PRCV 2023

空间感知与计算实验室 spAtial Sensing & Computing Lab

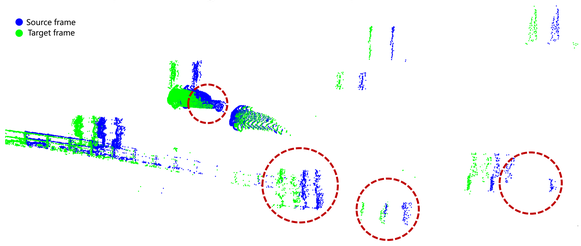
GMA3D: Local-Global Attention Learning to Estimate Occluded Motions of Scene Flow

Zhiyang Lu and Ming Cheng*

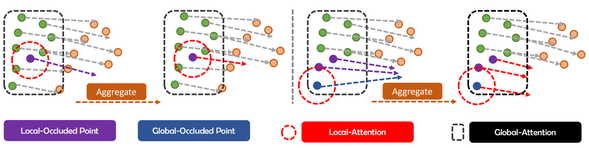
Xiamen University, Xiamen, 361005, China

Introduction

In real-world scenes, due to the sparsity and uneven distribution of point clouds, occlusion points often exist between two consecutive frame point clouds. These occlusion points can be broadly categorized into two types: **local occlusion points** and **global occlusion points.**



- Inferring the motion information of the **local occlusion** points leveraging the local rigidity of motion.
- Utilizing global attentive matching to estimate the scene flow of the **global occlusion** points.



- The Transformer architecture, commonly leveraged to model the global features in sequences, can be adapted to scene flow data which is treated as consecutive frames of sequence data by employing the **offset aggregation** methods.

Experimental Results

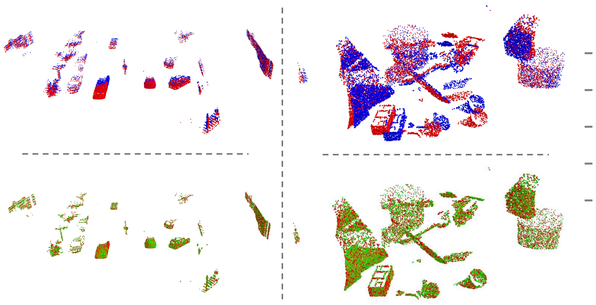
Table 1. Quantitative results of our GMA3D on the version of occlusion datasets Flyingthings3Do and KITTIo. All the models in the table are only trained on the occluded Flyingthings3Do and tested on the occluded KITTIo without any fine-tune.

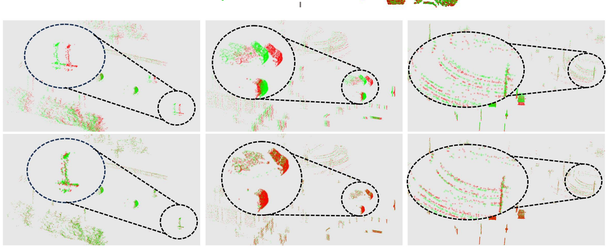
Dataset	Method	Sup.	EPE3D(m)↓	Acc Strict↑	Acc Relax↑	Outliers↓
FlyThings3Do	PointPWC-Net[28]	Self	0.3821	0.0489	0.1936	0.9741
	3D-OGFLow[14]	Self	0.2796	0.1232	0.3593	0.9104
	PointPWC-Net[28]	Full	0.1552	0.4160	0.6990	0.6389
	SAFIT[20]	Full	0.1390	0.4000	0.6940	0.6470
	OGSF[13]	Full	0.1217	0.5518	0.7767	0.5180
	FESTA[24]	Full	0.1113	0.4312	0.7442	–
	3D-OGFLow[14]	Full	0.1031	0.6376	0.8240	0.4251
	Estimation&Propagation[25]	Full	0.0781	0.7648	0.8927	0.2915
	Bi-PointFlowNet[1]	Full	0.0730	0.7910	0.8960	0.2740
	WM[23]	Full	**0.0630**	0.7911	0.9090	0.2790
	GMA3D(Ours)	Full	0.0683	**0.7917**	**0.9171**	**0.2564**
KITTIo	PointPWC-Net[28]	Self	0.3821	0.0489	0.1936	0.9741
	3D-OGFLow[14]	Self	0.2796	0.1232	0.3593	0.9104
	SCOOP[8]	Self	0.0470	**0.9130**	0.9500	0.1860
	PointPWC-Net[28]	Full	0.1180	0.4031	0.7573	0.4966
	SAFIT[20]	Full	0.0860	0.5440	0.8200	0.3930
	OGSF[13]	Full	0.0751	0.7060	0.8693	0.3277
	FESTA[24]	Full	0.0936	0.4485	0.8335	–
	3D-OGFLow[14]	Full	0.0595	0.7755	0.9069	0.2732
	Estimation&Propagation[25]	Full	0.0458	0.8726	0.9455	0.1936
	Bi-PointFlowNet[1]	Full	0.0650	0.7690	0.9060	0.2640
	WM[23]	Full	0.0730	0.8190	0.8900	0.2610
	GMA3D(Ours)	Full	**0.0385**	0.9111	**0.9652**	**0.1654**

Table 2. Quantitative results of our GMA3D on the version of non-occluded datasets FT3Ds and KITTIs.

Dataset	Method	Sup.	EPE3D(m)↓	Acc Strict↑	Acc Relax↑	Outliers↓
FlyThings3Ds	PointPWC-Net[28]	Full	0.0588	0.7379	0.9276	0.3424
	FLOT[15]	Full	0.0520	0.7320	0.9270	0.3570
	PV-RAFT(baseline)[26]	Full	0.0461	0.8169	0.9574	0.2924
	GMA3D(Ours)	Full	**0.0397**	**0.8799**	**0.9727**	**0.2293**
KITTIs	PointPWC-Net[28]	Full	0.0694	0.7281	0.8884	0.2648
	FLOT[15]	Full	0.0560	0.7550	0.9080	0.2420
	PV-RAFT(baseline)[26]	Full	0.0560	0.8226	0.9372	0.2163
	GMA3D(Ours)	Full	**0.0434**	**0.8653**	**0.9692**	**0.1769**

Visualization on the KITTIs dataset (left) and FlyThings3Ds dataset (right).



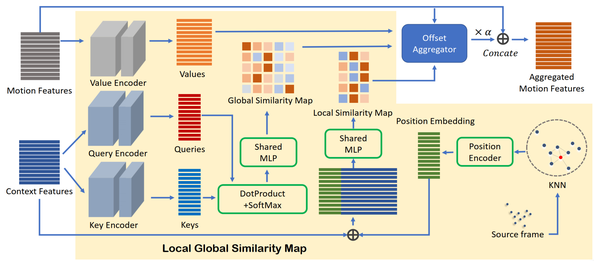


Qualitative results on the KITTIo dataset of occluded version :

- Top: source point cloud (green) and target point cloud (red).
- Bottom: warped point cloud (green)utilizing the predicted flow and target point cloud (red).

Methodology

Utilizing the LGSM (Local Global Similarity Map) module to quantify the semantic similarity between the local and global features of the point cloud in the first frame:

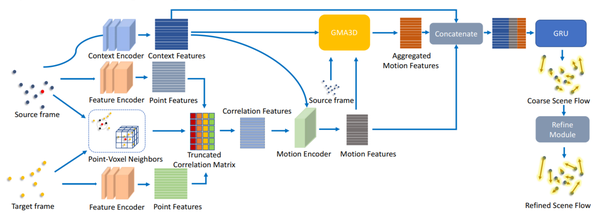


- **Global Similarity Map**: mapping the context features to the query feature map and key feature map, the global attention map produced by the dot product is applied with Softmax and MLP .
- **Local Similarity Map**: the local similarity map is calculated by utilizing a Local GNN to model point pairs in both feature space and Euclidean space.

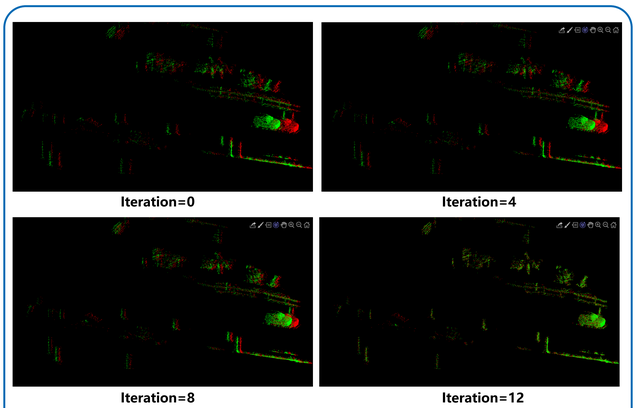
$$\mathbf{g}_{offset} = h_{l,b,r}(\mathbf{y} - (\mathbf{g}_{local} + \mathbf{g}_{global})),$$

$$\tilde{y_i} = y_i + \alpha(\mathbf{g}_{offset}),$$

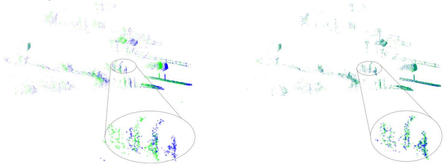
Offset Aggregation: maps are weighted sum with the motion features projected from the value encoder through the Offset Aggregator (OA) to output local and global aggregated motion features.



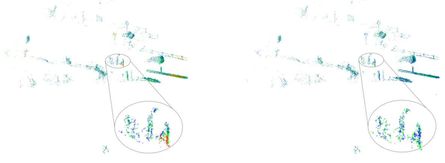
Our module can smoothly integrate into various scene flow backbones for more robust scene flow inference, such as PV-RAFT, FlowStep3D and so on.



Visualization of various iterative times. The green points represent translated source frame point cloud S + f while red points stand for target frame point cloud T.



(a) Input point clouds (b) Ground-truth flows



(c) OGSF (d) Ours(GMA3D)

The red points indicate the points that have been warped by incorrect flows (the first frame point cloud +incorrect scene flow whose EPE >0.1m) .

Conclusion

In this study, we introduce GMA3D, a novel approach to address motion occlusion in scene flow. GMA3D leverages a local-global motion aggregation strategy to infer motion information for both local and global occluded points within the source point cloud. Furthermore, GMA3D can enforce local motion consistency for moving objects, which proves advantageous in estimating scene flow for non-occluded points as well. Experimental results obtained from datasets involving both occluded and non-occluded scenarios demonstrate the superior performance and generalization capabilities of our GMA3D module.