1. Supplementary Material

1.1. Transformation Synchronisation for VO

For the task of visual odometry, we seek to obtain a trajectory of absolute poses for a driving sequence, from a collection of relative poses. While an obvious way of doing so is applying relative pose between consecutive frames and chaining them together, this is highly inaccurate due to compounding error or individual outliers.

As such, we exploit that our network estimates relative pose between arbitrary overlapping pairs, due to the nature of road planes capturing a significant portion of images in a driving sequence, unlike many self-supervised methods which are limited to adjacent images. However, our method involves estimating camera-relative pose (from ground-relative predictions) between each image in a sequence and the following five images temporally, providing a collection of camera-relative poses at varying distances. Here we summarise a form of transformation synchronisation, the task of optimising absolute pose from a set of relative poses, in the presence of noise and incomplete data.

In particular, we employ the SE(3) spectral motion syncronisation method proposed by Arrigoni et al. [4], but we summarise these details here. In general, we form our collection of camera-relative poses for n cameras into:

$$\mathbf{X} = \begin{pmatrix} \mathbf{I}_4 & \mathbf{M}_{12} & ... & \mathbf{M}_{1n} \\ \mathbf{M}_{21} & \mathbf{I}_4 & ... & \mathbf{M}_{2n} \\ ... & ... & ... & ... \\ \mathbf{M}_{n1} & \mathbf{M}_{n2} & ... & \mathbf{I}_4 \end{pmatrix}, \mathbf{M}_{ij} = \begin{pmatrix} \mathbf{R}_{ij} & \mathbf{t}_{ij} \\ \mathbf{0} & 1 \end{pmatrix}.$$

Given our chosen frame separation offset of zero to four, our relative poses occupy the five blocks of $\mathbf{M}_{ij}$ above and below the main diagonal of $\mathbf{X}$, with the remaining entries unoccupied. The situation for missing relative poses can be written as $\mathbf{L} = ((\mathbf{D} - \mathbf{A}) \otimes \mathbb{1}_{4\times4}) \circ \mathbf{X}$, where $\mathbf{A}$ and $\mathbf{D}$ is the degree and adjacency matrix of $\mathbf{X}$, $\mathbb{1}_{4\times4}$ is a matrix composed of ones, and $\otimes$ and $\circ$ denote the Kronecker and Hadamard products respectively [4]. We note that $\mathbf{U}$ is a basis for the null-space of $\mathbf{L}$. We have that the optimisation problem for absolute poses solves:

$$\min_{\mathbf{U}^T\mathbf{U}=m\mathbf{I}_4} \left\| \widehat{\mathbf{L}}\mathbf{U} \right\|_F^2,$$

where m denote eigenvalues for associated eigenvectors of $\mathbf{X}$, and F is the Frobenius norm. That is, least-squares is used to solve $\widehat{\mathbf{L}}\mathbf{U} = \mathbf{0}$. As illustrated by Arrigoni et al. [4], outliers are tackled with Iteratively Reweighted Least Squares. We write the resulting absolute poses as $\widehat{\mathbf{R}}_i = [\widehat{\mathbf{R}}_1, \widehat{\mathbf{R}}_2, ..., \widehat{\mathbf{R}}_m]$ and $\widehat{\mathbf{t}}_i = [\widehat{\mathbf{t}}_1^T, \widehat{\mathbf{t}}_2^T, ..., \widehat{\mathbf{t}}_m^T]$.

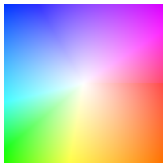

Figure 1. Optical flow field colour scheme. Pure white indicates zero pixel displacement.

1.2. Qualitative Performance Videos

Please see the video[1] we have made available for the qualitative performance of our approach on sequence 09 of the KITTI VO benchmark. The bottom two images show the input pair to our network, and the top image is the composition of the middle image warped into the perspective of the bottom with our pose estimation. This result was obtained using the *HEM Train+Test* method where we have fine-tuned our model with the Homography Estimation Module and utilised it at inference time to boost performance. Our method performs very competitively with comparable methods, while only requiring to estimate 9 pose parameters, avoiding the need to estimate tens of thousands of dense depth or optical flow parameters. To the best of our knowledge, *we are the first to leverage the geometry of the road towards relative pose estimation and visual odometry*.

We also provide another video[2] for the same conditions, except we have increased the frame separation from 5 to 10, which is a significantly larger separation in pose. Despite the network not being trained for this much larger frame separation (we trained for upto 5 frames between input images), we can still obtain good relative pose estimation on large segments of road, with dips in performance occurring where the road planarity assumption or appearance becomes challenging. Most comparable approaches for monocular self-supervised relative pose have *so far been highly restricted to temporally close frame-to-frame predictions*. Our approach leverages local geometry roads with a more general parameterisation which allows for high flexibility in the pose between both cameras.

In future work, we plan to investigate further how far we can take this flexibility and whether we can apply a more complex geometric model to assist pose estimation. Further, we plan to train on data where the camera pairs are in unusual relative positions (e.g. two cameras facing towards each other, at opposite ends of a junction), and whether this could facilitate more unusual applications, such as pose estimation between two different vehicles.

1.3. Additional Qualitative Results

In Figures 2 and 3 we show additional visual results for the pre-training perceptual loss stage. In the last example

[1] https://youtu.be/VrLbDH8LTFc (accessed 30/08/22)
[2] https://youtu.be/DtA6ll8NtSg (accessed 30/08/22)

for Fig. 2, accuracy is likely reduced due to a vehicle in reverse. Preceding examples show that our method is often able to outperform the pseudo ground truth, where we did not have knowledge of the camera-to-road distance and orientation, and thus assumed the calibrated values.

In Fig. 3 we show further strengths and limitations of our initial phase of training - these are perspective warp compositions for sequences 13, 15 and 16 (ground truth is unavailable for these sequences), where we show compositions at two sequential and close time frames in a driving video. Example one shows that our initial phase is able to predict overall relative pose accurately, despite little road being present in one of the image pairs (this example in Fig. 4 is for the second right-hand corner from the top-left, for sequence 13). Hence, we are able to handle cornering and outlying examples approximately in preparation for the refinement stage of our method. Example two and three illustrate cases where our primary assumption, that the road is locally planar, is challenged by quick changes in road gradient (e.g. speed bumps or cresting hills). For examples at either side of these outliers, our method performs well and gross trajectories appear robust in these regions (see Fig. 4 and KITTI sequence 13 for these examples). Example four illustrates a fail case for the initial phase. This could be due to mismatching from similar repetitive features such as lines, faster speeds and dynamic shadows from trees.

We observe that our initial phase of training can struggle on faster parts of the road. The KITTI visual odometry sequences are somewhat biased towards slower urban speeds. The network independent modeling we achieve from our homography estimation module is advantageous for tackling such bias. For the next examples we illustrate that initial training is robust in the presence of dynamic vehicles, narrow roads and cluttered scenes.

1.4. Localised Trajectory Evaluation

The trajectories in our main paper are computed over entire sequences. In practice, visual odometry will drift significantly over long sequences and therefore in Fig. 4 we provide evaluation over smaller sections of road (approximately 200 frames each) for our pre-training phase. We note that for sequences 11, 12, 13, 14 and 15, the ground truth is taken from the KITTI visual odometry benchmark. Evaluating over sequences 03, 09, 10, 11, 12, 13, 14 and 15 respectively, we note that overall our method performs very well on these smaller sub-sections.

1.5. Segmentation and Optical Flow Performance

In Figs. 5, 6 and 7 we show the performance of the semantic segmentation and FlowNet2 on the visual output of our initial phase of training. On the left column we illustrate the input $(\mathbf{I}_{i\to j}, \mathbf{I}_j)$ to the optical flow and on the right column we show the segmented optical flow results for the road-plane region. Fig. 1 illustrates the optical flow vector field colour scheme.

We note that the flow in the road region is largely coherent. In Fig. 6 we show in the first two examples where the optical flow is focused on regions with dynamic shadows. Shadows from moving vehicles will not be correctly cross-projected and represents a limitation in our modeling. However, our network is still able to produce close alignments in the presence of such noise and we suggest that automatic identification of features such as dynamic shadows could be a useful side-affect of our method. Additionally, we note that we generally see more misalignment in the background road-plane, which is reflected in many of the optical flow visualisations. Overall, the segmented optical flow performs very well on our predicted compositions, regardless of the shape of the warps or the content of the images, thus our refinement stage is not heavily limited by the optical flow and segmentation methods which we chose to utilise.

1.6. Architecture and Training Details

We refer to the network components in Fig. 1 (the feature extraction, matching and regression blocks). Using a geometric matching architecture by Rocco et al. [25] we estimate ground-relative pose from overlapping image-pairs. For the feature extraction network we used ResNet101 to estimate feature maps for each of the input views separately. Subsequently L2-normalisation is performed across the feature channel dimension. The matching component calculates similarity scores between both feature maps to form a correlation volume, which is followed by ReLU and L2-normalisation. The regression stage takes these putative matches to regress our ground-relative pose parameters, and is constructed by two successstive 2D convolutional layers (with batch nornalisation and ReLU). Three fully connected layers compute our nine ground-plane pose parameters.

The regression network is initialised with default values, and feature extraction parts use ImageNet weights initially. Our optimiser used SGD with a learning rate of 10^{-4} and a batch size of 16. The weights of the pretrained VGG-16 network are frozen. For the second phase of fine-tuning with L_{HEM} we also use priors to resolve scale ambiguity but only utilise this for the height of camera i.

For the network input we used 240^2 resolution and in Eqn. (8) for scales $s = 1, 2$ we used $120^2, 240^2$ respectively. Most pose estimators concatenate input images channel-wise but we put them separately into the feature extraction backbones in a siamese fashion, and fusing of feature maps only occurs in the correlation volume. For convolution layers in the regression component we use kernel sizes of 7 and 5, with 128 and 64 output channels respectively. Excluding the last output dimension, we used a size of 5000 for the fully connected layers' input and output feature dimensions.

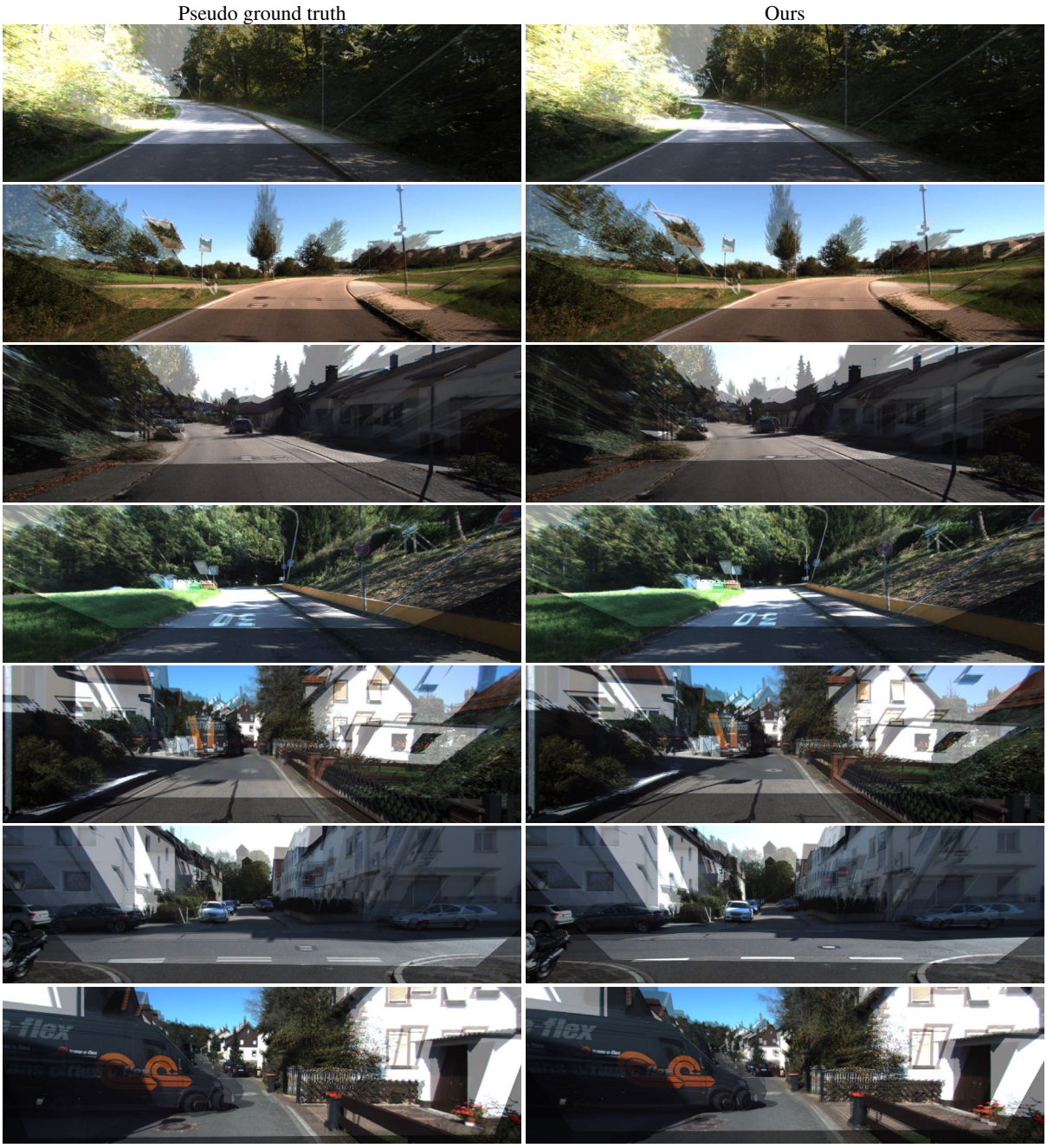

Figure 2. Further KITTI qualittative results for our pre-training phase. Images are a composition of one network input with its counterpart warped with the estimated relative pose. For pseudo ground truth we assume that the road plane is orientated as per the default calibrated camera values. Overall we show accurate results for subsequent refinement, even in the presence of cluttered scenes and sharp motions.

Figure 3. We illustrate further compositions for temporally close image-pairs n and $n + x$ with inference on our initial pre-training phase model. Each image is composed of a source image with its warped counterpart, and the grountruth for these images are unavailable as we use the KITTI visual odometry test set. Example one illustrates a cornering example where little road is captured in the initial pair, yet approximate motion is retained, and useful for our refinement stage. Example two and three show outlying examples where our local planarity assumption is challenged by sharper transitions between planes. Example four shows a limitation where performance is lower due to higher speeds (see text for further discussion). The final examples illustrate good overall performance in various road and motion conditions, such as cluttered scenes and cornering.

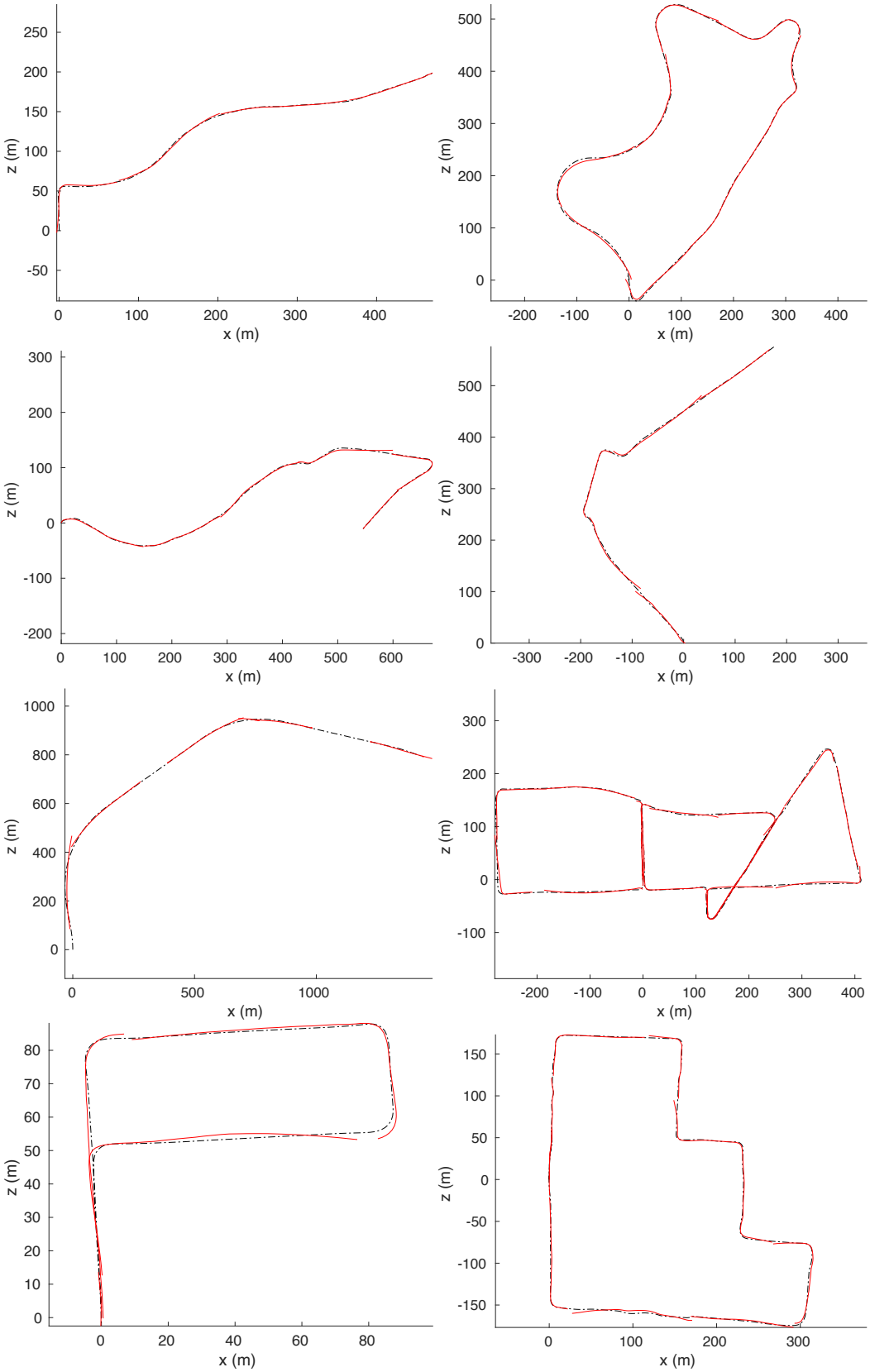

Figure 4. We evaluate our pre-trained model for trajectory on short-subsections (200 frames) for sequences 03, 09, 10, 11, 12, 13, 14 and 15 respectively. We note that we perform particularly well on the challenging sequence 13 where local planarity is challenged significantly.

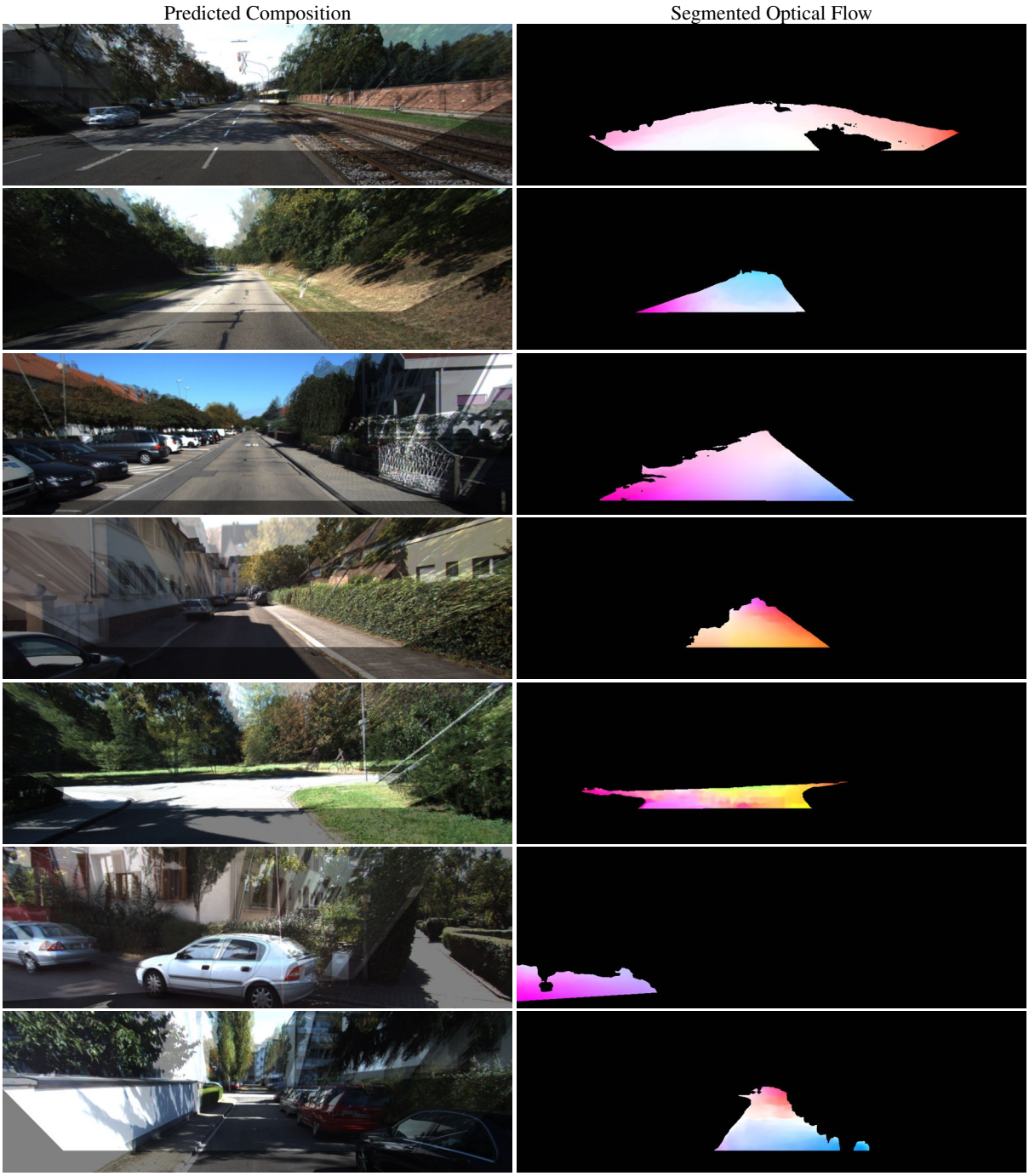

Figure 5. Pre-training performance with optical flow and segmentation accuracy. Left: composition of one network input with the warped counterpart based on the pre-training phase relative pose prediction. Right: accuracy of the segmented optical flow on these predicted initial compositions. Overall the segmented optical flow performs very well on compositions of all shapes, sizes and content.

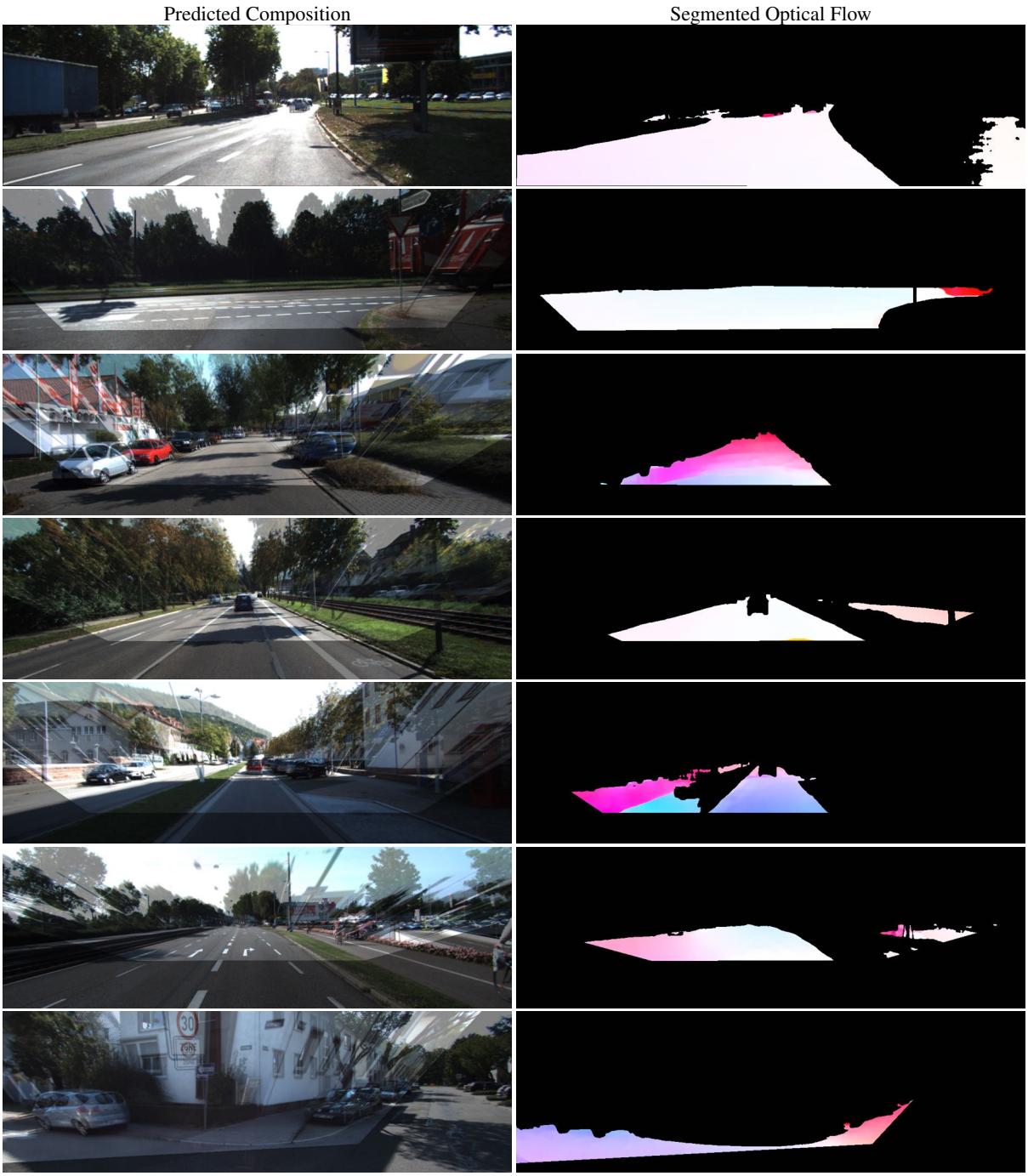

Figure 6. Pre-training performance with optical flow and segmentation accuracy. Left: composition of one network input with the warped counterpart based on the pre-training phase relative pose prediction. Right: accuracy of the segmented optical flow on these predicted initial compositions. Overall the segmented optical flow performs very well on compositions of all shapes, sizes and content. We note first two examples here where dynamic shadows from vehicles cause error in our warped alignment, which is a limitation of our current approach.

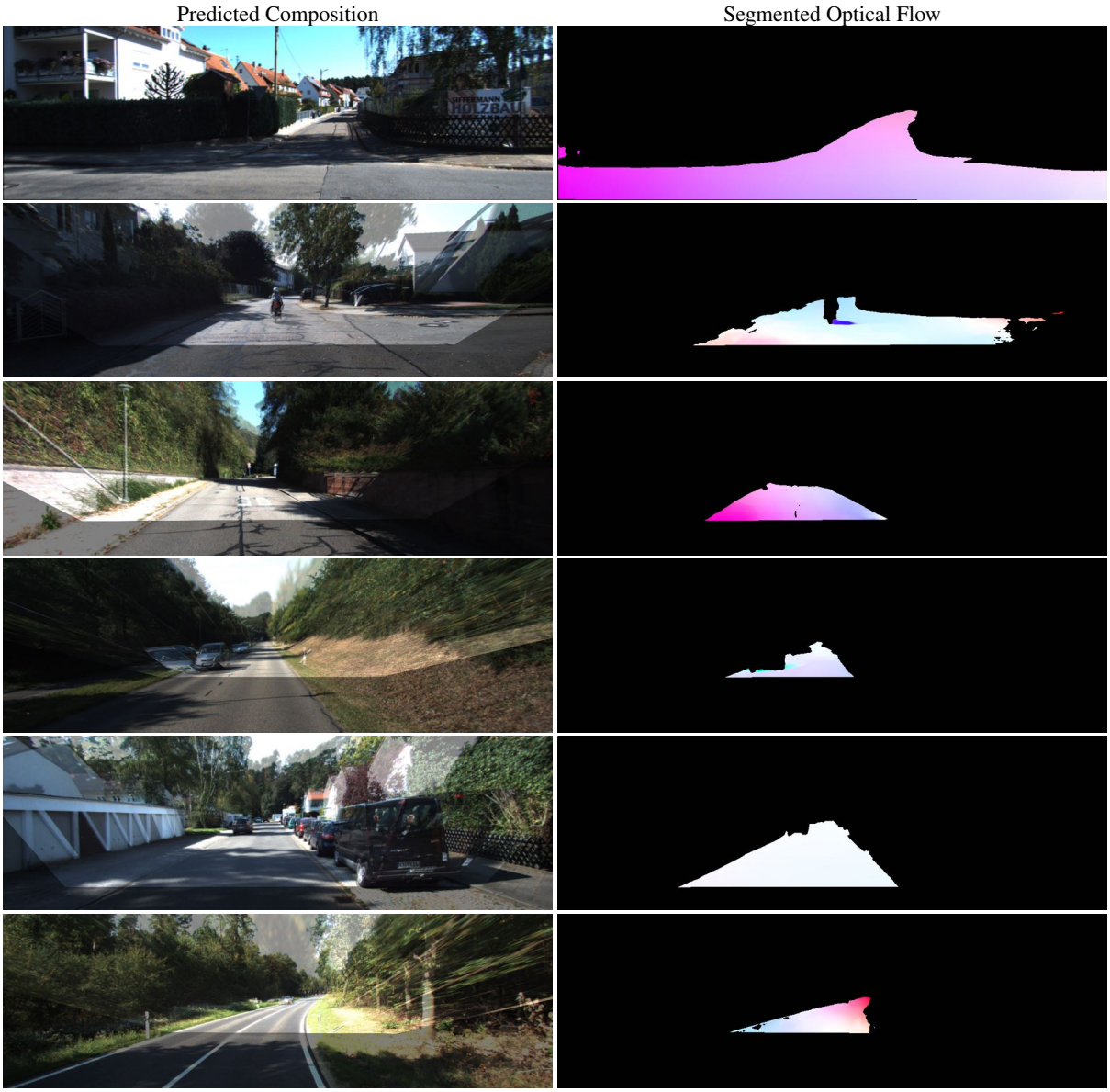

Figure 7. Pre-training performance with optical flow and segmentation accuracy. Left: composition of one network input with the warped counterpart based on the pre-training phase relative pose prediction. Right: accuracy of the segmented optical flow on these predicted initial compositions. Overall the segmented optical flow performs very well on compositions of all shapes, sizes and content.