DirectShape: Direct Photometric Alignment of Shape Priors for Visual Vehicle Pose and Shape Estimation

Rui Wang[1], Nan Yang[1], Jörg Stückler[2], Daniel Cremers[1]

***Abstract*— Scene understanding from images is a challenging problem encountered in autonomous driving. On the object level, while 2D methods have gradually evolved from computing simple bounding boxes to delivering finer grained results like instance segmentations, the 3D family is still dominated by estimating 3D bounding boxes. In this paper, we propose a novel approach to jointly infer the 3D rigid-body poses and shapes of vehicles from a stereo image pair using shape priors. Unlike previous works that geometrically align shapes to point clouds from dense stereo reconstruction, our approach works directly on images by combining a photometric and a silhouette alignment term in the energy function. An adaptive sparse point selection scheme is proposed to efficiently measure the consistency with both terms. In experiments, we show superior performance of our method on 3D pose and shape estimation over the previous geometric approach and demonstrate that our method can also be applied as a refinement step and significantly boost the performances of several state-of-the-art deep learning based 3D object detectors. All related materials and demonstration videos are available at the project page https://vision.in.tum.de/research/vslam/direct-shape.**

I. INTRODUCTION

3D scene understanding is a fundamental task with widespread applications in robotics, augmented reality and autonomous driving. For autonomous vehicles it is critical to observe the poses and 3D shapes of other cars for navigation planning and control. Yet the inference of such object properties from images is a challenging task due to camera projection, variability in view-point, appearance and lighting condition, transparent or reflective non-lambertian surfaces on cars, etc. The community therefore has so far mainly cogitated upon estimating bounding boxes which only contain coarse information on the object poses and sizes. Although with the advances in computer vision 2D object detection has gradually evolved to delivering finer grained results such as instance segmentations, 3D methods are still focusing on estimated bounding boxes.

In this paper, we address joint 3D rigid-body pose estimation and shape reconstruction from a single stereo image pair using 3D shape priors, as illustrated in Fig. 1. Shape priors allow for confining the search space over possible shapes and make vision based pose and shape estimation more robust to effects such as occlusion, lighting conditions or reflective surfaces. We use volumetric signed distance functions (SDF) to implicitly represent the shape of exemplar car models and a linear low-dimensional subspace embedding of the shapes. While previous works made it possible to align the 3D shape geometrically to point clouds estimated by stereo reconstruction [1], [2], we infer shapes and poses directly from images, thus avoid introducing the errors from stereo matching into the pipeline. In our case, the shape prior is aligned with detected cars in stereo images using photometric and silhouette consistency constraints which we formulate as a non-linear least squares problem and optimize using the Gauss-Newton method. Experiments demonstrate superior performance of our method over the previous approach that uses geometric alignment with dense stereo reconstructions. Moreover, as learning based 3D object detectors have become more popular, we also show that our method can be applied as a refinement step that significantly boosts the performances of all the tested methods.

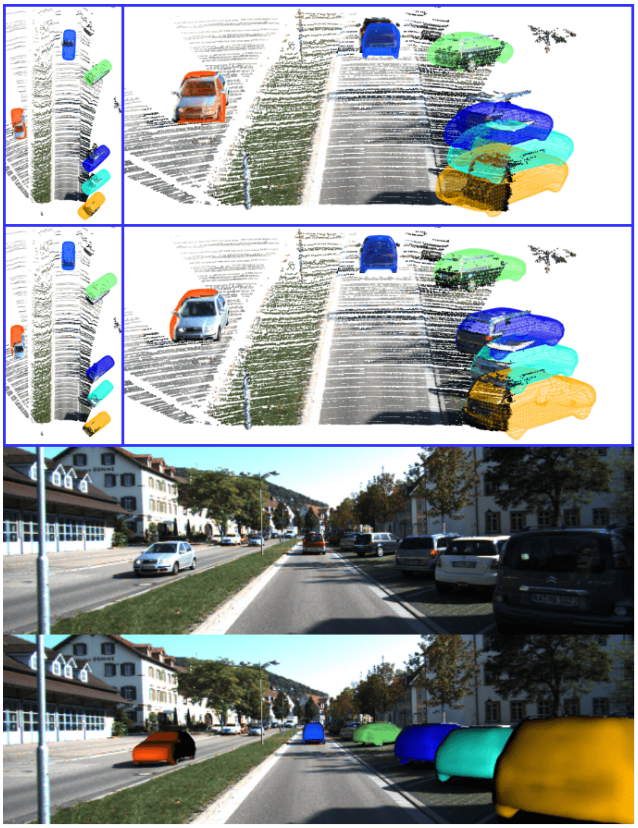


Fig. 1: We propose to jointly estimate 3D vehicle poses and shapes directly on image intensities using shape priors. Top: The initial and our estimated 3D poses and shapes in 3D. Note that the LiDAR point clouds are only for visualization and are not used in our method. Bottom: The input image and its overlay with our results.

[1]R. Wang, N. Yang and D. Cremers are with the Department of Computer Science, Technical University of Munich, Garching bei München, 85748, Germany and Artisense Corporation, 350 Cambridge Avenue 250, Palo Alto, CA 94306, USA. {wangr, yangn, cremers}@in.tum.de

[2]J. Stückler is with Max Planck Institute for Intelligent Systems Tübingen, Tübingen, 72076, Germany. joerg.stueckler@tuebingen.mpg.de

In summary, our contributions are:

- A novel approach for joint 3D pose and shape estimation that delivers more precise and fine grained results than 3D bounding boxes that are commonly estimated by most of the current methods.
- Our approach works directly in image space and thus avoids introducing errors from intermediate steps. It delivers superior performance over the previous approach that uses a geometric formulation for alignment.
- A fully differentiable formulation that operates directly between image and SDF based 3D shape embedding.
- Our method can be applied together with state-of-the-art learning based approaches and significantly boosts the performances of all the tested methods.

II. Related Work

3D object detection. Many successful object detectors have been focusing on localizing objects by 2D bounding boxes [3]–[7] and later by segmentation masks [8]–[10] in images. As object detection in 2D matures, the community starts to target at the much more challenging 3D object detection task [11]–[19]. Chen et al. [12] present 3DOP that first generates 3D object proposals in stereo reconstructions and then scores them in the 2D image using Faster-RCNN [4]. By combining the 3D orientations and dimensions regressed by a network with 2D geometric constraints, Mousavian et al. [15] largely improve the stability and accuracy of 3D detection. While recent deep learning based methods continue to boost the quality of the estimated 3D bounding boxes [16]–[18], Kundu et al. [19] first propose to use a network to regress the 3D pose and shape at the same time, yet the shape is not evaluated in their paper. At the current stage, learning based approaches still can only provide a coarse estimate of the object pose and shape. Moreover, it is an open research question how to assess the quality of learning based detections. Optimization based methods can refine coarse detections and introspect the quality of the fit of the measurements to the model. We believe that is where they come to the stage. Based on the coarse 3D poses estimated by the learning based approaches, our optimization pipeline jointly refines the poses and estimates precise 3D shapes, which contain much more information than 3D bounding boxes and we believe are more useful for applications such as obstacle avoidance and new view synthesis.

3D scene understanding. The availability of large-scale 3D model databases, capable 3D object detectors and fast rendering techniques have spawned novel interest in the use of geometric methods for 3D scene understanding. Salas-Moreno et al. [20] integrate object instances into RGB-D SLAM. The object instances are included as additional nodes in pose graph optimization which finds a consistent camera trajectory and object pose estimate. Geiger and Wang [21] infer 3D object and scene layout from a single RGB-D image by aligning CAD object models with the scene. The approach in [22] detects cars using a CNN-based detector, estimates dense depth using multi-view stereo and aligns a 3D CAD model to the detected car using depth and silhouette constraints. Closely related to our approach, Engelmann et al. [1], [2] use 3D shape embeddings to determine pose and shape of cars which are initially detected by 3DOP [12]. They also embed volumetric signed distance fields (SDF) of CAD models using PCA and formulate a non-linear least squares problem. Their data term, however, relies on a dense stereo reconstruction and measures the distance of reconstructed points to the object surface. It is thus susceptible to the errors of the black-box stereo reconstruction algorithm. Our approach does not require dense stereo matching but directly fits 3D SDF shape embeddings through photometric and silhouette alignment to the stereo images. While silhouette alignment has been used previously to align 3D object models [23]–[25], these methods mainly focus on controlled settings such as only one dominated object appears in the image. By combining silhouette alignment with photometric alignment and explicitly addressing occlusions, we target at the much more challenging real-world traffic scenarios.

III. Proposed Method

A. Notations

Throughout this paper, $\mathbf{p}$ and $\mathbf{X}$ respectively denote image pixels and 3D points. Subscripts o and c define coordinates in object and camera coordinate system. 3D rigid body transformations $\mathbf{T}_a^b = [\mathbf{R}_a^b, \mathbf{t}_a^b; 0, 1] \in \mathrm{SE}(3)$ transforms coordinates from system a to system b, where $\mathbf{R}$ and $\mathbf{t}$ are the 3D rotation matrix and translation vector. In our optimization, 3D poses are represented by their twist coordinates in Lie-algebra $\boldsymbol{\xi} \in \mathfrak{se}(3)$ and 3D shapes are represented by a SDF voxel grid $\boldsymbol{\Phi}$. PCA is performed to embed 3D shapes into a low dimensional space [1], [2] and thus each shape can be represented by $\boldsymbol{\Phi}(\mathbf{z}) = \mathbf{V}\mathbf{z} + \boldsymbol{\Phi}_{mean}$, where $\mathbf{V}$ is the transpose of the subspace projection matrix, $\mathbf{z}$ the shape encoding vector and $\boldsymbol{\Phi}_{mean}$ the mean shape of the gathered object set. The SDF value of a location within the 3D grid is obtained by trilinear interpolation to achieve subvoxel accuracy. While more sophisticated nonlinear shape encodings like Kernel PCA and GP-LVM [26]–[31] exist, we find for cars the PCA model is sufficient. Nevertheless, our formulation is agnostic to the shape representation, thus can be easily adopted for other SDF based shape encodings.

B. System Overview

An overview of our system is shown in Fig. 2. Given a stereo frame with the object segmentations in both images, the quality of the current estimate of the object pose and shape can be measured by two energy terms: (1) by projecting the current shape to both images, a silhouette alignment term E_{silh} measures the consistencies of the projections with the corresponding segmentation masks; (2) object pixels in the left image are warped to the right image and the photometric consistency term E_{photo} measures the color differences. Based on domain knowledge, we add priors terms on the object pose and shape. Our final energy function combines the terms above and is optimized using the Gauss-Newton method. In the following we present the details of each energy term.

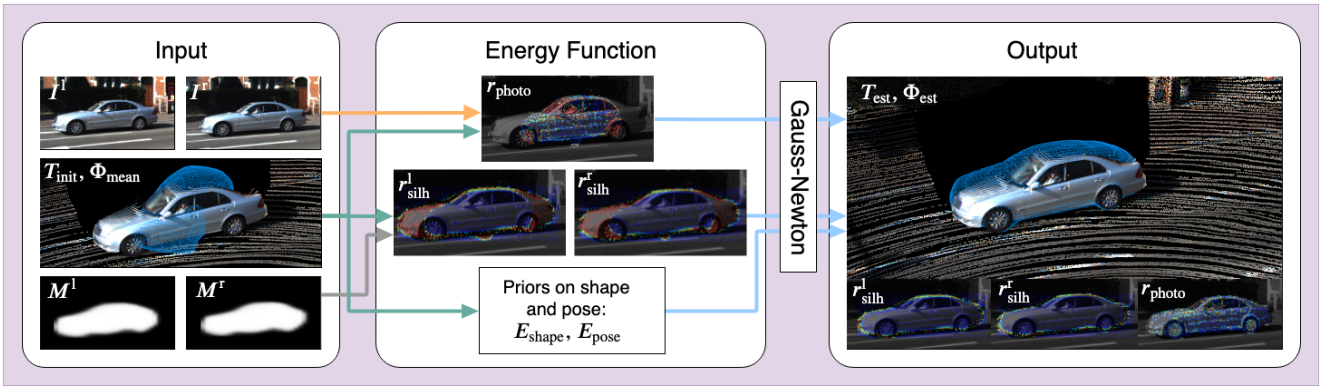


Fig. 2: System overview. As input our method takes a stereo frame I^l, I^r, an initial object pose T_{init}, the learned mean shape Φ_{est}, and the object segmentation masks M^l, M^r. Based on the current pose and shape, the object is projected to I^l and I^r and the consistencies between the projections and the segmentation masks are measured by the silhouette alignment residuals r_{silh} (Sec.III-C). Meanwhile, the object pixels in I^l can be warped to I^r. The color consistencies are measured by the photometric consistency residuals r_{photo} (Sec.III-D). The two terms together with the prior terms (Sec.III-E) are formulated as a non-linear energy function and optimized using the Gauss-Newton method. As output, our method delivers refined object pose T_{est} and shape Φ_{est}.

C. Silhouette Alignment Term

The silhouette alignment term measures the consistency between the image segmentation masks $\mathbf{M}^{l/r}$ and the object masks obtained by projecting the 3D SDF shape embedding $\mathbf{\Phi}$ into the images based on its current shape and pose estimate. Denoting the value of the shape projection mask at pixel $\mathbf{p}$ by $\pi(\mathbf{\Phi}, \mathbf{p})$ (details later) which holds values close to 1 inside and 0 outside the object, the consistency with $\mathbf{M}^{l/r}$ can be expressed by

$$E_{silh}^{l/r} = \frac{1}{|\Omega|} \sum_{\mathbf{p} \in \Omega} r_{silh}^{l/r}(\mathbf{p}), \tag{1}$$

$$r_{silh}^{l/r}(\mathbf{p}) = -\log\big(\pi(\mathbf{\Phi}, \mathbf{p}) p_{fg}(\mathbf{p}) + (1 - \pi(\mathbf{\Phi}, \mathbf{p})) p_{bg}(\mathbf{p})\big), \tag{2}$$

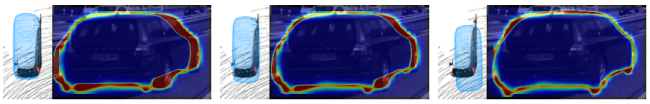
where Ω is the set of the pixels of this object instance, p_{fg} and p_{bg} are the foreground and background probabilities from $\mathbf{M}^{l/r}$. Ideally, if at $\mathbf{p}$ the shape projection coincides with the object segmentation, the value inside the log is close to 1, leading to a small silhouette alignment residual $r_{silh}(\mathbf{p})$; Otherwise the value inside the log is a positive number close to 0, resulting in a large $r_{silh}(\mathbf{p})$. Examples of the silhouette alignment residuals in the left and right images are shown in the middle of Fig. 2 (higher residuals are denoted in red).

Our requirement on the shape projection function $\pi(\mathbf{\Phi}, \mathbf{p})$ is its differentiability wrt. $\mathbf{\Phi}$ and $\mathbf{p}$. Inspired by [30], [32], we define it as

$$\pi(\mathbf{\Phi}, \mathbf{p}) = 1 - \prod_{\mathbf{X}_o} \frac{1}{e^{\mathbf{\Phi}(\mathbf{X}_o)\zeta} + 1}, \tag{3}$$

where $\mathbf{X}_o$ are sampled 3D points along the ray through the camera center and $\mathbf{p}$, ζ is a constant that defines the smoothness of the projection contour.

It is worth noting that the silhouette alignment term only confines the projections of 3D shapes to 2D silhouettes, which is not sufficient to resolve the ambiguity between 3D shapes and poses. Especially in our single-frame stereo setting, even with class-specific priors regularizing the estimated pose, the 3D model often drifts away to better fit the 2D silhouettes, as illustrated in Fig. 3. We thus propose in the next section to further enforce photometric consistency



(a) Iteration 1. (b) Iteration 2. (c) Iteration 3.

Fig. 3: Ambiguity between 3D shape and pose when using only the silhouette alignment term. While fitting the shape projection (harder contour) to the object segmentation (softer contour), the 3D pose drifts away as shown in the bird's-eye view on the left.

to favor poses and shapes that give less color inconsistencies between the left and right images.

D. Photometric Consistency Term

By warping the object pixels from the left image I^l to the right I^r, the photometric consistency term measures the color consistencies. For each pixel within the shape projection, we determine the depth to the object surface through raycasting and finding the intersection with the zero-level set in the SDF. Using this depth and the current object pose, the pixels are transformed from I^l to I^r. Under the brightness constancy assumption, when the pose and shape estimates of an object are correct, the corresponding pixel intensities in the two images should be the same. Our photometric consistency term is formally defined as:

$$E_{photo} = \frac{1}{|\Omega'||N_{\mathbf{p}}|} \sum_{\mathbf{p} \in \Omega'} \sum_{\tilde{\mathbf{p}} \in N_{\mathbf{p}}} \omega_{\mathbf{p}} ||r_{photo}(\tilde{\mathbf{p}})||_\gamma, \tag{4}$$

$$r_{photo}(\tilde{\mathbf{p}}) = \mathbf{I}_r\big(\Pi_c(\mathbf{R}_l^r \Pi_c^{-1}(\tilde{\mathbf{p}}, d_{\mathbf{p}}) + \mathbf{t}_l^r)\big) - \mathbf{I}_l\big(\tilde{\mathbf{p}}\big), \tag{5}$$

where Ω' is the set of the pixels that have intersecting rays with the current shape surface, $N_{\mathbf{p}}$ is a small image neighborhood around $\mathbf{p}$. For each pixel $\tilde{\mathbf{p}}$ in I^l, we warp it to I^r based on the current depth of its central pixel $d_{\mathbf{p}}$ and the relative 3D rotation $\mathbf{R}_l^r$ and translation $\mathbf{t}_l^r$. $\Pi_c(\cdot)$ and $\Pi_c^{-1}(\cdot)$ are the camera projection and back-projection functions that transform 3D coordinates to pixel coordinates and vice versa. The photometric residual $r_{photo}(\tilde{\mathbf{p}})$ is guarded by the Huber norm $||\cdot||_\gamma$ and an image gradient based weighting $\omega_{\mathbf{p}} = c^2/(c^2 + ||\nabla \mathbf{I}_l(\mathbf{p})||_2^2)$, where c is a constant and $\nabla \mathbf{I}_l(\mathbf{p})$ is the image gradient at $\mathbf{p}$. An example of the photometric consistency residuals are shown in the middle of Fig. 2.

The idea behind the photometric consistency term is analogous to direct image alignment applied in recent direct visual odometry (VO) and SLAM methods [33]–[36]. The difference is that instead of directly optimizing for the depth of each pixel independently, in our case the pixel depths are implicitly parameterized by the object pose and shape, i.e., $d_{\mathbf{p}} = d(\mathbf{p}, \mathbf{z}, \mathbf{T}_c^o)$, which brings challenges when deriving the derivative of r_{photo} wrt. $\mathbf{z}$ and $\mathbf{T}_c^o$. Nevertheless, the analytical Jacobians are still achievable and the thorough derivations are provided on our project page.

E. Prior Terms

As cars can only locate on road surface and rotate along the axis that is perpendicular to the road, we encode this domain knowledge with two priors for the pose estimation. Besides, since cars cannot have randomly diverse shapes, we further regularize the estimated shape to be close to our mean shape. Our prior term is therefore defined as:

$$E_{prior} = \lambda_1 E_{shape} + \lambda_2 E_{trans} + \lambda_3 E_{rot}, \tag{6}$$

$$E_{shape} = \sum_{i=1}^{K} (\frac{z_i}{\sigma_i})^2, \tag{7}$$

$$E_{trans} = (\mathbf{t}_o^c(y) - g(\mathbf{t}_o^c(x,z))(y))^2, \tag{8}$$

$$E_{rot} = (1 - (\mathbf{R}_o^c[0,-1,0]^\top)^\top \mathbf{n}_g)^2, \tag{9}$$

where $\lambda_{1,2,3}$ are scalar weighting factors, σ_i is the Eigen value of the i-th principal component; $g(\mathbf{t}_o^c(x,z))(y)$ is the height of the road plane at position $\mathbf{t}_o^c(x,z)$ so that E_{trans} pulls the bottom of the car close to the ground plane; $\mathbf{R}_o^c[0,-1,0]^\top$ is the direction vector of the negative object y-axis and $\mathbf{n}_g$ is the normal vector of the ground surface. E_{rot} penalizes a large difference between the two directions.

F. Adaptive Point Sampling

In previous works, the silhouette alignment term was computed densely for all the pixels on GPUs [32]. While the same implementation principal can be applied to the photometric consistency term, we observe that the dense pixel field contains highly redundant information that contributes only minor to both terms. Recent direct VO methods adopt the idea to sample pixels with sufficient gradients and meanwhile favor a more spatially uniform distribution [35]–[38]. Besides reducing the computational burden, this strategy also suppresses the ambiguous information being added to the system. We observe that due to the reflections on the car surfaces, this strategy becomes even more relevant in our case and can drastically improve the convergence of the photometric term. One issue, though, is the sampling strategy proposed in [36] is adaptive and sometimes can still give very imbalanced spatial distributions. This is undesired for the silhouette alignment if too few pixels are sampled from the object boundary area, as the corresponding 3D parts will not be well constrained. We thus modify the adaptive sampling in [36] to a two-round pipeline: The image is first discretized into a regular grid and a threshold for each cell is computed based on the gradient magnitudes of the pixels within it. The image is then re-discretized using smaller cell size and

Fig. 4: Adaptive point sampling. Pixels are sampled to meet the desired density for each object, preferring pixels with high image gradient (green) but meanwhile maintaining a close to uniform distribution (red).

pixels with gradients above the threshold are selected (green in Fig. 4). This is identical as in [36]. In the second round, we select the pixel with the highest gradient (red in Fig. 4) for each cell that doesn't get any sample from the previous round. To ensure a consistent density for object instances with different sizes in image, we compute the numbers to sample proportionally to the area of their bounding boxes as $0.05 \times height \times width$.

G. Occlusion Handling

When a car is occluded by other cars (the most common case in traffic scenarios), we can extract an occlusion mask using the bounding box of the occluded car and the segmentation masks of the occluding cars, as illustrated in Fig 5a-5c. To this end, we sort all the cars appearing in the image based on the bottom coordinates of their 2D bounding boxes and thus they are ordered roughly according to their distances to the camera. Then for each car we check if there is any closer car overlapping with it and if yes we extract the occlusion mask according to Fig 5a-5c. The occlusion mask is used in the computations of both the silhouette alignment term and the photometric consistency term to exclude the samples from the occluded area, as shown in Fig 5d and 5e.

H. Optimization

Our final energy function is defined as the weighted sum of the previously defined terms

$$E = \lambda_{silh} E_{silh}^l + \lambda_{silh} E_{silh}^r + E_{photo} + E_{prior}, \tag{10}$$

where λ_{silh} is a scalar weighting factor. Note that all the energy terms have quadratic forms except for E_{silh}, which prevents the application of 2nd-order optimization methods. While in the previous works E_{silh} is typically optimized using 1st-order methods like gradient descent, we reformulate it as an iteratively reweighted least squares problem as

$$E_{silh}^{l/r} = \frac{1}{|\Omega|} \sum_{\mathbf{p} \in \Omega} \omega'_{\mathbf{p}} (r_{silh}^{l/r}(\mathbf{p}))^2, \tag{11}$$

where $\omega'_{\mathbf{p}} = 1/r_{silh}(\mathbf{p})$ is recalculated in each iteration based on the value of $r_{silh}(\mathbf{p})$ of the current iteration. E is thus optimized using Gauss-Newton for the variables $[\boldsymbol{\xi}_c^o; \mathbf{z}]$, where $\boldsymbol{\xi}_c^o$ are the twist coordinates of the 3D rigid-body pose of the object in the camera coordinate system and $\mathbf{z}$ is the shape encoding vector. It is worth pointing out that previous works [19], [39] stated that the process of rendering from 3D shape to image is not differentiable due to the non-linearity and hidden relationship between the two domains, and thus opt for workarounds such as finite difference. We claim that

(a) Input image with bounding box. (b) Segmentation masks and bounding box. (c) Occlusion mask. (d) Silhouette alignment residuals. (e) Photometric consistency residuals.

Fig. 5: Occlusion handling. For each object detection, we check if its 2D bounding box is overlapped by the segmentation mask of any other object that is closer to the camera (5b). Such overlapping part is considered as the occlusion mask (5c) and is used in the computation of both the silhouette alignment (5d) and the photometric consistency residuals (5e) to exclude pixels from the occluded part.

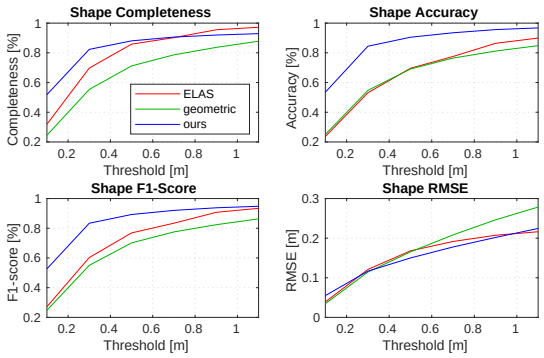


Fig. 6: Quantitative evaluation on shape reconstruction. Our approach outperforms the geometric approach [1] in all measures.

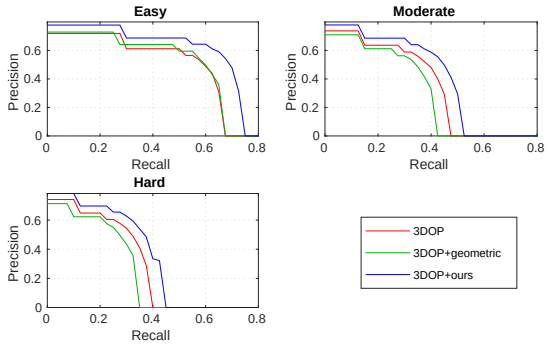


Fig. 7: Quantitative evaluation on pose estimation in comparison to the geometric approach [1]. IoU=0.5 is used for computing these precision-recall curves.

such process is actually fully differentiable. Please refer to our project page for the details of all the analytical Jacobians.

IV. Experiments

Our method aims at estimating more precise and accurate geometric properties than 3D bounding boxes for cars. To demonstrate this ability, we separately evaluate its performances on 3D shape estimation and 3D pose refinement, where the KITTI Stereo 2015 [40] and 3D Object [41] benchmarks are adopted for the two tasks respectively. Throughout all our experiments, Mask-RCNN [9] is used to generate the input object segmentation masks and we set the weighting factors to $\lambda_{silh} = 12$, $\lambda_1 = 10$, $\lambda_2 = 10$ and $\lambda_3 = 10^7$. The CPU implementation of our optimization runs at around 100ms per object and 180ms per frame on KITTI. Porting to GPU may yield further runtime improvement.

A. Shape Estimation

We first compare our method to the previously proposed geometric approach [1], which fits the same PCA model to the point cloud estimated by the dense stereo reconstruction method ELAS [42]. To our knowledge it is so far the only method that provides object shape evaluation. To be consistent with [2], [44], we measure *completeness* (the percentage of ground-truth (GT) points which have at least one estimated point within a certain distance τ), *accuracy* (the percentage of estimated points which have at least one GT point within τ) and F_1 score ($2 \cdot completeness \cdot accuracy/(completeness + accuracy)$) for the GT segments in the KITTI Stereo 2015 benchmark. To more precisely measure the accuracy, we additionally compute the root mean square error (RMSE): For each GT point, we search within τ and compute the distance to the closest point from our estimate. The RMSE is only computed for those GT points which have matched estimated points. These four metrics wrt. different τ are displayed in Fig. 6, where our method outperforms the geometric approach in all the four evaluations[1]. Compared to the laboratorial settings in [23]–[25] where only one dominant car appears in the image, KITTI Stereo 2015 is much more challenging and contains cars with severe truncation, occlusion and at large distance. This explains the degraded performances of the geometric approach, as it lacks occlusion handling and also the stereo reconstruction gets drastically more noisy in faraway areas.

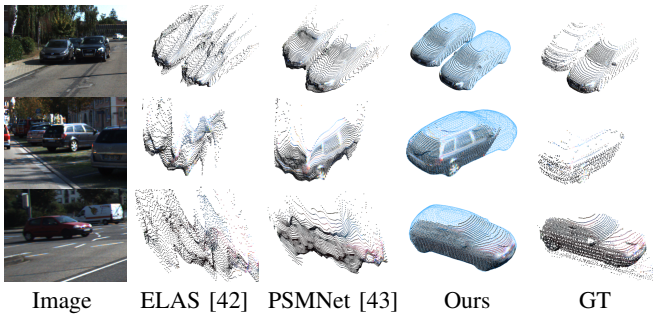


Fig. 8: Qualitative results on 3D shape refinement. We compare our method to a classical (ELAS [42]) and a SotA deep learning based (PSMNet [43]) stereo matching method.

In Fig. 8, we qualitatively compare our method to ELAS [42] and a recent deep learning based stereo reconstruction method PSMNet [43]. Although deep learning with strong supervision has significantly improved the reconstruction quality, it still suffers at large distances. The results in Fig. 8 validate the idea of introducing shape priors into the pipeline. More qualitative results can be found in Fig. 10.

[1]When comparing to Fig. 9 in [2], it is worth noting that the results there are obtained from a subset of KITTI Stereo 2015.

Method	AP_{bv} (IoU=0.5)			AP_{bv} (IoU=0.7)			AP_{3D} (IoU=0.5)			AP_{3D} (IoU=0.7)		
	Easy	Mode	Hard	Easy	Mode	Hard	Easy	Mode	Hard	Easy	Mode	Hard
Mono3D [14]	11.70	9.62	9.32	2.06	1.91	1.39	9.55	7.72	7.23	0.62	0.75	0.76
Mono3D + Ours	**23.53**	**16.54**	**15.30**	**5.21**	**4.02**	**3.84**	**18.88**	**14.31**	**11.73**	**2.61**	**2.09**	**2.17**
Deep3DBox [15]	29.99	23.74	18.81	9.96	7.69	5.29	26.94	20.51	15.85	5.82	4.08	3.83
Deep3DBox + Ours	**44.89**	**29.99**	**24.41**	**12.35**	**8.88**	**7.49**	**38.40**	**25.39**	**20.02**	**6.50**	**4.38**	**4.04**
3DOP [12]	48.73	35.20	30.95	12.63	9.07	7.12	40.76	28.92	24.31	5.38	3.76	3.25
3DOP + Ours	**59.40**	**39.43**	**33.54**	**19.98**	**13.40**	**11.34**	**50.16**	**34.66**	**29.31**	**11.38**	**7.36**	**6.34**
MLF [16]	55.03	36.73	31.27	22.03	13.76	11.60	47.88	29.48	26.44	10.53	5.69	5.39
MLF + Ours	**63.10**	**37.97**	**31.84**	**25.58**	**15.25**	**11.97**	**55.12**	**34.78**	**29.44**	**14.59**	**8.42**	**7.26**

TABLE I: Average precision of bird's eye view (AP_{bv}) and 3D bounding boxes (AP_{3D}), evaluated on the KITTI 3D Object validation set. Note that the KITTI Object Benchmark updated its evaluation script in 2017 which causes some inconsistent numbers in this table and in the original papers.

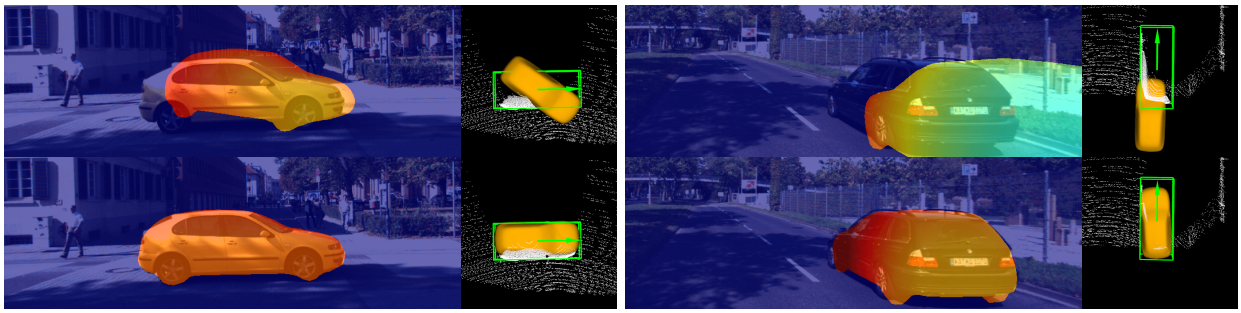


Fig. 9: Qualitative results of 3D pose refinement. Each column shows the initial and the optimized pose (overlapped with the input image and also in bird-eye view). GT poses are denoted by green boxes. Note that point clouds are not used in our optimization.

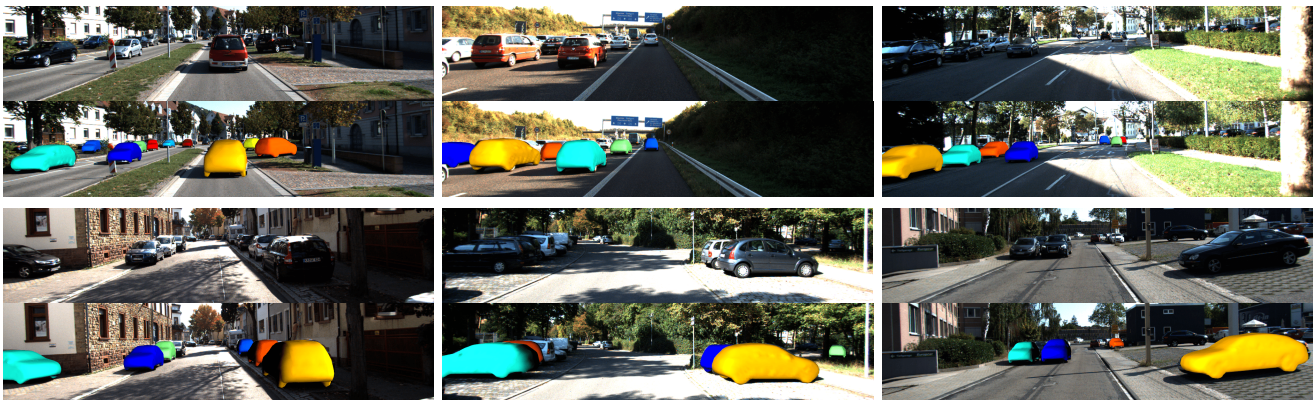


Fig. 10: Qualitative results of 3D shape and pose estimation.

B. Pose Refinement

Same as above, we first compare our method to the geometric approach [1]. Based on the GT provided by KITTI 3D Object, we compute the precision-recall curves for 3D bounding boxes for the three pre-defined difficulties. The results are shown in Fig 7, where our method delivers better 3D pose estimates for all the difficulties. In the categories Moderate and Hard, as many cars are occluded and only part of the object 3D points can be reconstructed, fitting the 3D model to the incomplete point cloud would generally worsen the pose estimation. The qualitative results of our 3D pose refinement on two example images can be found in Fig. 9.

In the next experiment we demonstrate the pose refinements of our method over 4 deep learning based 3D detectors, namely Mono3D [14], Deep3DBox [15], 3DOP [12] and MLF [16]. We use the validation splits provided by [12], [15] and compute the average precisions for bird's eye view (AP_{bv}) and 3D bounding boxes (AP_{3D}). The results in Table I shows that our method hugely boosts the performances of all the tested methods under all the settings, which demonstrates the effectiveness of our method on 3D pose refinement. Some qualitative results of our method in challenging real-world scenarios can be found in Fig. 10.

V. Conclusions

We propose a new approach for joint vehicle pose and shape estimation based on an energy function combining photometric and silhouette alignment. Our method delivers much more precise and useful information than the current 3D detectors that focus on estimating bounding boxes. In our experiments we demonstrate superior performance over the previous geometric method in both pose and shape estimation. We also demonstrate that our approach can significantly boost the performance of learning-based 3D object detectors. In future work, we are planning to extend our approach to a local window of multiple frames and integrate it into a visual SLAM system.

REFERENCES

[1] F. Engelmann, J. Stückler, and B. Leibe, "Joint object pose estimation and shape reconstruction in urban street scenes using 3D shape priors," in *German Conference on Pattern Recognition (GCPR)*. Springer, 2016, pp. 219–230.

[2] ——, "SAMP: Shape and motion priors for 4D vehicle reconstruction," in *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*. IEEE, 2017, pp. 400–408.

[3] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 9, pp. 1627–1645, 2009.

[4] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2015, pp. 91–99.

[5] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Advances in neural information processing systems*, 2016, pp. 379–387.

[6] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 764–773.

[7] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6517–6525.

[8] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[9] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2980–2988.

[10] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár, "Panoptic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9404–9413.

[11] S. Satkin and M. Hebert, "3DNN: Viewpoint invariant 3D geometry matching for scene understanding," in *IEEE International Conference on Computer Vision, (ICCV)*, 2013, pp. 1873–1880.

[12] X. Chen, K. Kundu, Y. Zhu, A. Berneshawi, H. Ma, S. Fidler, and R. Urtasun, "3d object proposals for accurate object class detection," in *NIPS*, 2015.

[13] S. Song and J. Xiao, "Sliding shapes for 3d object detection in depth images," in *European Conference on Computer Vision (ECCV)*, 2014, pp. 634–651.

[14] X. Chen, K. Kundu, Z. Zhang, H. Ma, S. Fidler, and R. Urtasun, "Monocular 3D object detection for autonomous driving," in *IEEE CVPR*, 2016.

[15] A. Mousavian, D. Anguelov, J. Flynn, and J. Kosecka, "3D bounding box estimation using deep learning and geometry," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7074–7082.

[16] B. Xu and Z. Chen, "Multi-level fusion based 3D object detection from monocular images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2345–2353.

[17] P. Li, X. Chen, and S. Shen, "Stereo R-CNN based 3D object detection for autonomous driving," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[18] Y. Wang, W.-L. Chao, D. Garg, B. Hariharan, M. Campbell, and K. Q. Weinberger, "Pseudo-LiDAR from visual depth estimation: Bridging the gap in 3D object detection for autonomous driving," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8445–8453.

[19] A. Kundu, Y. Li, and J. M. Rehg, "3D-RCNN: Instance-level 3D object reconstruction via render-and-compare," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3559–3568.

[20] R. F. Salas-Moreno, R. A. Newcombe, H. Strasdat, P. H. J. Kelly, and A. J. Davison, "SLAM++: Simultaneous Localisation and Mapping at the Level of Objects," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1352–1359.

[21] A. Geiger and C. Wang, "Joint 3D Object and Layout Inference from a single RGB-D Image," in *German Conference on Pattern Recognition (GCPR)*, ser. Lecture Notes in Computer Science, vol. 9358. Springer International Publishing, 2015, pp. 183–195.

[22] R. Ortiz-Cayon, A. Djelouah, F. Massa, M. Aubry, and G. Drettakis, "Automatic 3D Car Model Alignment for Mixed Image-Based Rendering," in *International Conference on 3D Vision (3DV)*, 2016.

[23] R. Sandhu, S. Dambreville, A. Yezzi, and A. Tannenbaum, "A nonrigid kernel-based framework for 2D-3D pose estimation and 2D image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 6, pp. 1098–1115, 2011.

[24] V. A. Prisacariu, A. V. Segal, and I. Reid, "Simultaneous monocular 2D segmentation, 3D pose recovery and 3D reconstruction," in *Proc. of the Asian Conf. on Computer Vision (ACCV)*, 2013.

[25] A. Dame, V. A. Prisacariu, C. Y. Ren, and I. D. Reid, "Dense reconstruction using 3D object shape priors," in *Proc. of the IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[26] S. Dambreville, Y. Rathi, and A. Tannenbaum, "A framework for image segmentation using shape models and kernel space shape priors," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 30, no. 8, pp. 1385–1399, 2008.

[27] R. Sandhu, S. Dambreville, A. Yezzi, and A. Tannenbaum, "A nonrigid kernel-based framework for 2D-3D pose estimation and 2D image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 33, no. 6, pp. 1098–1115, 2011.

[28] V. A. Prisacariu and I. Reid, "Nonlinear shape manifolds as shape priors in level set segmentation and tracking," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 2185–2192.

[29] ——, "Shared shape spaces," in *International Conference on Computer Vision (ICCV)*. IEEE, 2011, pp. 2587–2594.

[30] V. A. Prisacariu, A. V. Segal, and I. Reid, "Simultaneous monocular 2D segmentation, 3D pose recovery and 3D reconstruction," in *Asian Conference on Computer Vision*. Springer, 2012, pp. 593–606.

[31] S. Zheng, V. A. Prisacariu, M. Averkiou, M.-M. Cheng, N. J. Mitra, J. Shotton, P. H. Torr, and C. Rother, "Object proposals estimation in depth image using compact 3d shape manifolds," in *German Conference on Pattern Recognition (GCPR)*. Springer, 2015, pp. 196–208.

[32] A. Dame, V. A. Prisacariu, C. Y. Ren, and I. Reid, "Dense reconstruction using 3D object shape priors," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 1288–1295.

[33] J. Engel, J. Sturm, and D. Cremers, "Semi-dense visual odometry for a monocular camera," in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 1449–1456.

[34] J. Engel, T. Schöps, and D. Cremers, "LSD-SLAM: Large-scale direct monocular SLAM," in *European Conference on Computer Vision (ECCV)*. Springer, 2014, pp. 834–849.

[35] R. Wang, M. Schwörer, and D. Cremers, "Stereo DSO: Large-scale direct sparse visual odometry with stereo cameras," in *International Conference on Computer Vision (ICCV), Venice, Italy*, 2017.

[36] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 40, no. 3, pp. 611–625, 2018.

[37] N. Yang, R. Wang, J. Stuckler, and D. Cremers, "Deep virtual stereo odometry: Leveraging deep depth prediction for monocular direct sparse odometry," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 817–833.

[38] X. Gao, R. Wang, N. Demmel, and D. Cremers, "LDSO: Direct sparse odometry with loop closure," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 2198–2204.

[39] R. Zhu, C. Wang, C.-H. Lin, Z. Wang, and S. Lucey, "Object-centric photometric bundle adjustment with deep shape prior," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 894–902.

[40] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[41] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the KITTI vision benchmark suite," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[42] A. Geiger, M. Roser, and R. Urtasun, "Efficient large-scale stereo matching," in *Asian conference on computer vision*. Springer, 2010, pp. 25–38.

[43] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5410–5418.

[44] C. Zhou, F. Güney, Y. Wang, and A. Geiger, "Exploiting object similarity in 3D reconstruction," in *Proc. of the IEEE Int. Conf. on Computer Vision (ICCV)*, 2015.