P1-20

18th International Conference on Machine Vision Applications (MVA)
Hamamatsu, Japan, July 23-25, 2023.

ViTVO: Vision Transformer based Visual Odometry with Attention Supervision

Chu-Chi Chiu, Hsuan-Kung Yang, Hao-Wei Chen, Yu-Wen Chen, and Chun-Yi Lee

Elsa Lab, Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan

Abstract

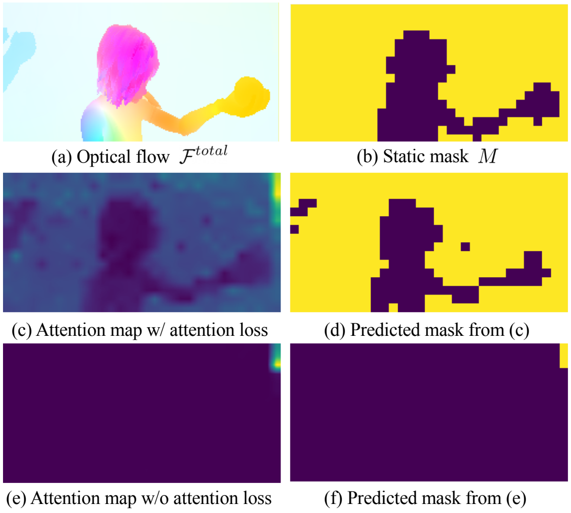
In this paper, we develop a Vision Transformer based visual odometry (VO), called ViTVO. ViTVO introduces an attention mechanism to perform visual odometry. Due to the nature of VO, Transformer based VO models tend to overconcentrate on few points, which may result in a degradation of accuracy. In addition, noises from dynamic objects usually cause difficulties in performing VO tasks. To overcome these issues, we propose an attention loss during training, which utilizes ground truth masks or self supervision to guide the attention maps to focus more on static regions of an image. In our experiments, we demonstrate the superior performance of ViTVO on the Sintel validation set, and validate the effectiveness of our attention supervision mechanism in performing VO tasks.

1 Introduction

VO is a crucial process for various applications, but conventional VO methods often face challenges in environments with dynamic objects, leading to degradation in accuracy. To address this issue, VO methods based on convolutional neural networks (CNNs) have been explored, utilizing additional dynamic masks to filter out noise from moving objects. However, these methods still suffer from some drawbacks, such as not always focusing on static regions or requiring additional segmentation models for detecting dynamic regions [1–5].

Self-attention based mechanisms have been investigated to enable deep learning-based VO models to focus on static regions. These mechanisms utilize self-attention layers to guide the models, allowing them to have relatively wider fields of view and better distinguish between dynamic and static regions than pure CNN-based approaches [1, 6, 7]. Among these, Transformer-based VO methods have gained prominence due to the successes of Vision Transformer (ViT) in various vision tasks [8–12]. However, these models still suffer from overly focusing on a sparse set of pixels.

To tackle this issue, we propose a new Vision Transformer-based visual odometry approach called ViTVO, which adopts an extra attention supervision mechanism to enforce the model to focus on the static regions of the image. With this attention supervision, our ViTVO can extract global information in earlier layers and stably attend to static regions, as in Fig. 1.


(a) Optical flow $\mathcal{F}^{total}$ (b) Static mask M

(c) Attention map w/ attention loss (d) Predicted mask from (c)

(e) Attention map w/o attention loss (f) Predicted mask from (e)

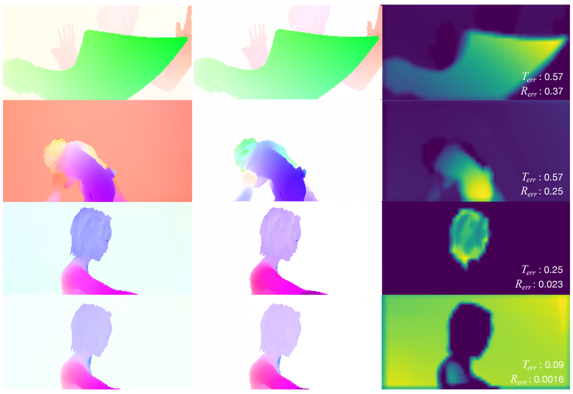
Figure 1: The attention maps of ViTVO w/ and w/o the use of the proposed attention loss via the highlighted patch regions.

We validate the effectiveness of the proposed ViTVO through a series of experiments, comparing it to other baselines. We train all models on a dataset generated by the data generation workflow proposed in [13] and validate them on the Sintel dataset [14]. The quantitative results show that ViTVO delivers better VO performance in terms of rotational and translational errors, and the saliency maps further validate the effectiveness of the proposed attention supervision method.

2 Preliminary

In this section, we highlight the issues of conventional VO techniques and introduce the main motivations behind our proposed methodology. As discussed in Section 1, dynamic objects might cause noises for VO models, while such models have traditionally been implemented by either CNN based approaches [1–7, 13, 15, 16] or Transformer based approaches [8–12]. Although those prior arts have achieved promising performance, they inherently suffer from several issues that prevent them from being able to effectively eliminate the impacts of dynamic objects. We next elaborate on the above issues separately as in the following.

CNN is deficient in attention capability. Due to architectural limitations, CNN based VO models generally lack effective manners to deal with the influences of dynamic objects. As a CNN model tend to regress itself and optimize its final projection, such an objec-



(a) Optical flow $\mathcal{F}^{total}$ (b) Object flow $\mathcal{F}^{obj}$ (c) VONet saliency map

Figure 2: The saliency maps of VONet for testcases of Sintel [14].

tive does not necessarily encourage a CNN to filter out the regions of dynamic objects, unless it is tailored to do so. This deficiency might cause CNN based VO models prone to be affected by dynamic objects, especially when those objects occupy a significant portion of an image. Without correct clues about background regions to be referenced, they could predict incorrectly under such scenarios and potentially be misguided.

To demonstrate the above issue, we select a recent CNN based VO approach, called VONet [17], to illustrate the attention regions focused by it. Fig. 2 visualizes the saliency map of VONet on some testcases from the Sintel dataset [14], and compares the prediction errors of different cases. It can be observed from Fig. 2 that the prediction errors of the first three rows, which correspond to the cases with dynamic objects, are relatively higher than those from the bottom row, in which VONet is not mislead by dynamic objects. In addition, even for image frames from the same video clip (i.e, the third and the fourth rows), VONet may focus on either the foreground or background regions, as illustrated in Fig. 2. This causes the prediction errors of VONet to fluctuate across different image frames. Although some researchers proposed to incorporate additional attention mechanisms to CNN-based models, the performance enhancement is still limited [1, 6, 7, 18].

The nature of VO causes Transformers to be over-concentrated. To borrow the benefits of the attention mechanism and to overcome the limitations posed by CNN-based VO models, some recent researchers turned their focus and introduced Transformer-based VO techniques [8–12], and were able to deliver promising results. Nevertheless, due to the self-attention nature of Transformers, such models tend to concentrate on only few points, as depicted in Fig. 1 (e). The rationale is that according to the perspective-n-point (PnP), only few paired correspondences in certain background area are feasible for VO approaches to infer the camera motion, even being in lack of any global information. In practice, however, overly relying on certain points for VO prediction may cause Transformer-based VO models to lose generalizability, as the camera pose changes derived from those points might be biased and vulnerable to noises or imperfect feature extraction. Moreover, the above tendency is not advantageous to model training, as the mechanism of Transformers is formulated as follows:

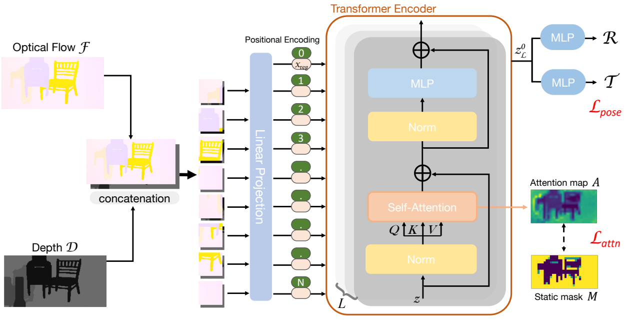


Figure 3: An overview of the proposed ViTVO framework.

$$\frac{\partial SA(z)}{\partial W^v} = \alpha \frac{\partial SA(z)}{\partial z}, \tag{1}$$

where $SA(\cdot)$ represents the self-attention function, α is the attention weights, z is the input feature embedding before each transformer block, and W^v is the weights for projecting z to the values v of a Transformer block [19]. It can be observed that the attention derivative has positive correlation with α. If the attention weights α is overly concentrated on certain specific points, only a limited set of parameters can be effectively updated during each training iteration. This, in turn, can result in slow convergence. The above issues motivate us to investigate solutions to Transformer-based VO models.

2.1 ViTVO

Fig. 3 depicts an overview of the ViTVO framework. Given an optical flow map $\mathcal{F}_i$ derived from two image frames I_i and I_{i+1}, and a depth map $\mathcal{D}_i$ of I_i, the main objective of the ViTVO framework is to infer the corresponding camera rotation $\mathcal{R}_i$ and translation $\mathcal{T}_i$.

2.1.1 The refined self-attention module

The attention weights α in each encoder layer are computed based on the pairwise similarity between the respective key K and query Q, and are used to scale the values V. The procedure of the self-attention scheme can be formulated as $SA(z) = \alpha V$, where $\alpha = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)$, and d is the scaling factor that represents the dimension of the key embedding for normalizing the dot-product value of the query Q and key K. To encourage the model to focus more on the static regions for prediction, an additional attention loss is applied at all of the encoder layers for refining the attention weights α. The weights of different patches are coalesced to form the attention weight map A as:

$$A = concat([\alpha^0, \alpha^1, ..., \alpha^N]). \tag{2}$$

In order to meet the attention supervision objective discussed above, this coalesced attention weight map A can be refined and guided during the training phase by either (a) a binary cross entropy loss, or (b) a regularizer term. These types of losses are applied to scenarios when explicit supervision annotations exist or not.

Binary cross entropy for supervision. We compare A with the ground truth static mask M, and treat it as a segmentation task which contains two classes (i.e., static and dynamic regions). The binary cross entropy loss is expressed as follows:

$$\mathcal{L}_{bce} = \sum_{l=1}^{\mathcal{L}} M \cdot \log A_l + (1 - M) \cdot \log(1 - A_l). \quad (3)$$

Regularization for self-supervision. If the ground truth static mask is not available, the attention map A can still be supervised by a regularization loss term as:

$$\mathcal{L}_{reg} = -\sum_{l=1}^{\mathcal{L}} \sum_{j=0}^{N} (A_l[j]/max(A_l[j]))/N, \quad (4)$$

where $A_l[j]$ represents α^j from the l^{th} layer of the encoder. This loss function design can prompt the attention map A to cover as many static regions as possible, and prevent the situation that A only concentrates on certain patch regions.

2.1.2 MLP Decoder for Regression

The extracted features z_L from the last encoder block is used for performing a regression task to predict the rotation $\mathcal{R}_i$ and the translation $\mathcal{T}_i$ vectors of the camera. Similar to BERT's and ViT's [*class*] token, a learnable embeddings is prepended to the sequence of the embedding patches (i.e., $z_0^0 = x_{reg}$) to indicate that the output from the last layer of the transformer encoder z_L^0 is used to serves as the feature representation. This feature representation is then used for predicting the final camera pose $\mathcal{R}_i$ and $\mathcal{T}_i$. Two separate MLP layers are employed to generate these predictions, which can be formulated as the following,

$$\tilde{\mathcal{R}}_i = \text{MLP}(\text{LN}(z_L^0)), \quad (5)$$

$$\tilde{\mathcal{T}}_i = \text{MLP}(\text{LN}(z_L^0)), \quad (6)$$

where LN stands for the layer normalization operation.

The loss function used for optimizing the model is a supervised L_2-norm loss, which calculates the difference between the six degrees of freedom (DoF) ground truth pose $(\mathcal{R}_i, \mathcal{T}_i)$ and the predicted pose $(\tilde{\mathcal{R}}_i, \tilde{\mathcal{T}}_i)$, which is denoted as $\mathcal{L}_{pose}$.

2.2 Total Loss Function

The total loss of ViTVO is expressed as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{pose} + \lambda_{attn}\mathcal{L}_{attn}, \quad (7)$$

where λ_{attn} is a scaling factor. $\mathcal{L}_{attn}$ is set to $\mathcal{L}_{bce}$ when the ground truth static masks are available. Otherwise, $\mathcal{L}_{attn}$ is set to $\mathcal{L}_{reg}$ for performing self supervision.

3 Experimental Results

3.1 Evaluation Metrics

The following evaluation metrics are adopted for measuring the performance in the experiments: (1) the average L_1 rotational error R_{err} and the average L_1 translational error T_{err}, and (2) the mean intersection over union (mIoU) between the ground truth static mask M and the predicted binary mask generated from the attention map of ViTVO. In order to achieve this, an attention weight map A is turned into a predicted binary mask M^A according to the values of each pixel entry, expressed as the following equation:

$$M^A = \begin{cases} 1 & \text{if } ||A|| > \epsilon \\ 0, & \text{otherwise,} \end{cases} \quad (8)$$

where the comparison with threshold ϵ is carried out in a pixel-wise fashion. To determine the value of ϵ, we first design a set of candidate thresholds $\{\epsilon_1, \epsilon_2, \cdots, \epsilon_N\}$, and use them to predict a set of binary masks $\{M_1^A, M_2^A, \cdots, M_N^A\}$. We then compute a coarse static mask M_{init} according to [20, 21], and select a $\epsilon_i, i \in [1, N]$ that generates a prediction mask M_i^A with the highest mIoU with M_{init}. M_i^A is used to compute the final mIoU with the ground truth M.

3.2 Quantitative Results

In this section, we compare the proposed ViTVO against the VONet [17] and the PWVO [13] baselines, and report the quantitative results in Table 1. We provide two versions of ViTVO for comparison: with an without the use of $\mathcal{D}_{i+1}$ as its input. It can be observed that our proposed ViTVO is able to outperform the VONet baseline in terms of R_{err} and T_{err} by considerable margins. To fairly compare with PWVO, we additionally introduce a depth map $\mathcal{D}_{i+1}$ to to input of ViTVO. The results indicate that ViTVO is able to achieve better performance than PWVO in terms of R_{err}, but is unable to outperform PWVO in terms of T_{err}. The rationale behind this might be due to the fact that PWVO additionally employs an ego flow loss during training [13], which causes PWVO to be sensitive to the predictions of pose translation. Therefore, PWVO might tend to optimize the translational errors, resulting in its better performance in terms of T_{err}.

3.3 Qualitative Results

Visualization of the saliency map In this section, we compares the attention weight map A with the saliency map generated from the baseline VONet with the Grad-CAM [22] approach. Both maps highlight the regions that contribute to the final prediction of the camera poses and can be referred as an evidence for understanding the focus areas of the models. The

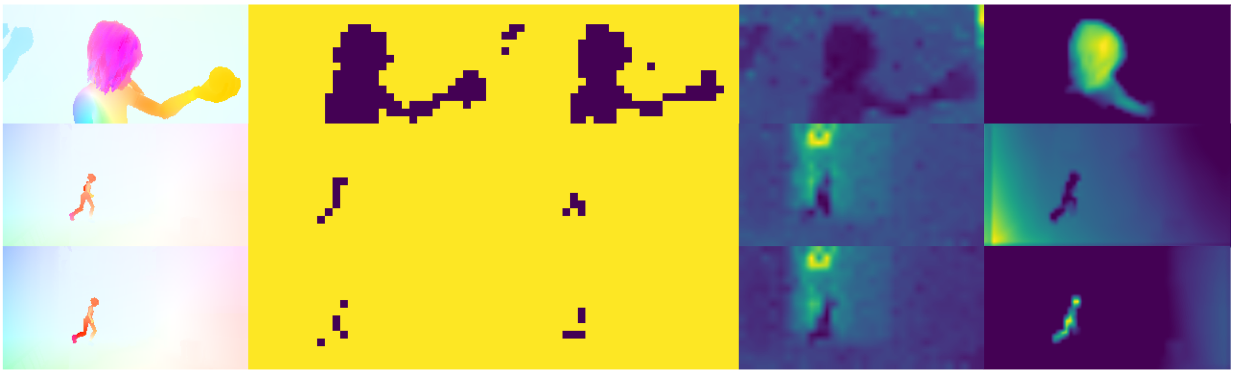

(a) Optical flow $\mathcal{F}$ (b) Ground truth static mask M (c) Predicted binary mask M_i^A (d) ViTVO attention weight map A (e) VONet saliency map

Figure 4: A comparison of ViTVO and VONet through their saliency maps. It can observed that the predicted masks from ViTVO (i.e., column (c), which are derived according to Eq. (8)) are highly correlated with the ground truth static masks (i.e., column (b)), which validates the effectiveness of our approach.

Table 1: Comparison of ViTVO and the baselines in terms of R_{err} and T_{err}. It can be observed that ViTVO outperforms VONet when using the same inputs, and delivers better performance than PWVO in terms of R_{err}.

	Input				Error	
	$\mathcal{F}_i$	$\mathcal{D}_i$	$\mathcal{D}_{i+1}$	$\mathcal{K}$	R_{err}	T_{err}
VONet	✓	✓			0.145	0.141
PWVO	✓	✓	✓	✓	0.081	**0.043**
ViTVO	✓	✓			0.071	0.092
ViTVO	✓	✓	✓		**0.064**	0.075

results are shown in Fig. 4. It can be observed the predicted masks M_i^A and the attention weight map A from our proposed ViTVO closely aligns to the ground truth static mask M. On the other hand, the saliency maps generated from VONet sometimes highlight the regions that dynamic objects located in, which may potentially affect the quality and correctness of the final prediction as discussed in the Sec. 2. It is worth to be noted that, as shown in the third and forth rows of Fig. 4, the saliency maps from VONet highlight completely inverse regions even when the input optical flow map is similar and is from consecutive frames. On the contrary, the proposed ViTVO generate consistent attention weight map and implies that the ViTVO could perform more stable than VONet.

3.4 Ablation Study

In this section, we ablatively examine the effectiveness of our proposed attention loss, and report the results in Table 2. As described in Section 2.1, $\mathcal{L}_{attn}$ is set to the binary cross entropy loss $\mathcal{L}_{bce}$ when the ground truth static masks are available. Otherwise, it is set to the regularization loss $\mathcal{L}_{reg}$. The second row of Table 2 suggests that by simply applying the attention loss term $\mathcal{L}_{reg}$ during training, the performance improves considerably. Moreover, the third row reveals that by setting an appropriate value of λ_{attn}, the performance in terms of R_{err}, T_{err}, and mIoU can be further enhanced. If the ground truth static masks are available (i.e., $\mathcal{L}_{bce}$ is adopted), ViTVO is able to achieve its best performance in terms of the three metrics, as shown in the fourth row of Table 2.

Table 2: An analysis for the effectiveness of the proposed λ_{attn}. The results are evaluated on the validation sets of Sintel [14].

Loss Design	R_{err}	T_{err}	mIoU
w/o $\mathcal{L}_{attn}$	0.162	0.113	3.202
w/ $\mathcal{L}_{reg}$ ($\lambda_{attn} = 1.00$)	0.078	0.092	69.87
w/ $\mathcal{L}_{reg}$ ($\lambda_{attn} = 0.35$)	0.075	0.090	77.94
w/ $\mathcal{L}_{bce}$ ($\lambda_{attn} = 0.05$)	**0.071**	**0.083**	**78.54**

4 Conclusion

In this paper, we propose ViTVO, a Vision Transformer-based VO framework that tackles a key problem in conventional VO methods: dynamic objects in input observations, which create difficulties in estimating camera motions due to noise. Previous efforts employed semantic mask strategies but suffered from architectural limitations. To train ViTVO in identifying noisy regions, we introduce an attention supervision mechanism that allows the model to focus on static areas when performing VO tasks. We conducted experiments on the Sintel validation set for ViTVO and baseline approaches, along with ablation analyses to support our design choices. Our quantitative and qualitative results show that ViTVO achieves favorable outcomes, and even when trained on a synthetic dataset, ViTVO can be transferred to an unfamiliar Sintel validation set without requiring any fine-tuning.

5 Acknowledgements

The authors acknowledge the National Science and Technology Council in Taiwan (grant number MOST 111-2223-E-007-004-MY3), NVIDIA Corporation and NVIDIA AI Technology Center for the donated GPUs, and the National Center for High-Performance Computing for the computational and storage resources.

References

[1] H. Damirchi, R. Khorrambakht, and H. D. Taghirad. Exploring self-attention for visual odometry. *arXiv*, abs/2011.08634, 2020.

[2] W. Ye, X. Lan, S. Chen, Y. Ming, X. Yu, H. Bao, Z. Cui, and G. Zhang. PVO: Panoptic visual odometry. *ArXiv*, abs/2207.01610, 2022.

[3] X.-Y. Kuo, C. Liu, K.-C. Lin, E. Luo, Y.-W. Chen, and C.-Y. Lee. Dynamic attention-based visual odometry. In *Proc. IEEE Int. Conf. on Intelligent Robots and Systems (IROS)*, pages 5753–5760, 2020.

[4] S. Lee, S. Im, S. Lin, and I. S. Kweon. Instance-wise depth and motion learning from monocular videos. *ArXiv*, abs/1912.09351, 2019.

[5] S. Shen, Y. Cai, W. Wang, and S. Scherer. DytanVO: Joint refinement of visual odometry and motion segmentation in dynamic environments. *ArXiv*, 2022. doi: abs/2209.08430.

[6] F. Xue, Q. Wang, X. Wang, W. Dong, J. Wang, and H. Zha. Guided feature selection for deep visual odometry. In *Proc. Asian Conf. on Computer Vision (ACCV)*, Dec. 2018.

[7] B. Wang, C. Chen, C. X. Lu, P. Zhao, N. Trigoni, and A. Markham. AtLoc: Attention guided camera localization. *arXiv:1909.03557*, 2019.

[8] X. Li, Y. Hou, P. Wang, Z. Gao, M. Xu, and W. Li. Transformer guided geometry model for flow-based unsupervised visual odometry. *Neural Computing and Applications*, 33:8031–8042, Jan. 2021.

[9] A. Klochkov and I. Drokin. Transformer-based deep monocular visual odometry for edge devices. In *Proc. the 31st FRUCT Conf.*, page 422–428, Apr. 2022.

[10] E. Parisotto. An empirical evaluation of sequence-based deep learning architectures for visual odometry. 2018.

[11] N. Kaygusuz, O. Mendez, and R. Bowden. AFT-VO: Asynchronous fusion transformers for multi-view visual odometry estimation. *ArXiv*, abs/2206.12946, 2022.

[12] R. Cao, Y. Wang, K. Yan, B. Chen, T. Ding, and T. Zhang. An end-to-end visual odometry based on self-attention mechanism. In *Proc. IEEE Int. Conf. on Power, Intelligent Computing and Systems (ICPICS)*, pages 406–410, Jul. 2022.

[13] H.-W. Chen, T.-H. Liao, H.-K. Yang, and C.-Y. Lee. Pixel-wise prediction based visual odometry via uncertainty estimation. In *Proc. IEEE Winter Conf. on Applications of Computer Vision (WACV)*, pages 2518–2528, Oct. 2023.

[14] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In *Proc. European Conf. on Computer Vision (ECCV)*, pages 611–625, Oct. 2012.

[15] R. Zhu, M. Yang, W. Liu, R. Song, B. Yan, and Z. Xiao. DeepAVO: Efficient pose refining with feature distilling for deep visual odometry. *arXiv*, abs/2105.09899, 2021.

[16] W. Zhou, H. Zhang, Z. Yan, W. Wang, and L. Lin. DecoupledPoseNet: Cascade decoupled pose learning for unsupervised camera ego-motion estimation. *IEEE Trans. Multimedia*, Feb. 2022.

[17] W. Wang, Y. Hu, and S. A. Scherer. TartanVO: A generalizable learning-based vo. In *Proc. Conf. on Robot Learning (CoRL)*, Nov. 2020.

[18] E. Parisotto, D. S. Chaplot, J. Zhang, and R. Salakhutdinov. Global pose estimation with an attention-based recurrent network. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition Workshop (CVPRW)*, pages 237–246, Jun. 2018.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Proc. Conf. on Neural Information Processing Systems (NeurIPS)*, Dec. 2017.

[20] Z. Yin and J. Shi. GeoNet: Unsupervised learning of dense depth, optical flow and camera pose. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1983–1992, Jun. 2018.

[21] L. Liu, G. Zhai, W. Ye, and Y. Liu. Unsupervised learning of scene flow estimation fusing with local rigidity. In *Proc. Int. Joint Conf. on Artificial Intelligence (IJCAI)*, page 876–882, Aug. 2019.

[22] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, pages 618–626, Oct. 2017.