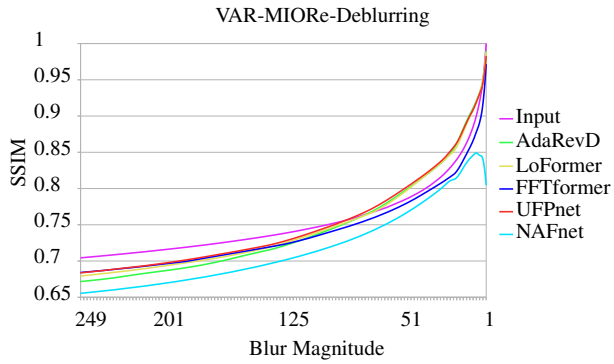

Figure 11. Trend of SSIM performance for Motion Deblurring on *VAR-MIORe*.

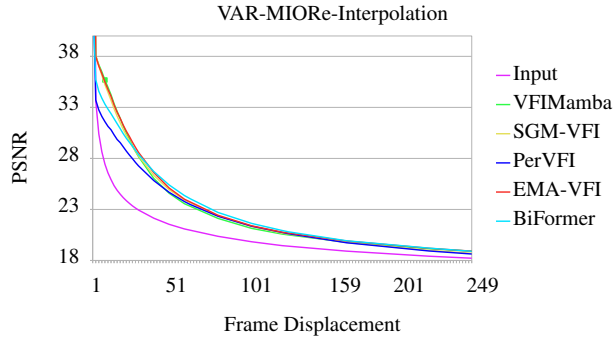

Figure 12. Trend of PSNR performance for Video Frame Interpolation on *VAR-MIORe*.

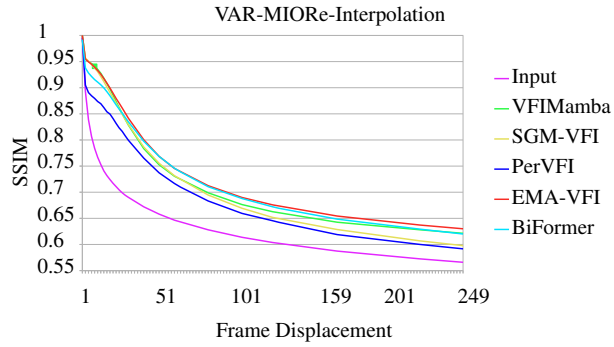

Figure 13. Trend of SSIM performance for Video Frame Interpolation on *VAR-MIORe*.

SSIM on *VAR-MIORe*, respectively. They offer more granular information, containing additional data points compared to the main paper Table 3.

Video Frame Interpolation:

Figure 15 displays the radar chart PSNR and SSIM performances visualization of the video frame interpolation task on *MIORe*, linked to the main paper Table 4. Figures 12 and 13 showcase the PSNR and SSIM values on *VAR-*

Motion Source	Ego-Camera					Scene Motion								
Motion Type	Static	Oxyz Translation		Yaw/Pitch/Roll Rotation		Static Background		Dynamic Entities						
Components	Stable	Parallel	Radial	Chaotic Shake	Tracking	Flat	Depth-Variable	Vehicles	Humans	Animals	Liquid	Fire	Foliage	Salient Subjects
MIORe	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
VAR-MIORe	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
GoPro [20]	✓	✓	✓	✓			✓	✓	✓		✓		✓	
RealBlur [21]				✓		✓	✓							
Vimeo90K [29]	✓	✓			✓	✓	✓	✓	✓	✓	✓	✓		✓
X4K1000FPS [24]	✓	✓	✓		✓		✓	✓	✓	✓	✓		✓	✓
KITTI [10]	✓	✓	✓				✓	✓	✓					
Sintel [3]		✓	✓		✓		✓		✓	✓			✓	✓

Table 13. Comparison of motion types present in our novel datasets *MIORe* and *VAR-MIORe* against the other state-of-the-art datasets in the literature for our tasks of choice.

Tasks	All ***(Ours)***		Motion Deblurring		Video Frame Interpolation		Optical Flow Estimation	
Dataset	***MIORe***	***VAR-MIORe***	GoPro [20]	RealBlur [21]	Vimeo90K [29]	X4K1000FPS [24]	KITTI [10]	Sintel [3]
Camera	Chronos 2.1-HD		GoPro Hero4	2 × Sony A7RM3	multiple	Phantom Flex4K	custom	Blender
Lenses	4		1	1	unspecified	1	1	virtual
Camera FPS	1000		240	2	30	1000	10	24
Rendered FPS	28-1000	4-1000	18-34	80	30	240	10	24
Blur Intensity	3-35	0-249	7-13	real	-	-	-	-
Max Flow	95	1932	135	47	65	288	355	414
Total Size	52218	83250	3214	4556	73171	4888	400	1628
Resolution	1920 × 1080		1280 × 720	680 × 773	448 × 256	768 × 768	1242 × 375	1024 × 436

Table 14. The comprehensive comparison of our novel multi-task *MIORe* dataset variants against the existing single-task datasets.

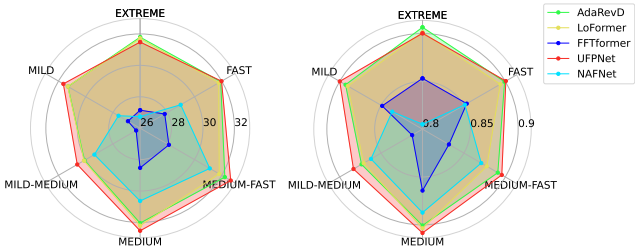

Figure 14. PSNR (left) and SSIM (right) performances of Motion Deblurring state-of-the-art methods on *MIORe*.

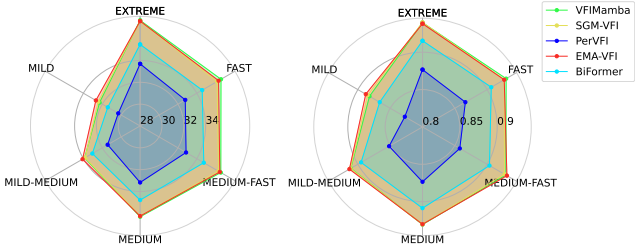

Figure 15. PSNR (left) and SSIM (right) performances of Video Frame Interpolation state-of-the-art methods on *MIORe*.

MIORe, respectively. They offer a better depiction of our extensive experiments; having all the data points available, it represents the complete version of Table 5 from the main paper. We have chosen to present these graphic representations in the supplementary, since the exact numbers offered in the experiments section of the main paper are more relevant to our takeaways and insights.

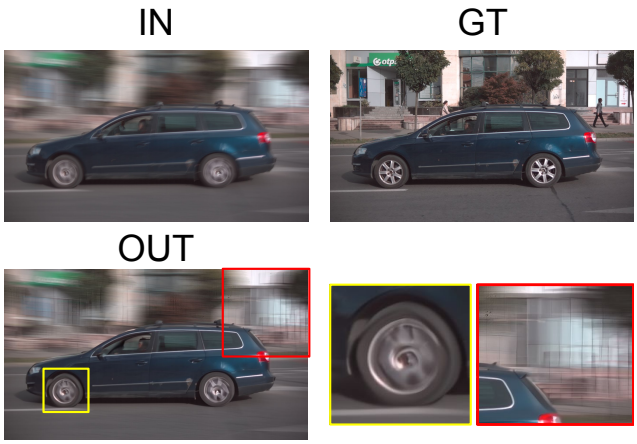

Figure 16. Qualitative analysis of FFTformer [14] on one frame of our *MIORe* dataset. Top-left quadrant displays the input blurry image. Top-right quarter shows the equivalent ground truth label. Bottom-left part represents the output of the model, and on the bottom-right side one may observe the highlighted degradations of the model. In this case, FFTformer creates a grid-like degradation due to the overwhelming blur intensity. It is to be noted that some particular movement types are underrepresented in multiple datasets that we compared to. In this particular case, we may notice how models trained on [20] are unable to correctly restore the spinning wheel. This effect is persistent with all the models we have benchmarked for motion deblurring.

7.4. OF Pseudo-GT Labels Limitations

Frame averaging: We provide the pseudocode 1. OF computation was partially inspired from [24]. We synthesize blur following [20]. Differently, instead of having a fixed window size for OF, we adapt it for each sequence.

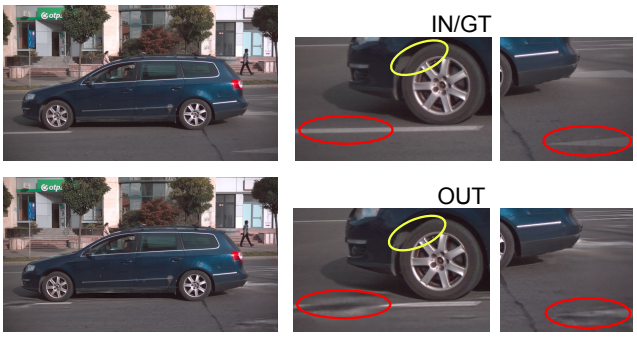

Figure 17. Qualitative analysis of FFTformer [14] on one zero-blur frame of our *VAR-MIORe* dataset. It is visible that the model is unaware that the input is sharp, and needs no deblurring, hence the artifacts. Firstly, the model appears to warp objects and it fails to preserve the shape consistency of the wheel. Moreover, the model seems to smear the high-contrast lines on the road. Similar artifacts can be seen for the other benchmarked models.

Algorithm 1 MIORe blurring policy

procedure GETSEQOF(seq)
  $fs \leftarrow [model(selfEnsemble(im))$ **for** $im \in seq]$
  return $annotate(fs)$ **if not** $QC(fs)$ **else** fs
end procedure

procedure GETCONFIGURATION($\mathcal{D}$)
  $bkt \leftarrow \{flow_i : (window_i, stride_i)\}_{i=0..flow_{max}}$
  $configs_s \leftarrow bkt[mean(getSeqOF(s))]$ **for** $s \in \mathcal{D}$
  $return\ configs$
end procedure

We classified our flow magnitudes into 6 buckets (bkt), ranging from mild to extreme. The OF is produced using a $self\ ensemble$ technique that applies all possible rotations and flips to an img. The OF estimation $model$ yields the predicted masks which are restored to the original orientations. Then, by averaging all the computed flows, we reduce potential underlying biases.

Yet, the SOTA OF model may still fail in edge cases. We $annotate$ the OF when the offline quality control QC indicates us to. Inspired by [3], the QC method implies verifying segmentation and occlusion masks' features' alignment compared to the OF map. Figure 23 displays a variety of pseudo-GT labels. There are a few aspects to be noted, namely that the granularity of the model used may not be high enough, since some motion may have not been seen by the OF estimation models, which makes it unreliable in offering us high-fidelity annotations. At the same time, Figure 24 displays three the OF maps generated for consecutive frames, displaying inconsistencies in the pseudo-GT labels.

Figure 18. Qualitative analysis of NAFNet [4] on a sample frame of our *MIORe* dataset. Left column displays the output of the model, provided the blurry-rendered inputs, while on the right we have the corresponding ground truth sharp images.

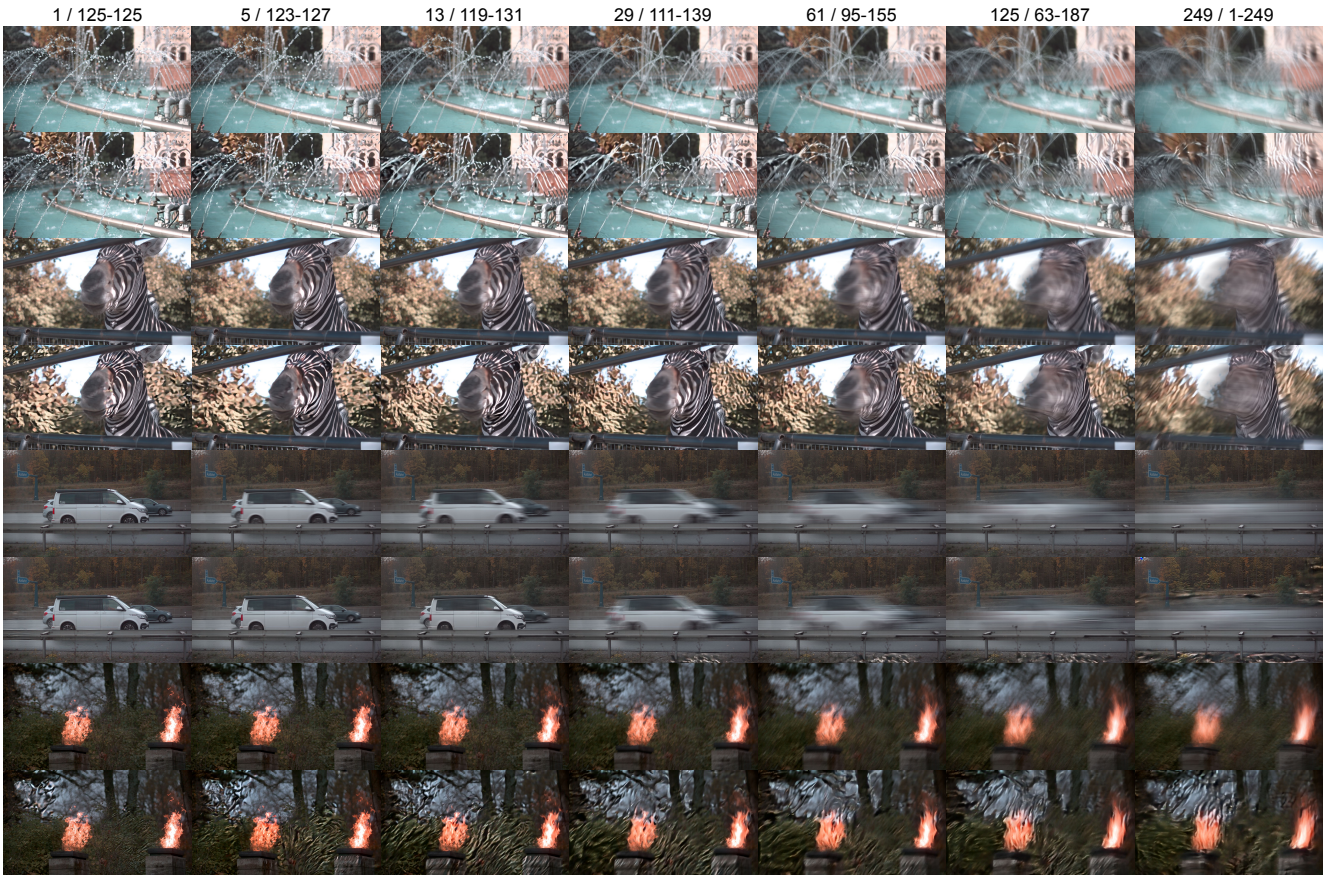

Figure 19. Qualitative evaluation of FFTformer [14] on several samples of from the our novel *VAR-MIORe* dataset. Odd rows show the equivalent ground truth label, while even rows depict the model's output. The numbering above the figure mentions how many frames have been averaged to achieve the blurry input, and the frame numbers respectively. To be marked are the dataset difficulty, through its extreme motion magnitude. The models benchmarked on *VAR-MIORe* display ineffectiveness when dealing with blurry images beyond a certain thresold.

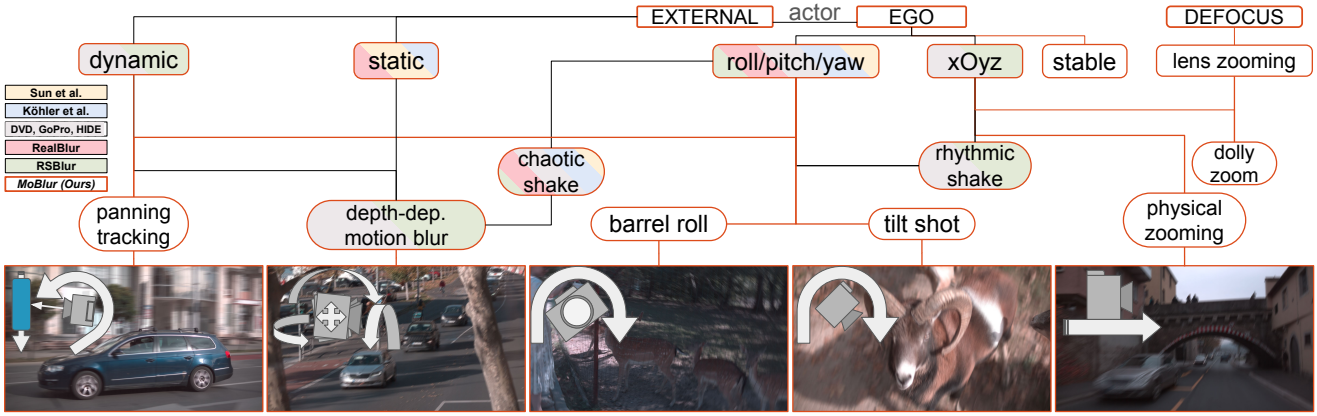

Figure 20. Visual explanation regarding motion types described in Section 3: *MIORe* introduces several motion types dependent both on ego-camera actions and the scene's subject. Therefore, the contrast between foreground, subject, and background are better highlighted. Referenced in this comparison are only motion deblurring datasets. The lines symbolize the interaction between the motion components to create a certain effect. The red borders of the white-background text bubbles indicate the presence of such effects only in our datasets. This representation extends the concepts presented in Figure 7 of the main paper, and compares the existing motion types and patterns in *MIORe* with some the ones of previous deblurring datasets, some of which mentioned in Subsection 4.1.

Figure 21. Video Frame Interpolation visual results of SGM-VFI [15] on several samples from our *VAR-MIORe* dataset, having the maximum offset of 0.25 seconds between the two input frames. The left and right frames are the inputs, and the middle frame is split between the middle ground truth, and the model's prediction, which is surrounded by a red border.

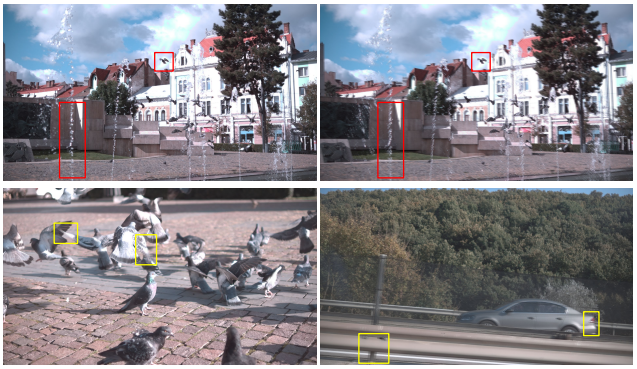

Figure 22. Video Frame Interpolation visual results of VFI-Mamba [32] on two samples from our *MIORe* dataset, having the regularized offset of 9 frames (0.009 seconds) between the two input frames. Top-left quadrant depicts the overlaid motion differences between the first input frame and the model's prediction. Top-right quadrant displays the visibly bigger offset between the prediction and the last input frame. From the first row, one may conclude that some models fail to precisely interpolate the middle frame, but the quality of the prediction is very good nonetheless, with no distortions. The bottom row further shows the offset between the ground truth labels and the predictions. Although small, the differences are significant, since the maximum pixel travel distance is approximately of 40 units for both the birds and vehicle image.

Figure 23. Optical Flow Pseudo-Ground-Truth generated labels.

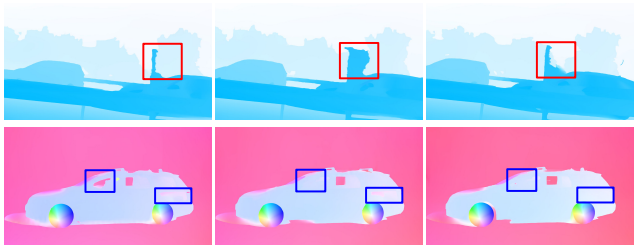

Figure 24. Generated OF pseudo-GT labels. Purpose of this visualization is to raise awareness of the inconsistencies found across different frames. Due to reflections, shadows, large offsets, or model suboptimalities there are several pixels that are being classified as belonging to different entities. Unfortunately, there is no immediate solution to mitigating this problem, and it is required to further study this subject in the future.