Learning to Fuse Monocular and Multi-view Cues for Multi-frame Depth Estimation in Dynamic Scenes (CVPR 2023)

2023.11.16

Review of Mono-MVS

$$L_{\text{consistency}} = \sum M \left| D_t - \hat{D}_t \right|. \qquad M = \max\left(\frac{D_{\text{cv}} - \hat{D}_t}{\hat{D}_t}, \frac{\hat{D}_t - D_{\text{cv}}}{D_{\text{cv}}}\right) > 1.$$

rics are above predefined thresholds: (1) The static stereo photometric error using D_t, *i.e.*, $pe^t_{ts}(\mathbf{x}, D_t(\mathbf{x}))$. (2) The average temporal stereo photometric error using D^S_t, *i.e.*, $\overline{pe^t_{t'}}(\mathbf{x}, D^S_t(\mathbf{x}))$. (3) The difference between $D_t(\mathbf{x})$ and $D^S_t(\mathbf{x})$. Please refer to our supplementary materials for

2

ManyDepth [36]提出了一种自我发现的掩模，并用单眼深度来监督潜在的动态区域。
MonoRec [37]提出了一种运动分割网络发现可动的物体，然后三个指标中有两个大于阈值就来掩盖成本量中的动态区域，并仅使用单目图像特征来推断深度。
尽管动态结果比纯多帧估计更高，但它们的性能相当可比 [1,9,36]，甚至比他们提出的单目分支更差（如表 4 所示）。
此外，它们需要预先计算的实例掩码 [9, 37]，这会给网络训练或推理带来额外的计算负担。
还有另一种多帧方法[1]，它用大型单目网络指导多帧成本重建。然而，由于依赖单目网络预测，它的泛化能力弱于多帧方法（表3）。

Review of Mono-MVS

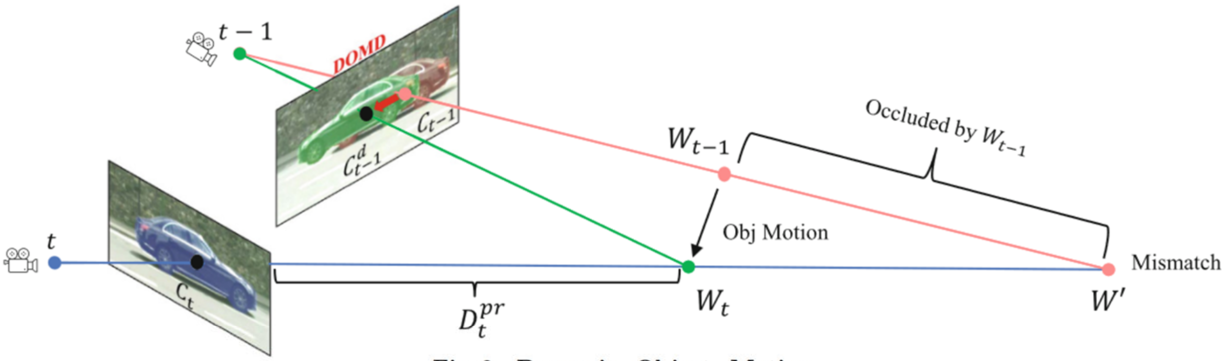


Fig. 3. Dynamic Object Motion Disentanglement: A dynamic object moves from W_{t-1} to W_t, C_{t-1} and C_t are corresponding image patches. D_t^{pr} is our depth prior prediction. Conventional methods tend to mismatch at W'. We re-project C_t to C_{t-1}^d with depth prior D_t^{pr} to replace C_{t-1} to disentangle the object motion. This solves the mismatch problem, making our cost volume and re-projection loss correctly converge at W_t.

3

[9]建议在计算成本体积之前使用单目深度校正图像平面中的动态对象位置（带有实例掩模）。

Contribution & Motivation

1. ManyDepth
2. MAGNet

1. Analyze multi-frame and monocular depth estimations in dynamic scenes and unveil their respective advantages in static and dynamic areas. Fuse Monocular and Multi-view Cues

2. Cross-Cue fusion module that utilizes the cros cue attention to encode non-local intrarelations from one depth cue to guide the representation of the other

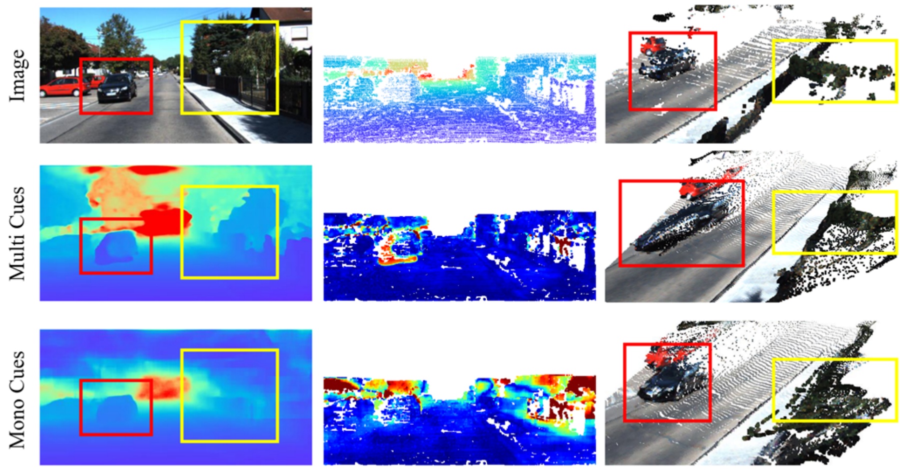
Method	Mono. Err.	Final Err.	Err. Redu.
Manydepth [36]	0.212	0.222	−4.72%
Dynamicdepth [9]	0.214	0.208	2.83%
MaGNet [1]	0.153	0.141	7.84%
Ours - Res.18	0.149	0.118	**20.81%**
Ours - Res.50	0.145	0.116	**20.00%**

Eval	Method	Backbone	Abs Rel	Sq Rel	RMSE	RMSE_{log}	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Overall	MonoRec [37]	Res-18	**0.158**	3.102	**7.553**	**0.227**	**0.854**	**0.931**	**0.961**
	MaGNet [1]	Effi-B5	0.208	2.641	10.739	0.382	0.620	0.878	0.942
	Ours	Res-18	**0.158**	**2.416**	9.855	0.299	0.747	0.894	0.947
Dynamic	MonoRec [37]	Res-18	0.544	16.703	16.116	0.482	0.460	0.667	0.798
	MaGNet [1]	Effi-B5	0.266	3.982	11.715	0.398	0.462	0.815	0.917
	Ours	Res-18	**0.234**	**3.611**	**11.007**	**0.331**	**0.576**	**0.835**	**0.921**

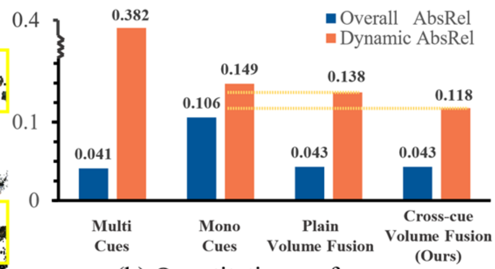
4

、尽管动态结果比纯多帧估计更高，但它们的性能相当可比 [1,9,36]，甚至比他们提出的单目分支更差（如表 4 所示）。
此外，它们需要预先计算的实例掩码 [9, 37]，这会给网络训练或推理带来额外的计算负担。
还有另一种多帧方法[1]，它用大型单目网络指导多帧成本重建。然而，由于依赖单目网络预测，它的泛化能力弱于多帧方法（表3）。

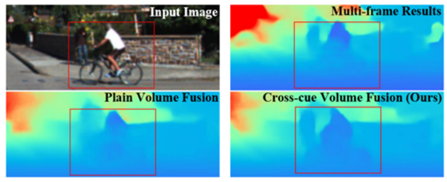
Analyses



(a) Multi-frame and monocular cues in the dynamic scene

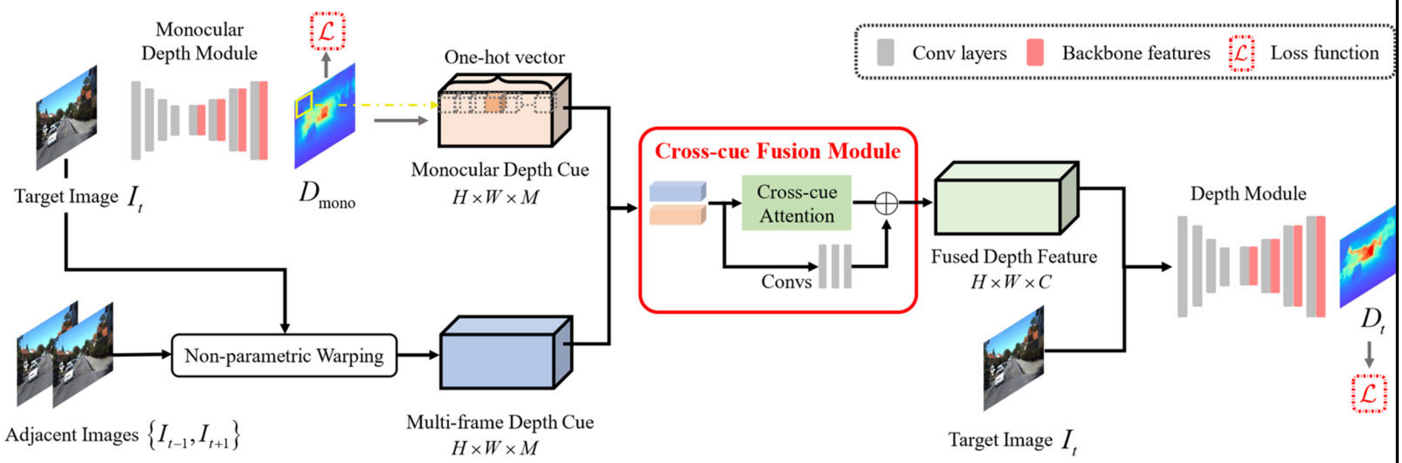


(b) Quantitative performance

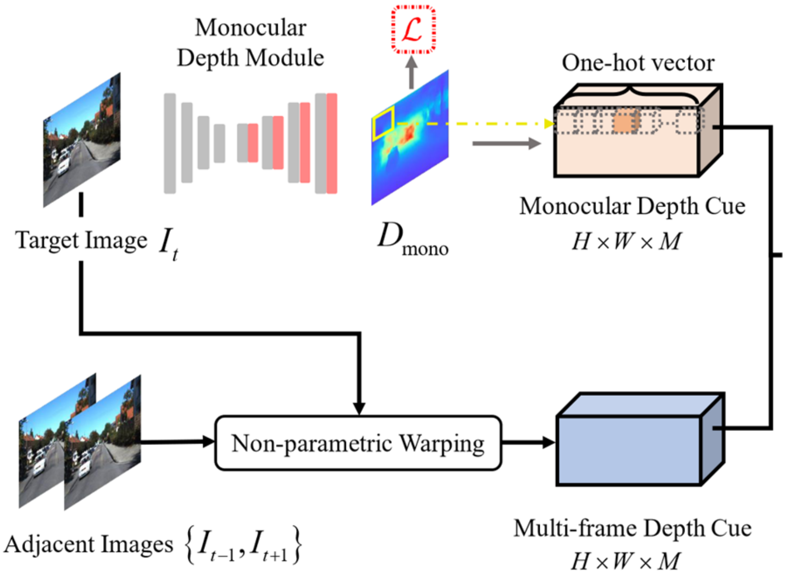


(c) Depth results in dynamic areas

Method: Overview of Network



Method: Representing Monocular and Multi-view Cues



$$C_{\text{mono},(i,j)}[k] = \{1 \mid d_{\text{mono}} \in (d_{k-1}, d_k]\}_{k=1}^{M},$$

SSIM

7

Method: The Cross-cue Fusion Module

$C_{\text{multi}}, C_{\text{mono}} \in \mathbb{R}^{H\times W\times M}$

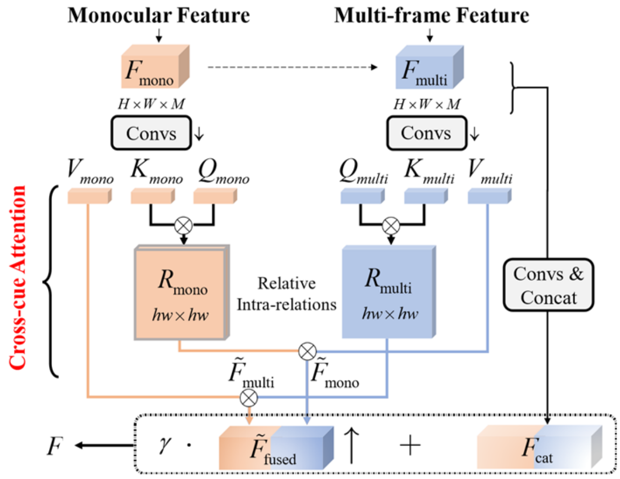
$F_{\text{multi}}, F_{\text{mono}} \quad \mathbb{R}^{h\times w\times M}$

$$F_{\text{cat}} = \text{Cat}(\text{Conv}(C_{\text{multi}}), \text{Conv}(C_{\text{mono}}))$$

$$R_{\text{mono}} = \text{Softmax}(Q_{\text{mono}} \otimes K_{\text{mono}}^{T})$$

$$\widetilde{F}_{\text{multi}} = R_{\text{mono}} \otimes V_{\text{multi}},$$

$$F = \gamma \widetilde{F}_{\text{fused}} \uparrow + F_{\text{cat}}.$$



这里好像是直接在图像上warp

Method: The Cross-cue Fusion Module

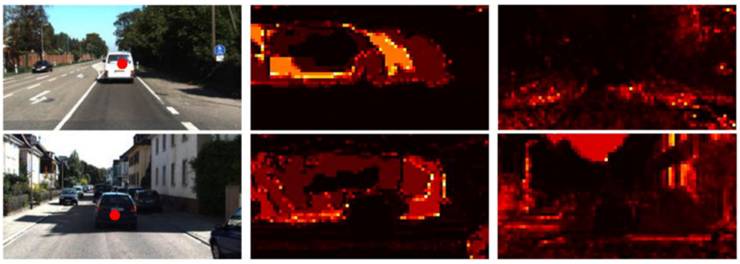
$$Q_{\text{mono}} = f(F_{\text{mono}}, \theta^{Q}_{\text{mono}}),$$

$$K_{\text{mono}} = f(F_{\text{mono}}, \theta^{K}_{\text{mono}}),$$

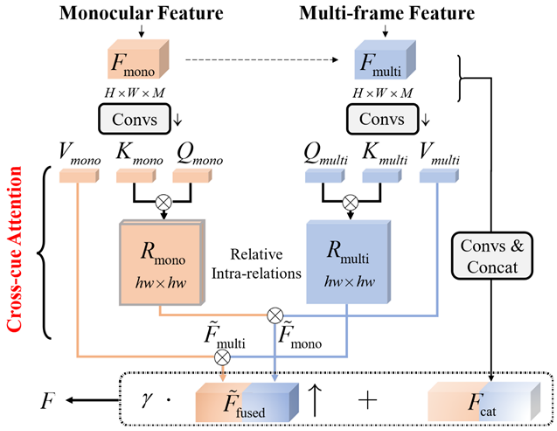
$$V_{\text{multi}} = f(F_{\text{multi}}, \theta^{V}_{\text{multi}}),$$

$$R_{\text{mono}} = \text{Softmax}(Q_{\text{mono}} \otimes K^{T}_{\text{mono}})$$

$$\widetilde{F}_{\text{multi}} = R_{\text{mono}} \otimes V_{\text{multi}},$$



(a) Input image (b) R_{mono} atten. map (c) R_{multi} atten. map



Rmono 单眼深度线索的非局部内部关系
Fmulti 表示受益于单眼深度线索的改进的多帧表示。
Rmulti代表多帧线索的内部关系，可用于改善单目深度特征Vmono。

Experiment

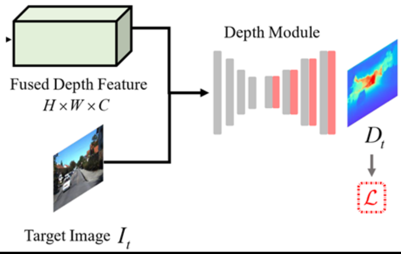
Eval	Method	Back.	Reso.	Sup.	Abs Rel	Sq Rel	RMSE	$RMSE_{log}$	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Overall	Manydepth [36]	Res-18	MR	M	0.071	0.343	3.184	0.108	0.945	0.991	0.998
	DynamicDepth [9]	Res-18	MR	M	0.068	0.296	3.067	0.106	0.945	0.991	0.998
	MonoRec [37]	Res-18	MR	D*	0.050	0.290	2.266	0.082	0.972	0.991	0.996
	Ours	Res-18	MR	D	**0.043**	**0.151**	**2.113**	**0.073**	**0.975**	**0.996**	**0.999**
	MaGNet [1]	Effi-B5	MR	D	0.057	0.215	2.597	0.088	0.967	**0.996**	**0.999**
	Ours	Effi-B5	MR	D	0.046	0.155	2.112	0.076	0.973	**0.996**	**0.999**
	MaGNet [1]	Effi-B5	HR	D	0.043	0.135	2.047	0.082	0.981	**0.997**	**0.999**
	Ours	Effi-B5	HR	D	**0.039**	**0.103**	**1.718**	**0.067**	**0.981**	**0.997**	**0.999**
Dynamic	Manydepth [36]	Res-18	MR	M	0.222	3.390	7.921	0.237	0.676	0.902	0.964
	DynamicDepth [9]	Res-18	MR	M	0.208	2.757	7.362	0.227	0.682	0.911	0.971
	MonoRec [37]	Res-18	MR	D*	0.360	9.083	10.963	0.346	0.590	0.882	0.780
	Ours	Res-18	MR	D	0.118	0.835	4.297	0.146	0.871	0.975	0.990
	MaGNet [1]	Effi-B5	MR	D	0.141	1.219	4.877	0.168	0.830	0.955	0.986
	Ours	Effi-B5	MR	D	**0.111**	**0.768**	**4.117**	**0.135**	**0.881**	**0.980**	**0.994**
	MaGNet [1]	Effi-B5	HR	D	0.140	1.060	4.581	0.202	0.834	0.954	0.982
	Ours	Effi-B5	HR	D	**0.112**	**0.830**	**4.101**	**0.137**	**0.885**	**0.978**	**0.992**

Table 1. **Quantitative comparisons on KITTI [10] Odometry dataset**. 'Back.' denotes the network backbone. 'Reso.' denotes the image resolutions, where 'MR' refers to the resolution of 256 × 512 and 'HR' is 352 × 1216. In the 'Sup.' column, 'M' are self-supervised methods, 'D*' refers to semi-supervised methods trained with pseudo GT depth, while 'D' denotes fully-supervised methods. Color blue denotes 'lower is better', while red means 'higher is better'. The best results are in **bold**.

使用KITTI Odometry数据集，据考是沿用MonoRec的backbone（MonoRec使用VSLAM得到的深度作为半监督）。因此数据和通常的KITTI深度数据集不太一样。

Ablation

#	Category	Variant	Dynamic					Overall
			Abs Rel	Abs Sq	RMSE	RMSE_{log}	$\delta \leq 1.25$	AbsRel
1	Depth with single cues	Pure multi-frame cues	0.382	7.167	10.292	0.35	0.509	0.041
2		Pure monocular cues	0.149	1.369	5.282	0.178	0.810	0.106
3	Volume fusion with masks	Self-discovered mask [36]	0.130	0.990	4.692	0.160	0.837	0.043
4		MaskNetwork [37]	0.220	2.896	6.299	0.223	0.735	0.040
5	Volume fusion without masks	Stack & 3D Convs	0.154	1.479	5.866	0.189	0.777	0.046
6		Stack & 3D U-Net [15]	0.155	1.444	5.762	0.191	0.772	**0.040**
7		Concat & 2D Convs	0.138	1.124	5.110	0.174	0.815	0.043
8		**Ours** CCF w./o. R_{multi}	0.124	0.939	4.610	0.154	0.855	0.043
9		**Ours** CCF w./o. R_{mono}	0.123	0.926	4.545	0.153	0.861	0.043
10		**Ours** CCF w./ only intra-cue self-attention	0.122	0.896	4.544	0.152	0.860	0.042
11		**Ours** CCF w./o. residual connection	0.130	0.961	4.616	0.157	0.840	0.048
12		**Ours** depth module w./o. I_t	0.126	0.954	4.636	0.155	0.844	0.042
13		**Ours** CCF - full	**0.118**	**0.835**	**4.297**	**0.146**	**0.871**	0.043



36 manydepth
37 monorec
有点比较好笑的是反而恶化了整体的表现

Thanks