A Compacted Structure for Cross-domain learning on Monocular Depth and Flow Estimation

Yu Chen¹, Xu Cao², Xiaoyi Lin³, Baoru Huang⁴, Xiao-Yun Zhou⁵, Jian-Qing Zheng⁶, Guang-Zhong Yang⁷

Abstract—Accurate motion and depth recovery is important for many robot vision tasks including autonomous driving. Most previous studies have achieved cooperative multi-task interaction via either pre-defined loss functions or cross-domain prediction. This paper presents a multi-task scheme that achieves mutual assistance by means of our Flow to Depth (F2D), Depth to Flow (D2F), and Exponential Moving Average (EMA). F2D and D2F mechanisms enable multi-scale information integration between optical flow and depth domain based on differentiable shallow nets. A dual-head mechanism is used to predict optical flow for rigid and non-rigid motion based on a divide-and-conquer manner, which significantly improves the optical flow estimation performance. Furthermore, to make the prediction more robust and stable, EMA is used for our multi-task training. Experimental results on KITTI datasets show that our multi-task scheme outperforms other multi-task schemes and provide marked improvements on the prediction results.

I. INTRODUCTION

3D scene understanding from single view video sequences is an important topic in robot vision, which consists of several classic computer vision tasks, including depth estimation, optical flow estimation, and visual odometry prediction. It plays a key role in many real-world applications, including autonomous driving [1], unmanned aerial vehicles (UAV) [2], and surgical navigation [3], [4].

Depth Estimation. Traditional method referred to as Structure from Motion (SfM) addresses depth and camera pose estimation simultaneously by keypoint matching. However, traditional methods tend to use hand-crafted features and often rely on sparse correspondence of low-level features [5], [6].

More recently, Convolutional Neural Networks (CNN) have shown to be able to establish dense mapping relationships between single view images and the corresponding depth information in a supervised manner [7], [8], [9]. Nevertheless, the requirement of depth annotations is a major challenge for supervised methods [10], [11], [12]. Without any dependence on depth annotations, self-supervised methods [13], [14], [15] are trained with only stereo images by leveraging the multi-view geometric constraints and capable of monocular depth estimation. Zhou et al. [16] proposed
a self-supervised monocular depth estimation paradigm by combining ego pose estimation and depth estimation tasks. By leveraging the appearance similarity indicators calculated by synthetic views [17] from temporal continuous frames and the current frame, these methods show promising performance w.r.t. other self-supervised methods. However, simple appearance loss is insufficient as supervision especially in illumination variations scenes. To this end, Xiong et al. [18] proposed several novel scale-consistent geometric constraints by jointly considering relative forward-backward poses consistency and depth reconstruction error to further improve the depth estimation performance.

**Optical Flow Estimation.** Optical flow estimation was traditionally treated as an optimization problem over the dense displacement fields between temporal adjacent image pairs [19], [20]. The optimization objective consists of two main terms: 1) **data** – that encourages correlations in visually similar regions; 2) **regularization** – that penalizes incoherent motion fields according to the motion priors. Whilst offering good performance in general, the inherent optimization process entails high computational cost making these methods difficult to be used for real-time applications [21].

Recently, CNN-based approaches have shown improved optical flow estimation results [22], [21], [23]. However, these supervised methods require dense per-pixel optical flow annotation, which is laborious. By optimizing photometric consistency and local flow smoothness, unsupervised methods overcome the need for annotations. Several recent methods have greatly improved prediction results by checking forward-backward consistency [24], filtering range map to omit occlusion regions [25] or using loss function based on occlusion-aware bidirectional flow estimation and the robust census transform [26]. Unlike coarse-to-fine optical flow prediction, such as PCW-Net [21], more accurate results have been obtained by methods applying RNN structures to optical flow prediction scenarios [23].

**Unsupervised Multi-task Network Scheme.** Depth and optical flow can be estimated in a joint manner due to their inherent geometric correspondences. Many approaches [27], [28], [29] have predicted depth and optical flow with separate networks trained by geometric constraints based consistency losses. Other methods [30], [31], [32], [33] have combined depth and optical flow networks by serving depth or optical flow estimation results as the 3D information supervision of the other task. Although efficient, the above methods omit the inherent correspondence between latent information of different domains, which also serve as important assistance to multi-task estimation.

To further improve the multi-task network performance, we propose a novel multi-task scheme for depth estimation and optical flow prediction. It enables a bidirectional interaction of latent information between the depth and optical flow networks, enabling mutual assistance between two tasks. Our contributions can be summarised as follows:

- We proposed **D2F** and **F2D** mechanisms, which allow the latent information of depth and optical flow prediction transfer to and assist the estimation of the other task.
- For the former mechanism, we adopt the EMA to the multi-task learning and further improve the performance of cross-task collaboration.
- Dual-head mechanism is designed to enhance the optical flow prediction results by implicitly disentangling the representation of foreground and background.

To prove the effectiveness of our multi-task scheme, experiments are conducted on KITTI [34] datasets. The experiments cover different combinations of D2F, F2D, EMA training approach, and dual-head mechanism based on the baseline scheme in Fig. 1(a). D2F and D2F enable the interchange of latent information between depth and optical flow estimation. Experiment results show that our D2F mechanism can significantly improve the optical flow prediction accuracy. At the same time, the road signs and streetlamps are more clearly told in depth estimation. Experiments also show that our F2D can further improve depth estimation accuracy. We further show that our EMA training approach can improve both depth and optical flow prediction results. Our dual-head mechanism enables the division of the motion field into rigid and non-rigid parts based on a divide-and-conquer manner and improves the optical flow accuracy significantly.

**II. RELATED WORK**

**A. Loss-level Collaboration in Parallel Multi-task Network**

Previous methods achieve cross-task interaction by applying cross-domain loss functions to multi-task learning scheme combined with separate single-task networks, as illustrated in Fig. 1(a). Ranjan et al. [30] introduced a competitive collaboration strategy that can classify moving objects and static background and effectively coordinate the training of the multi-task network. Benefiting from the framework, these tasks with inherent correspondence could reinforce each other. To reduce the violation of non-rigid motion flow caused by moving objects and occluded pixels, Wang et al. [31] constructed various masks for a valid consistency loss calculation, which remarkably improves the estimation of monocular depth, ego pose, and optical flow in temporal consecutive frames. Cao et al. [32] proposed photometric loss accompanied with bundle adjustment modules to achieve more accurate results. Zhao et al. [33] facilitated pose estimation by directly calculating camera pose from optical flow to improve depth estimation and optical flow prediction.

**B. Cross-task Boosting in Multi-task Network**

Other than cross-task collaboration based on the loss functions, other methods took further steps to boost multi-task performance by introducing information interaction between the depth and optical flow estimation network. GeoNet et al. [35], [36] derives rigid flow from depth and pose prediction, which is then fine-tuned by optical flow network as illustrated in Fig. 1(b). By grafting the depth estimation network behind the flow-motion network, the
method proposed by [37] refines the camera poses from GPS, IMU, or odometry algorithms and depth estimation with the confidential map as illustrated in Fig. 1(c). Yang et al. [38] have been estimated depth map based on optical flow from pretrained FlowNet with mid-point triangulation method. DRAFT [39] generates coarse depth prediction by triangulation from optical flow estimated by RAFT [40]. DRAFT then estimates the fine-grained depth and scene flow by joint leveraging coarse depth prediction and pyramid correlation. Different from the methods mentioned above, our cross-task exchange mechanism D2F and F2D enables multi-scale latent information interchange and facilitates multi-level cross-domain collaboration between depth and optical flow estimation.

As for multi-task feature-level collaboration, Chi et al. [41] introduced a feature-level collaboration mechanism for the stereo depth estimation, flow prediction, and pose detection networks as illustrated in Fig. 1(f). Whereas our scheme is dedicated to more challenging monocular depth estimation tasks, rather than stereo depth estimation. Hur et al. [42] proposed an architecture to estimate depth and 3D motion simultaneously from a single decoder. In addition to the previous cross-task boosting scheme, DENAO [43] proposed multi-scale exchange blocks which deeply coupled depth and optical flow branches according to epipolar geometry constraints, as illustrated in Fig. 1(d). The auxiliary optical flow improves the depth predictions and in turn yields large improvements in optical flow accuracy. However, by leveraging known ego camera pose and the proposed epipolar layer, DENAO is a partially supervised multi-task scheme. We proposed a fully unsupervised multi-task scheme with cross-task information exchange blocks. The geometry constraints between optical flow and depth predictions are leveraged implicitly, rather than the explicit 3D-to-2D projection and SVD decomposition-based triangulation in DENAO.

III. METHODOLOGY

Given a pair of consecutive frames \( (I_t, I_{t+1}) \) from an unlabeled monocular video, our method estimates optical flow \( F \) between the two frames \( (I_t, I_{t+1}) \) and the corresponding depth maps \( D_t \) and \( D_{t+1} \).

The overall architecture of the proposed method is illustrated in Fig. 2(a). The multi-task estimation network consists of a shared encoder extracting the common feature domain, a two-branch decoder regress in the depth and optical domain, and exchange blocks for cross-domain message interchange. As described in Sec. III-A the two branches in the cross domain interchange, respectively for estimation of depth and optical flow, collaborate in their tasks by transferring cross-domain latent information from multiple layers to each other.

EMA is embedded into our training framework to improve the depth and optical flow estimation as shown in Fig. 2(b), with the detail described in Sec. III-B. Following mean teacher method [44], the optical flow branch is further divided into student and teacher decoder. The structure of the two optical flow decoders is adopted from PWC-Net [21] with dual-head mechanism (Sec. III-C) incorporated in each decoder layer. The depth branch contains a depth decoder following the decoder structure of monodepth [45].

A separate camera pose prediction network is applied during training for jointly learning depth and camera pose based on geometric consistency.

A. Cross-domain Collaboration

Estimation of depth and optical flow both requires comprehensive understanding of stereo vision. We hypothesize that these two tasks can be solved from a common feature domain and thus utilize only one shared encoder. For further mutually benefiting the two tasks in the exchange blocks, depth and optical flow branches collaborate with each other via D2F and F2D mechanisms. D2F and F2D mechanisms help to refine the prediction from multiple levels of the decoder layers by providing each other with cross-domain information, rather than merely supervising the final prediction of the entire network.

1) D2F mechanism: In solitary optical flow prediction, the network determines the pixel shift by comparing the RGB information of two images. Benefiting from D2F mechanism,
the depth branch provides additional geometric information for optical flow prediction.

D2F mechanism provides depth information of the images for optical flow prediction. As shown in Fig. 3, the $i^{th}$ D2F mechanism takes dual-head masks $M^{i-1}$ received from the $(i-1)^{th}$ decoder layer, encoder feature maps $\Phi^{i}_{en}$, and depth feature maps $\Phi^{i}_{depth}$ from two frames of images as input to generate cost volume $C^{i}_{d}$. Together with the cost volume, the up-sampled optical flow from the $(i-1)^{th}$ decoder layer $F^{i-1}$ is further refined by the $i^{th}$ optical flow decoder layer to generate $F^{i}$.

As the pre-process, the D2F mechanism first generates fused feature map, $\Phi^{i}$ by $\Phi^{i} = \phi^{i} (\Phi^{i}_{en} \oplus \Phi^{i}_{depth})$ where $\oplus$ denotes the channel concatenation and $\phi^{i}$ denotes $1 \times 1$ convolutional layers. $\Phi^{i}$ will be then operated by dual-head mechanism to produce refined feature map $\Psi^{i}$, which is discussed in section III-C.

The cost volume $C^{i}_{d}$ is obtained through the correlation between $\Psi^{i}$ and $\Phi^{i}$ with a limited range of $d$ pixels:

$$C^{i}_{d}(\Psi^{i}, \Phi^{i}) = \frac{1}{N} \left( \Psi^{i} + o \right)^{\top} \left( \Phi^{i} + o \right), o \in [-d,d] \times [-d,d]$$

where $N$ denotes the length of the column vector $\Psi^{i}$.

2) F2D Mechanism: The F2D mechanism improves depth prediction with the stereo features extracted from optical flow.

As illustrated in Fig. 3, the F2D mechanism obtain $F^{i}$ to help improve the depth feature map $\Phi^{i}_{depth}$ and produce the refined depth feature map $\tilde{\Phi}^{i}_{depth}$.

We first normalize the value of $\Phi^{i}_{depth}$ to [0, 1]. The normalized feature maps are then applied to calculate the cost map $C_{0}$ following Equation (1) where $d$ is set to 0.

The output refined depth feature map $\tilde{\Phi}^{i}$ is calculated with $\tilde{\Phi}^{i} = \phi^{3} (\Phi^{i}_{en} \oplus \Phi^{i}_{depth})$ where $\phi^{3}$ denotes the $3 \times 3$ convolutional layer. The refined depth feature map will be obtained by the next depth decoder layer.

B. EMA for Cross-task Prediction

In our model, cross-domain information walks double-sided between depth and optical flow branches. Parameter fluctuation of any of these two branches will directly affect the other one. Moreover, D2F and F2D work at multiple decoder layers. The error of one layer will accumulate at the next layer, and will eventually cause significant bias on the final prediction result. These factors might cause the network not to converge easily to the optimal solution during training.

To make the prediction more robust and stable, we train and test the optical flow branch following EMA approach. EMA [44] was adapted to the CNN training to facilitate the semi-supervised classification result by averaging model weights instead of predictions. During training, the teacher model updates its parameters by using the EMA weights of the student model. To the best of our knowledge, this is the first application of EMA to cross-task prediction of depth and optical flow.

As illustrated in Fig. 2(b), the optical flow branch contains teacher and student decoders. Both decoders share the same architecture and initial parameters as well. The teacher decoder only works during training. During training, the student branch updates its parameters with back-propagation. For each training step, the teacher branch is updated in moving average approach according to the parameters of student branch. Only the feature maps predicted by the teacher decoder are given as the input of the F2D mechanism while the depth branch affiliates optical flow prediction by
feeding student decoder feature maps via D2F mechanism. After being trained, the model discards the student decoder and the F2D mechanism extracts optical flow maps from teacher decoder.

C. Dual-head Mechanism

Self-attention based framework [46], [47] avoids the locality in recurrent operations by making full use of global dependencies between input and output. The self-attention framework with multi-head attention mechanism can capture information from different representation subspaces and further achieve better performance. Zheng et al. [48] proposed a Residual Aligner (RA) module to decouple different motions of nearby objects with a multi-head mask. While in this paper, inspired by the use of multi-head attention and following the idea of divide-and-conquer, we adopted dual-head disentanglement for D2F mechanism to further improve the optical flow estimation performance by decoupling the rigid and non-rigid motions.

The dual-head mechanism is applied to divide optical flow into rigid and non-rigid motion field representation. As shown in Fig. 3, each dual-head mask is updated from the combination of the encoder feature map and the upsampled previous dual-head mask. While the first mask is updated from the encoder feature map only:

\[ M^i = \begin{cases} \varsigma \left( \mathcal{E}^3 \left( \text{ReLU} \left( \mathcal{E}^3 \left( \Phi_{en}^i \right) \right) \right) \right), & i = 1, 2, 3, 4 \\ \varsigma \left( \mathcal{E}^3 \left( \text{ReLU} \left( \mathcal{E}^3 \left( \Phi_{en}^i \right) \right) \right) \right), & i = 1 \end{cases} \] (2)

where \( \varsigma(\cdot) \) and \( \text{ReLU}(\cdot) \) respectively denotes sigmoid and ReLU operation.

The dual-head feature map of image is generated by the tensor product of and the warped previous dual-head mask . Meanwhile, and are first warped by in Fig. 4 and 5 illustrate the effect of D2F and dual-head mechanism to the result of optical flow and depth estimation. Fig. 6 further illustrates the trade-off between depth and optical flow prediction and adding different components to the network architecture.

D. Experiments Implementation

In this section, we validate the improvement of (1) the cross-task collaboration, (2) the dual-head mechanism, and (3) the EMA approach to depth evaluation and optical flow prediction. The network is trained on KITTI’s data split of Eigen et al. [49] following Zhou et al.’s [16] pre-processing. We train the network in the following steps:

1. The feature encoder and the depth branch are trained with the learning rate of \( 10^{-4} \) for 20 epochs.
2. The optical flow decoder branch and D2F blocks are further connected into the network. The network is trained with the learning rate of \( 10^{-4} \) for another 20 epochs.
3. Finally, we add the F2D block for depth evaluation assistance. The network is trained with the learning rate of \( 10^{-5} \) for 5 epochs.

We trained our network using depth loss following [50] and [21], respectively. Adam optimizer [51] is used in all of the training steps. The size of the images is \( 1024 \times 320 \). We validate the depth estimation result on the Eigen et al.’s testing split, and the optical flow prediction result on KITTI 2015 training set.

IV. RESULTS

A. Ablation Study

To validate the effect of each component of depth and optical flow prediction network, we tried different combinations of D2F mechanism, F2D mechanism, EMA approach, and dual-head mechanism in experiments. Tab. 1 shows the prediction results of the experiments with the various combinations, marked as I to VI. Fig. 4 and 5 illustrate the effect of D2F and dual-head mechanism to the result of optical flow and depth estimation. Fig. 6 further illustrates the trade-off between depth and optical flow prediction and adding different components to the network architecture.

Fig. 4: Illustration of the ablation study of optical flow prediction. From top to bottom: input color image and the prediction result of model I to VI, and the optical flow Ground Truth. In the orange boxes are vehicles or their respective optical flow.

Fig. 5: Illustration of the depth estimation. From top to bottom: input color image and the prediction result of model I to VI. In the yellow boxes, we show some road signs that are easy to be confused with the background buildings.
TABLE I: Ablation study indicating how the mechanisms we proposed affect the parameter amount, Frames Per Second (FPS), and prediction result of depth and optical flow. Baseline: monodepth2 [50] and PWC-Net [21] working separately. The parameter mount is the sum of these two networks. D2F: depth-to-flow mechanism. F2D: flow-to-depth mechanism.

| Model | Components | Parameters | FPS  | Depth | Optical Flow |
|-------|------------|------------|------|-------|--------------|
|      |            |            |      |        |              |
| I    | Baseline   | 19.97M     | -    | 0.115 | 8.20        |
|      | I + D2F    | 18.22M     | 12.8 | 0.117 | 6.73        |
| II   | II + F2D   | 19.00M     | 11.9 | 0.113 | 7.23        |
| III  | III + EMA  | 19.00M     | 11.9 | 0.112 | 7.32        |
| IV   | IV + D-H   | 18.78M     | 8.1  | 0.117 | 7.11        |
| V    | V + D-H    | 19.55M     | 7.7  | 0.112 | 6.84        |

D2F mechanism can significantly improve the accuracy of optical flow prediction, which can be proved by comparing experiment I and II in Tab. I and Fig. 4. Furthermore, it can be seen from Fig. 5 that D2F mechanism can also support the depth branch telling the road signs from the buildings. This might be because objects like road signs are closer to the camera and can generate optical flow different from the distant buildings, which can be detected by the optical flow branch.

The information generated from optical flow prediction facilitates depth estimation via F2D mechanism. This could be seen from the comparison of experiment I, II, and III in Tab. I. Experiment III and IV also demonstrates that EMA approach can benefit both depth and optical flow prediction.

Experiment V and VI, respectively compared to experiment II and IV, prove that dual-head mechanism supports the optical flow prediction accuracy. The supporting role of dual-head mechanism can also be seen in Fig. 4 and 5. As shown in Fig. 4 comparing to baseline, with or without D2F mechanism’s assistance, adding dual-head mechanism to the architecture can significantly improve the network’s ability to recognize moving objects like vehicles – even in the dark tunnel. Fig. 5 shows that during depth estimation, the network with dual-head mechanism can also help distinguish between buildings and road signs.

Fig. 6 tells how the different components affect the two tasks of the network. D2F mechanism significantly improves optical flow prediction while causing the depth estimation accuracy slightly drops. A similar effect was achieved with F2D, enhances the accuracy of depth prediction while mildly reducing the effect of optical flow prediction. EMA approach achieves a positive effect on both tasks and dual-head mechanism facilitates optical flow prediction with little impact on depth estimation.

B. Depth and Optical Flow Evaluation

The comparisons of monocular depth and optical flow estimation between our approach and other methods are summarized in Tab. II and Fig. 7 and Fig. 8.

In Tab. II and Fig. 8, we take our models in experiment V and VI of Tab. I to compare with other baselines. Our models achieve better or comparable results for depth and optical flow estimation. Moreover, benefit from shared encoder and small components, our models contain small amount of parameters and thus achieve higher or comparable FPS. Moreover, as shown in Fig. 7 in both depth and optical flow prediction, our method achieves significant improvement of detecting vehicles.

V. CONCLUSION

In this paper, we design a multi-task network that achieves mutual assistance between the two tasks of optical flow and depth prediction through bi-directional information interaction. Specifically, we design D2F and F2D modules for cross-domain information interaction and apply the EMA method to the training of the network. We further enhance the network’s ability to distinguish objects during multi-task prediction by applying dual-head mechanism to the optical flow branch. Besides improving the mean accuracy of depth and optical flow prediction, our method significantly facilitates the ability to tell objects like road signs and vehicles from the scene. In the future, our work can be applied to the visual perception of autonomous vehicles and the navigation of surgical robots.
TABLE II: Depth estimation comparison. Top row: unsupervised monocular depth estimation methods. Middle top row: unsupervised optical flow prediction methods. Middle bottom row: multi-task estimation methods. Bottom row: our method.

| Method            | Parameters | FPS | Scheme | abs rel | sq rel | rms | log rms | $\delta_1$ | $\delta_2$ | $\delta_3$ | epe | epe noc | F1(%) |
|-------------------|------------|-----|--------|---------|--------|-----|---------|------------|------------|------------|-----|---------|-------|
| zhou et al. [16]  | -          | -   | -      | 0.183   | 1.595  | 6.399| 0.270   | 0.734      | 0.902      | 0.359      | -  | -       | -    |
| monodepth2 [50]   | -          | -   | -      | 0.115   | 0.882  | 4.701| 0.190   | 0.879      | 0.961      | 0.982      | -  | -       | -    |
| DOP [31]          | -          | -   | -      | 7.04    | 24.21  | -    | -       | -          | -          | -          | -  | -       | -    |
| Unflow [26]       | -          | -   | -      | 8.10    | 23.27  | -    | -       | -          | -          | -          | -  | -       | -    |
| DF-Net [27]       | 151.15M    | -   | (a)    | 0.115   | 0.882  | 4.701| 0.190   | 0.879      | 0.961      | 0.982      | -  | -       | -    |
| GL-Net [28]       | 59.77M     | 7.8 | (a)    | 0.140   | 1.070  | 5.253| 0.217   | 0.827      | 0.943      | 0.977      | -  | -       | -    |
| CC [30]           | 59.77M     | 7.9 | (a)    | 0.147   | 0.936  | 4.348| 0.218   | 0.810      | 0.941      | 0.977      | -  | -       | -    |
| GeoNet [35]       | 118.52M    | 0.2 | (b)    | 0.139   | 1.297  | 5.879| 0.223   | 0.827      | 0.936      | 0.979      | -  | -       | -    |
| Yang. [36]        | -          | -   | (d)    | 0.117   | 1.000  | 4.859| 0.194   | 0.880      | 0.960      | 0.980      | -  | -       | -    |
| Ours (V)          | 18.78M     | 8.1 | (e)    | 0.112   | 0.859  | 4.798| 0.193   | 0.879      | 0.960      | 0.981      | 6.23| 2.88    | 19.88 |
| Ours (VI)         | 19.55M     | 7.7 | (e)    | 0.150   | 1.124  | 5.507| 0.223   | 0.806      | 0.933      | 0.973      | 8.98| -       | 26.01 |
| Mono2+PWC         | 20.0M      | -   | -      | 0.135   | 1.070  | 5.230| 0.210   | 0.841      | 0.948      | 0.980      | 8.35| 4.86    | -    |

Fig. 7: Depth and optical flow prediction illustration comparison between CC [30], monodepth2 [50]+PWC-Net [21], and our method (model V and VI). Column (a) shows the target image and its optical flow Ground Truth. Column (b), (c), and (d) shows the depth and optical flow prediction results of CC, monodepth2 and PWC-Net, and our method.

Fig. 8: Bubble chart comparing the trade-off of depth estimation, optical flow prediction, and parameter mount between different methods. The horizontal and vertical axis represents optical flow prediction non-occluded pixels endpoint error and depth prediction absolute relative error, while the size of the bubbles shows the parameter mounts.

REFERENCES

[1] J. Janai, F. Güney, A. Behl, A. Geiger, et al., “Computer vision for autonomous vehicles: Problems, datasets and state of the art,” Foundations and Trends® in Computer Graphics and Vision, vol. 12, no. 1–3, pp. 1–308, 2020.
[2] M. Pirvu, V. Robu, V. Licaret, D. Costea, A. Marcu, E. Slusanschi, R. Sukthankar, and M. Leordeanu, “Depth distillation: unsupervised metric depth estimation for uavs by finding consensus between kinematics, optical flow and deep learning,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 3215–3223.
[3] B. Huang, J.-Q. Zheng, A. Nguyen, D. Tuch, K. Vyas, S. Giannarou, and D. S. Elson, “Self-supervised generative adversarial network for depth estimation in laparoscopic images,” in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2021, pp. 227–237.
[4] J.-Q. Zheng, X.-Y. Zhou, C. Riga, and G.-Z. Yang, “Real-time 3-d shape instantiation for partially deployed stent segments from a single 2-d fluoroscopic image in fenestrated endovascular aortic repair,” IEEE Robotics and Automation Letters, vol. 4, no. 4, pp. 3705–3710, 2019.
[5] M. Kloft and A. Vedaldi, “Supervising the new with the old: learning sfm from sfm,” in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 698–713.
[6] 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), September 2018.
Proceedings of the IEEE international conference on computer vision, 2015, pp. 2650–2658.

[50] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, “Digging into self-supervised monocular depth estimation,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 3828–3838.

[51] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, 2014.