Saliency-aware Stereoscopic Video Retargeting (Supplementary materials)

A. Introduction

In this supplementary material, we provide more results and ablation studies to demonstrate the efficiency of our proposed **Saliency-aware Stereoscopic Video Retargeting**. In section B, we provide results to illustrate the importance of the salient object detection module. Section C presents more ablation study results, more failure cases, and the dilation operation. More qualitative and quantitative results are provided in section D.

B. Importance of Salient Objects Detection

As stated in the main paper, detecting salient objects as accurately as possible in a stereo video is an essential step of our model. In this work, we combine the Co-Saliency detection module (CoSD) [8], disparity information, and object detection method to accurately detect the salient objects.

We first show that using CoSD module [8] alone is not sufficient to accurately detect salient objects. CoSD model can perform well on videos from DAVIS [6] dataset, which it is being trained on, as shown by sample results in Figure 1. However, when we apply CoSD to detect the salient objects in videos from the KITTI stereo 2015 [5] and 2012 [3] datasets, the detection results are not accurate. Some examples of failure results are shown in Figure 2. As observed, the results are not accurate. There are two possible reasons that CoSD does not perform well for videos from KITTI stereo 2015 and 2012 datasets. Firstly, the camera used for capturing the videos in the DAVIS dataset is stationary, while the KITTI stereo 2015 and 2012 datasets are captured with moving cameras. Secondly, the CoSD model has not been trained on the videos from the KITTI stereo 2015 and 2012 datasets.

Similarly, detecting salient objects using just the disparity information is not sufficient. Figure 3 shows the disparity values for some videos from KITTI stereo 2012 and 2015 datasets. The cars are apparent in these examples. The main challenge with the disparity values is that although the salient objects are apparent visually, the noise produced by disparity values of other nearby parts of the scene, such as the road or the surrounding objects, shows that disparity information, by itself, is insufficient for detecting salient objects.

Figure 1. Results of CoSD on the Davis [6] dataset. The CoSD is trained on the Davis [6] dataset. So, the results are with high accuracy.

To gain better performance for salient objects detection, we first employ an object detection method, the efficient Yolov5 [9], to detect the location of the salient objects (cars). Then, we combine CoSD and disparity information, with Yolov5, to segment the salient object from the background. Figure 4 shows the results for some test frames. As can be seen from this figure, the key salient objects are segmented from the background.

C. Ablation Studies

In the main paper, due to space constraints, we only provide ablation studies for the saliency module (CoSD) and stereo video Transformer block. Here, we provide further ablation studies to illustrate the importance of the reconstruction block and the design of the loss functions in influencing the performance of stereo video retargeting.

C.1. Impact of the reconstruction block

As shown in Figure 5, we removed the reconstruction block from the left and right streams Figures 6, 7, and 8 show the results of this study. It is apparent that the reconstruction block affects the training, and its absence results in poor inference performance in most of the test cases. The effect of removing the reconstruction block is indirect. It changes the loss function, causing the actual loss cannot be

Figure 2. Failure results of CoSD on KITTI stereo 2015 [5] (rows #1 and #2) and 2012 [3] (rows #3 and #4) datasets. The CoSD is not trained on KITTI stereo datasets.

Figure 3. Disparities of example videos from KITTI stereo 2015 [5] (rows #1 and #2) and 2012 [3] (rows #3 and #4) datasets. The salient object (cars) are apparent in the disparity maps, but they could not be used as the saliency detection method because of the high disparity values of near objects.

Figure 4. Salient objects detected using the combination of CoSD, Yolov5, and disparity.

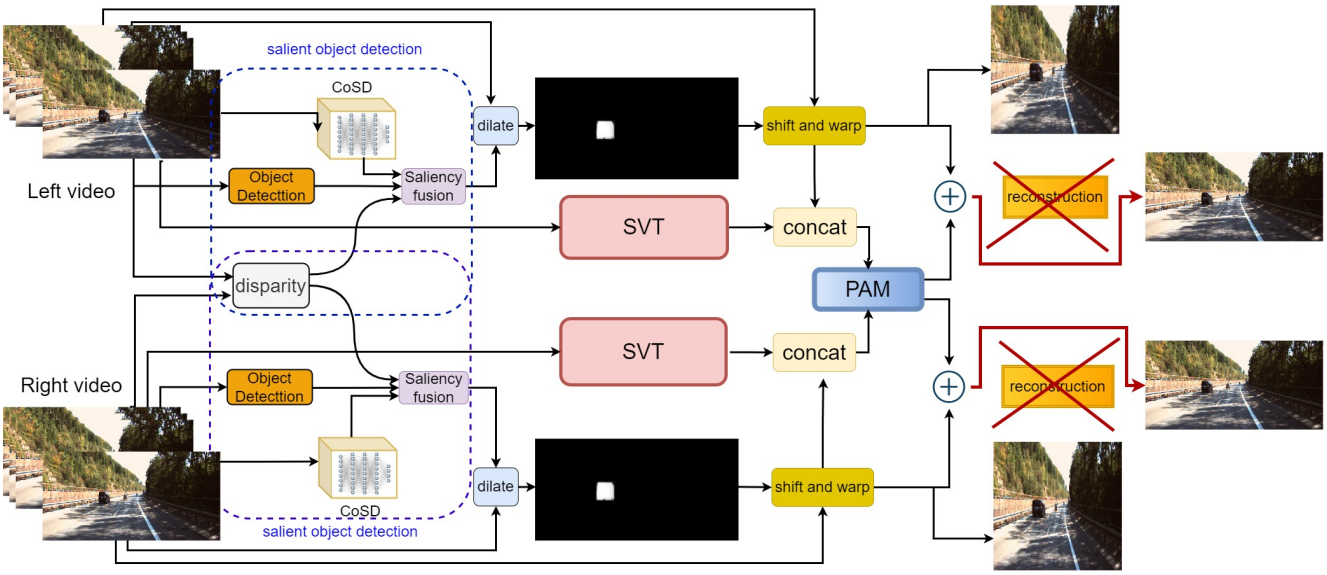

Figure 5. The architecture of the proposed method for stereo video retargeting after removing the reconstruction module.

computed. Therefore, it results in deformed objects.

C.2. Impact of removing the disparity-related loss

The total loss function used to train our model is the union of the three major losses:

$$loss = L_{VGG19}^{total} + \gamma L_{DWT} + (L_{smooth} + L_{photo}) \quad (1)$$

L_{VGG19}^{total} is the loss between VGG19 [7] features extracted from the source and retargeted frames, L_{DWT} is the 2D DWT decomposition loss, and $L_{smooth} + L_{photo}$ is the stereo smoothness and photometric losses.

We ablate with removing stereo-related losses and use the Disparity Distortion ratio (DDr) metric [4] to quantify the spatial and temporal depth distortion and changes. Specifically, we remove the $L_{smooth} + L_{photo}$ term from the loss function. Therefore, we trained the model with the

Figure 6. Ablation study. Impact of the reconstruction block for several test stereo videos. From left to right (columns): Input frame, our main model, our model after removing the reconstruction block.

following loss function:

$$loss_{ablation} = L_{VGG19}^{total} + \gamma L_{DWT} \quad (2)$$

Figure 9 shows the results of this study. We calculate the *DDr* values for a few randomly selected videos with

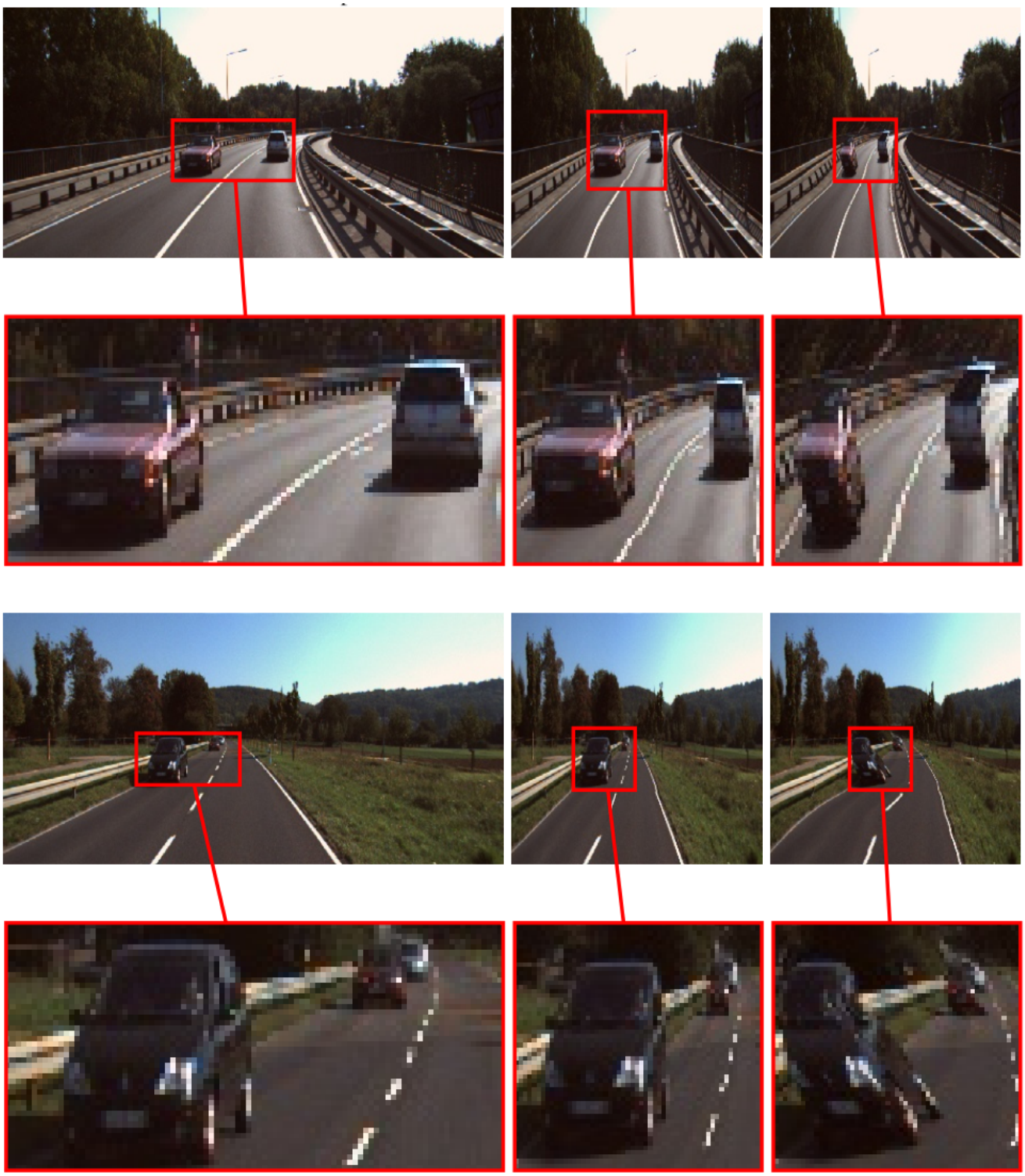

Figure 7. Ablation study. Impact of the reconstruction block for several test stereo videos. From left to right (columns): Input frame, our main model, our model after removing the reconstruction block.

50% size reduction along the horizontal dimension. We first retarget the stereo video. Then, we calculate the disparity map for each stereo frame pair and show it in the fourth column. The DDr values are written on the top right corner

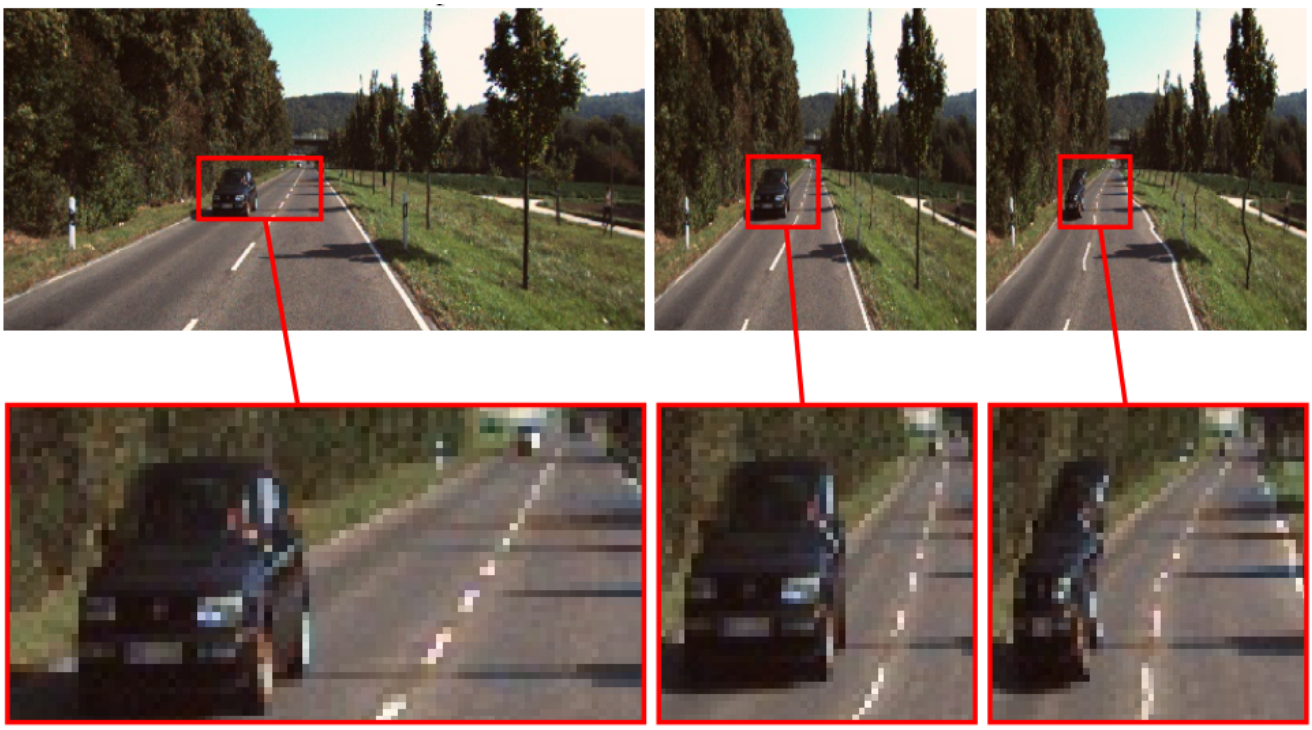

Figure 8. Ablation study. Impact of the reconstruction block for several test stereo videos. From left to right (columns): Input frame, our main model, our model after removing the reconstruction block.

of the respective disparity maps.

C.3. More Failure Cases

We discussed in the main manuscript that our method fails in some situations. When there is extreme retargeting cases like reducing the horizontal size for 50%. For example, when there are a lot of salient objects or the saliency detection module cannot detect all of the salient objects properly. Figure 10 shows examples of those cases where attention map is not created well enough and the retargeted results are not accurate. As shown in the first row, the obvious distortion is more on the structural preservation, e.g. the white lines on the road and the building structure. In the second row, a few objects at the back are detected (with darker gray) but they are distorted. This is due to the fact that the constraint on our loss function for objects that the saliency value is small (darker salient objects) is less effective and the model tries to preserve the main salient object and some distortions happen. In the future, we will work on both saliency detection and the loss function to avoid these kinds of distortions.

C.4. Dilation Operation

Before shifting the salient object to the appropriate position, we dilate them to recover parts of the salient object that could have been missed. Figure 11 illustrates the need for this dilation process. If the saliency detection method misses out on parts of the salient object, e.g., the car without one of its wheels, our method will not warp the whole object uniformly, causing deformation in the object of interest.

D. More Qualitative Results

In this section, we do experiments to prove the superior performance of our model on the videos from KITTI stereo 2012 [3] and 2015 [5] datasets. Several videos are randomly selected from both datasets for this study. We compare the performance of our method with four other methods: linear scaling, manual cropping, fast video [2], and seam carving [1] methods. Figures 12 and 13 compare the proposed method's results with four methods for 50% aspect ratio for two randomly selected videos from the KITTI stereo 2015 [5] test set. Each row belongs to the left frame of one of the videos in the dataset. It is apparent from this figure that our method can better preserve both the salient object and the background. As can be observed, the main object is resized less than the background.

Figures 14 and 15 show the retargeting results with horizontal size reduction of 30% and 20%, respectively. These visual results demonstrate that our method is superior to the other methods. These two aspect ratios require a lesser shift

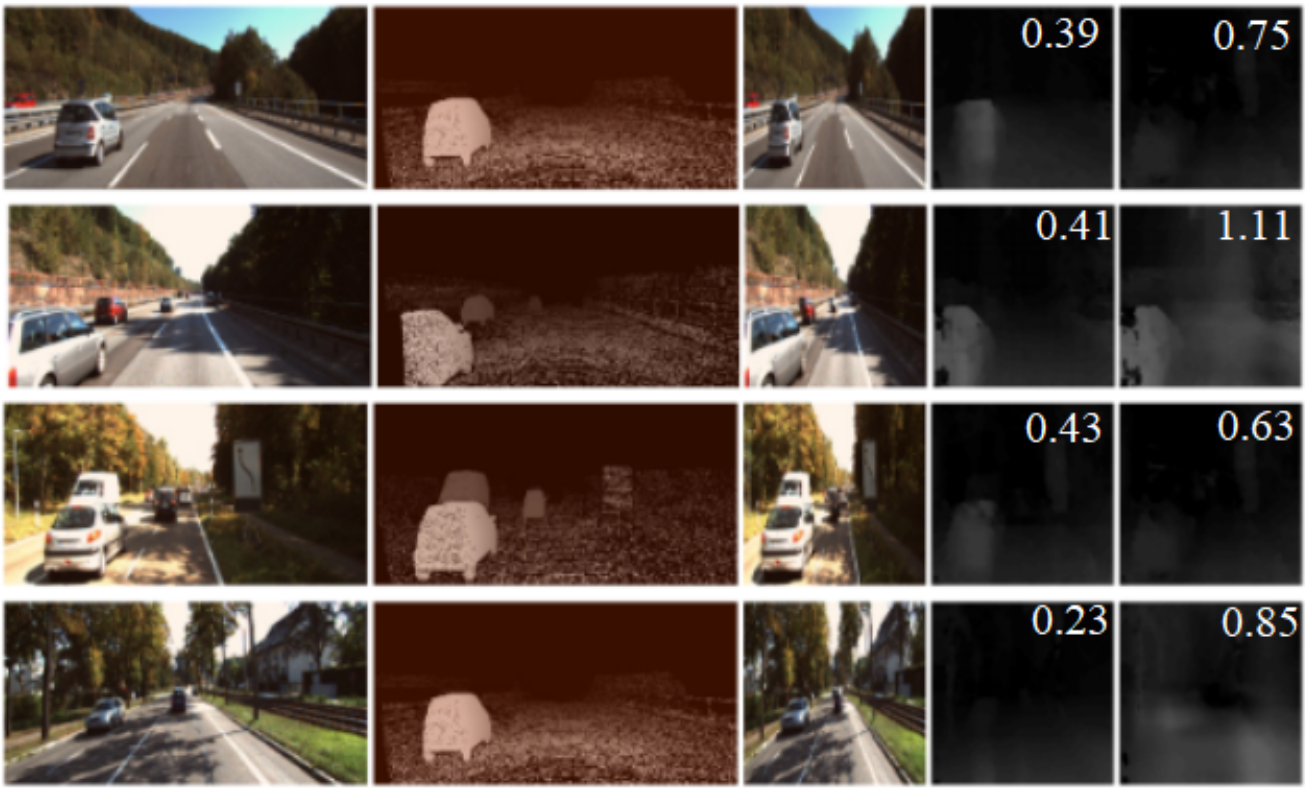

Figure 9. Impact of removing the stereo loss functions. From left to right: Input frame, disparity map, retargeted frame, disparity map calculated with using stereo related loss, disparity map calculated without using stereo related loss. The DDr values are written on top right corner of each disparity map.

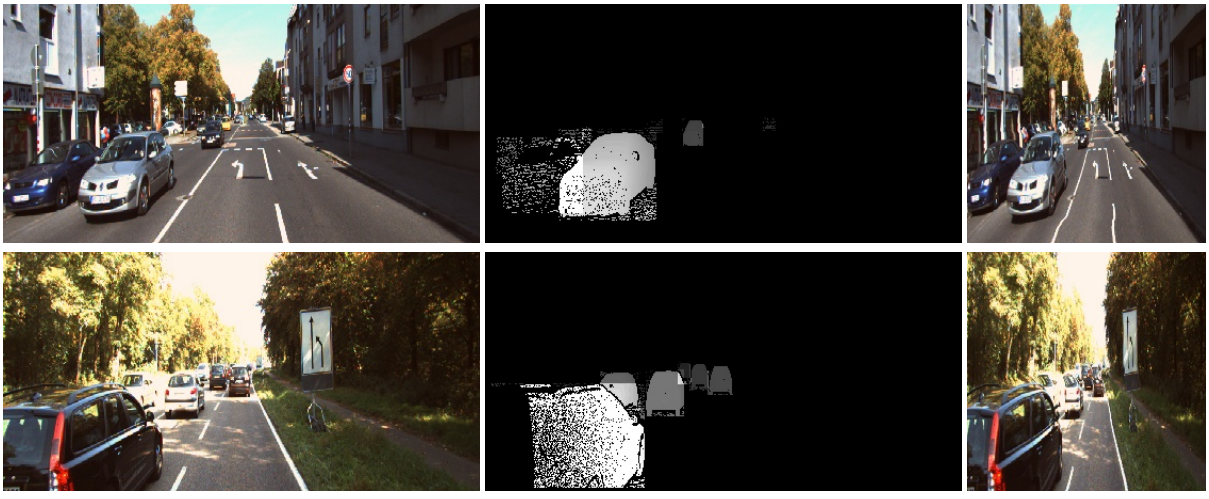

Figure 10. Failure cases where the attention map is not accurate enough and the retargeted results are not accurate.

of pixels. Thus, the main structure of the video frames is well-preserved by all methods.

References

[1] Shai Avidan and Ariel Shamir. Seam carving for content-aware image resizing. In *ACM SIGGRAPH 2007 papers*,

Figure 11. Dilation of the salient region. It helps to enlarge the object to warp without changing its shape.

pages 10–es. 2007. 6, 9, 10

[2] Zhu Chuning. Fast video retargeting based on seam carving with parental labeling. *arXiv preprint arXiv:1903.03180*, 2019. 6, 9, 10

[3] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 1, 2, 6

[4] Bing Li, Chia-Wen Lin, Boxin Shi, Tiejun Huang, Wen Gao, and C-C Jay Kuo. Depth-aware stereo video retargeting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6517–6525, 2018. 3

[5] Moritz Menze, Christian Heipke, and Andreas Geiger. Joint 3d estimation of vehicles and scene flow. *ISPRS annals of the photogrammetry, remote sensing and spatial information sciences*, 2:427, 2015. 1, 2, 6

[6] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 724–732, 2016. 1

[7] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[8] Yukun Su, Jingliang Deng, Ruizhou Sun, Guosheng Lin, and Qingyao Wu. A unified transformer framework for group-based segmentation: Co-segmentation, co-saliency detection and video salient object detection. *arXiv preprint arXiv:2203.04708*, 2022. 1

[9] Ultralytics. Ultralytics/yolov5: Yolov5 in pytorch; onnx; coreml; tflite. 1

Figure 12. More results. Qualitative results of stereo video retargeting on randomly selected videos from the KITTI stereo datasets for 50% reduced the horizontal video size. The results are shown for just the left frames. Left to right: original frame, linear scaling, manual cropping, seam carve [1], fast video [2], and ours.

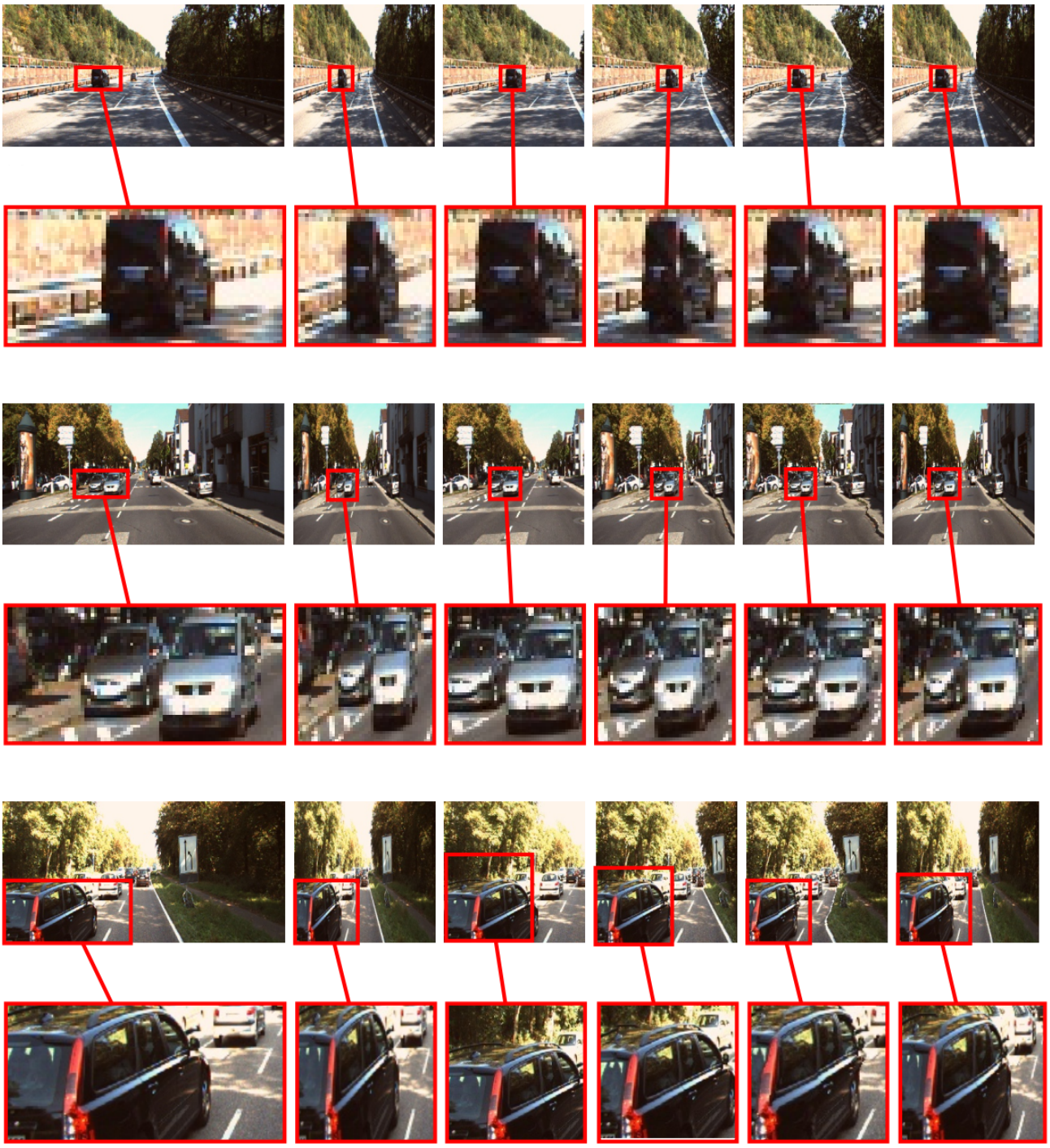

Figure 13. More results. Qualitative results of stereo video retargeting on randomly selected videos from the KITTI stereo datasets for 50% reduced the horizontal video size. The results are shown for just the left frames. Left to right: original frame, linear scaling, manual cropping, seam carve [1], fast video [2], and ours.

Figure 14. More results. Qualitative results of retargeting on randomly selected frames from the KITTI stereo datasets for 30% reduced the horizontal size.

Figure 15. More results. Qualitative results of retargeting on randomly selected frames from the KITTI stereo datasets for 20% reduced the horizontal size.