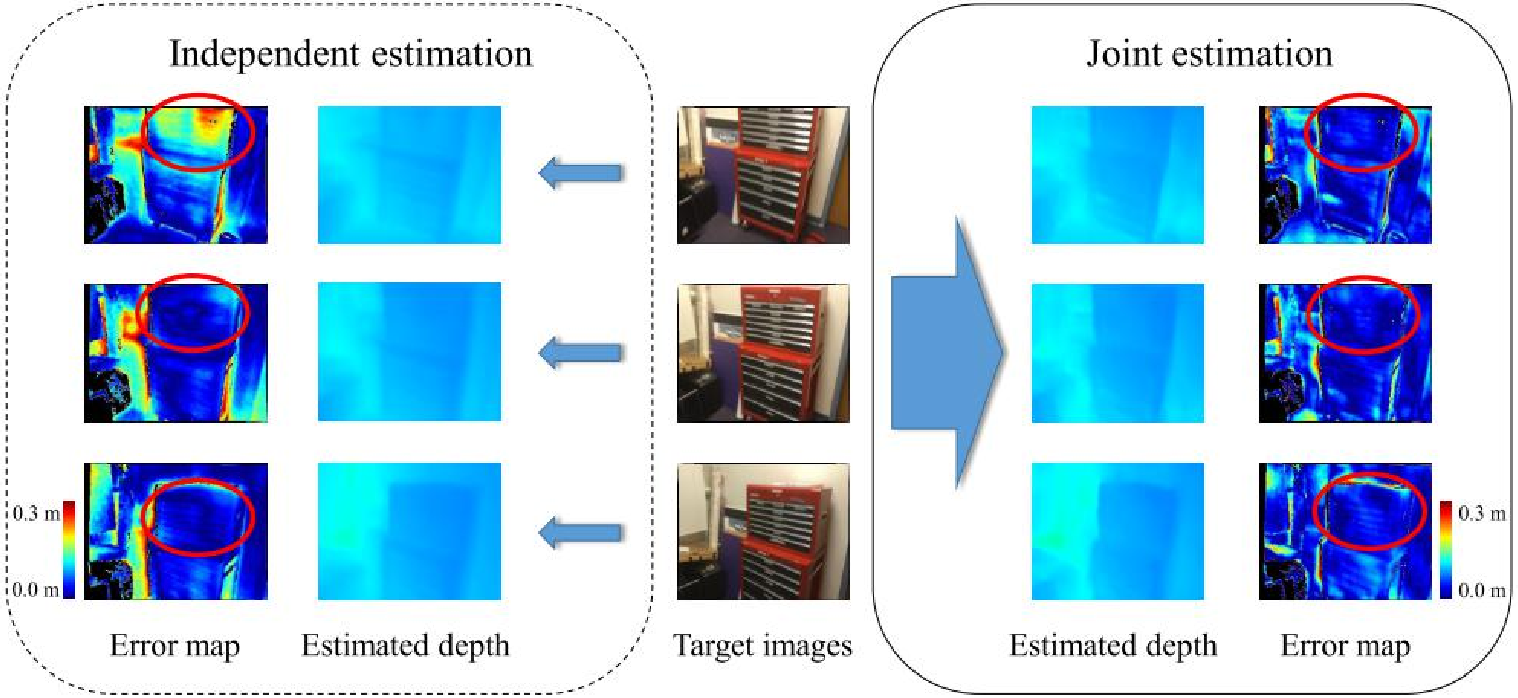
Multi-view Depth Estimation using Epipolar Spatio-Temporal Networks

Xiaoxiao Long[1] Lingjie Liu[2] Wei Li[3] Christian Theobalt[2] Wenping Wang[1,4]
[1]The University of Hong Kong [2]Max Planck Institute for Informatics
[3]Inceptio [4]Texas A&M University

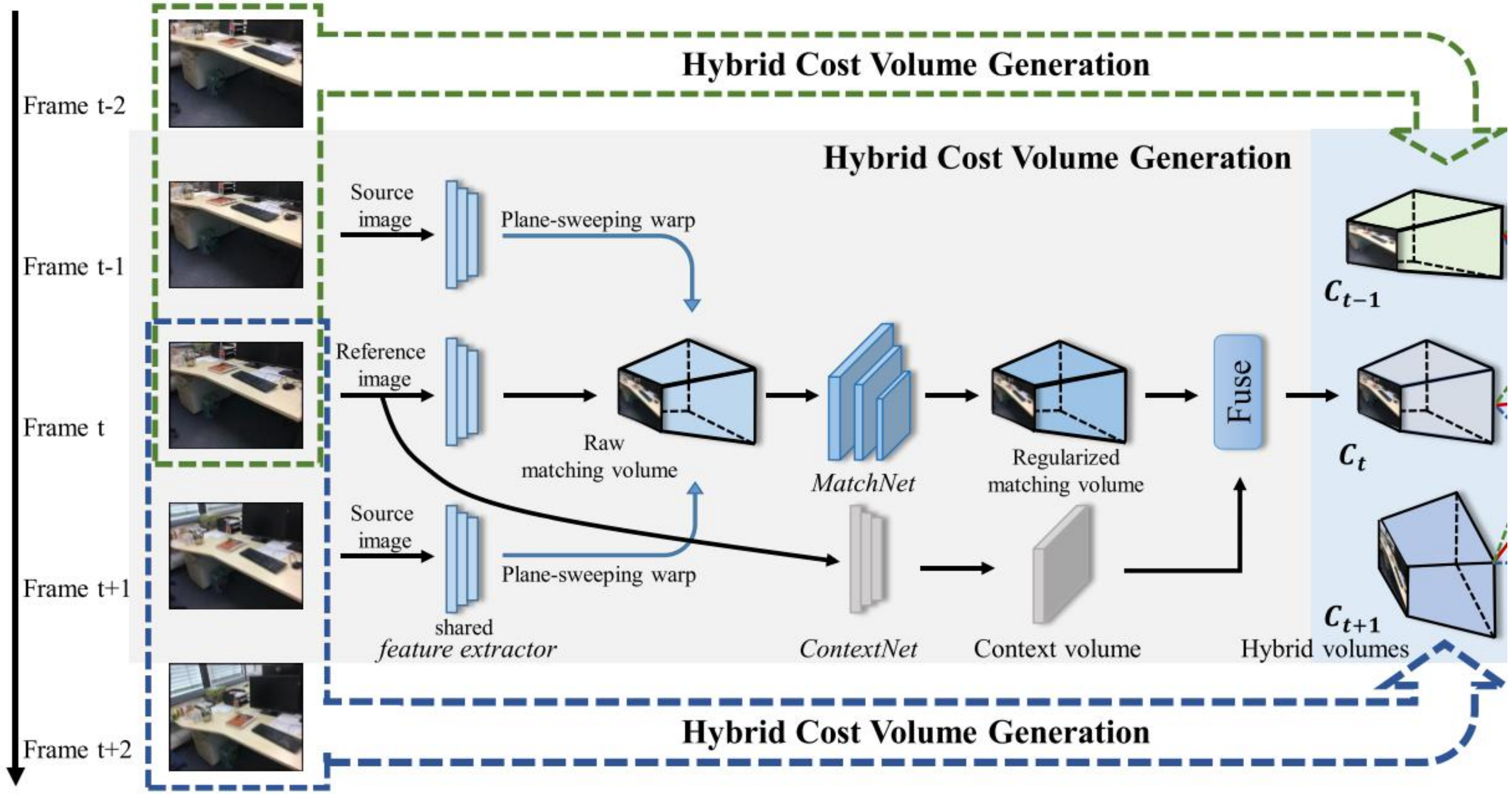
CVPR 2021

Motivation

- SOTA models mostly adopt a fully 3D convolution network for cost regularization and therefore require high computational cost.
- Most works estimate depth maps of individual video frames independently, without taking into consideration the strong relationship between frames.



Method: Hybrid Network



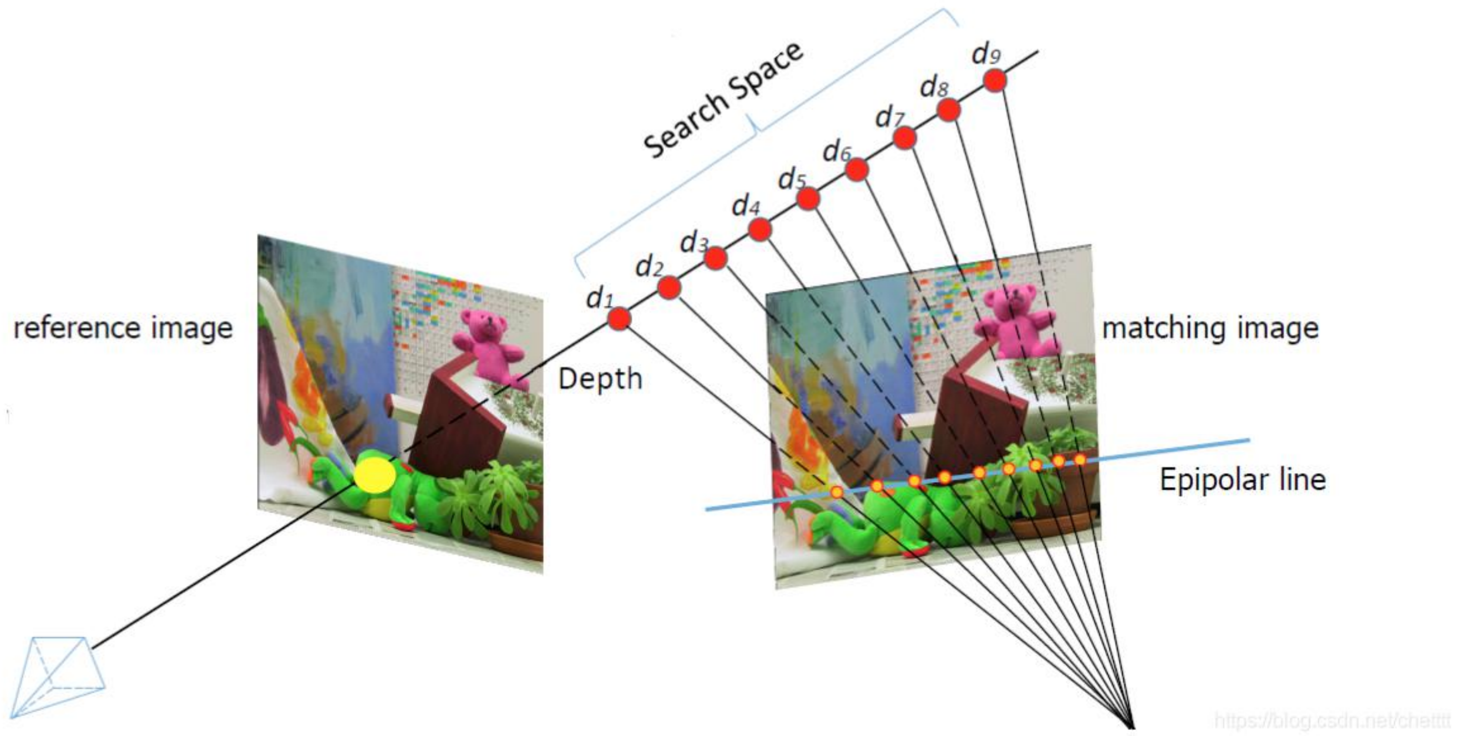
How to fuse?

Treat <u>context volume</u> as 1 channel in <u>regularized matching volume</u>. (Concatenate)
(global information) (local feature)

Method: EST Transformer

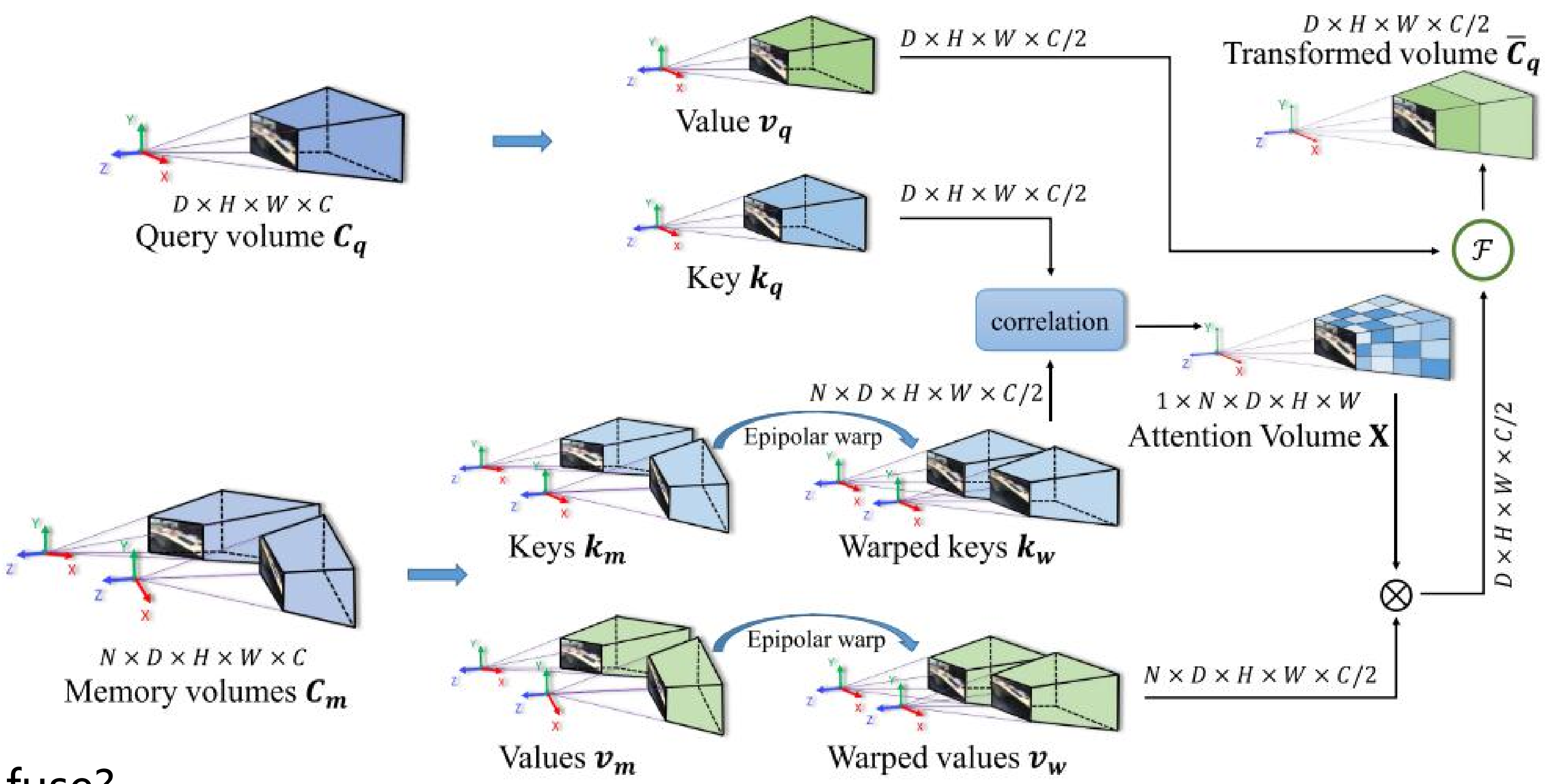
Photometric Consistency Assumption

A 3D point in world space will be projected into visible images, and the image textures near their projections should bear high similarity.



Epipolar Search

Method: EST Transformer

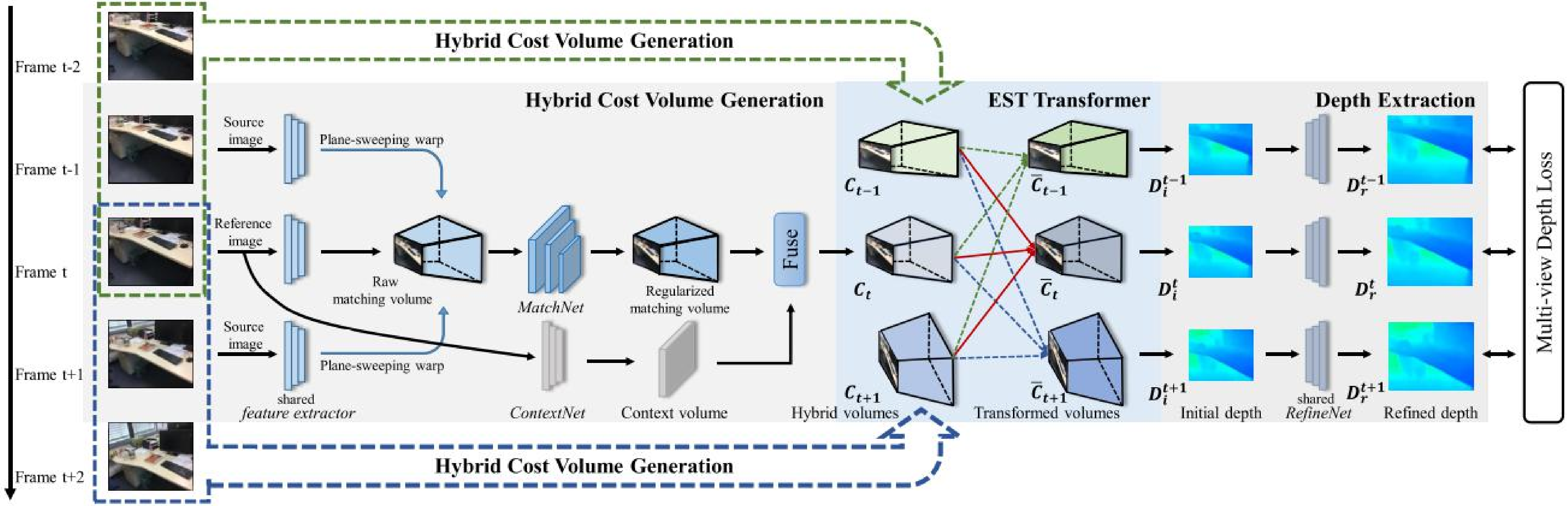


How to fuse?

$$f(v_q, y) = w \odot y + (1 - w) \odot g(v_q, r \odot y), y = \sum_{i=1}^{N} x_i v_w^i$$

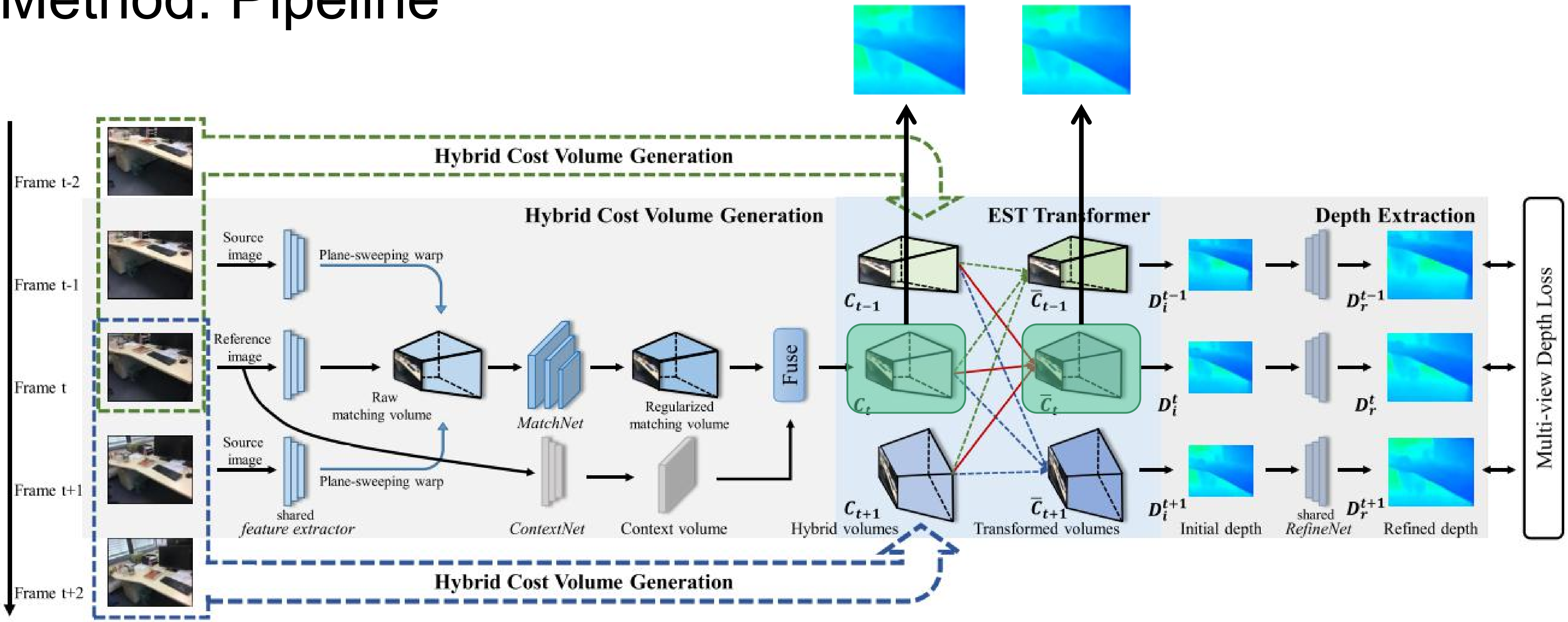
w、r是可学习的权重；g表示卷积操作

Method: Pipeline



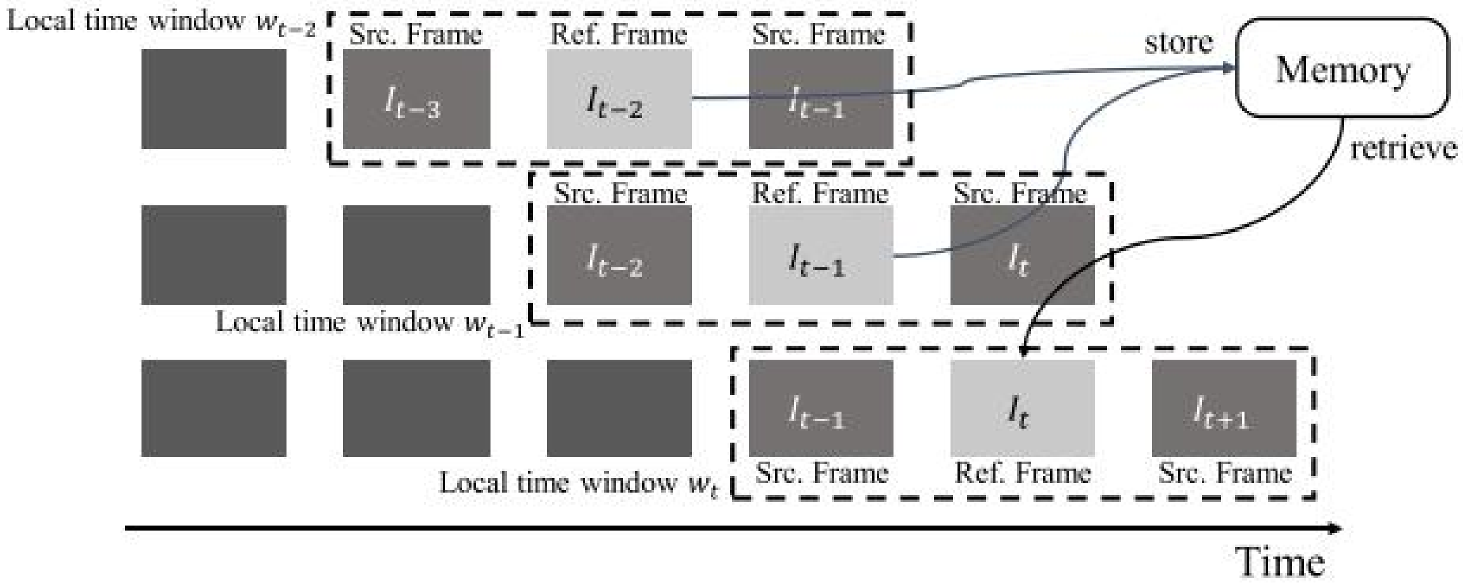
Takes a short video sequence with 5 frames as input and jointly estimate the depth maps of 3 target images with short-term temporal coherence.

Method: Pipeline



$$loss = \frac{1}{N} \sum_{s=0}^{3} \sum_{i=1}^{N} \lambda^{s-3} \left\| \mathbf{D}_s^i - \hat{\mathbf{D}}_s^i \right\|_1$$

Method: Inference acceleration

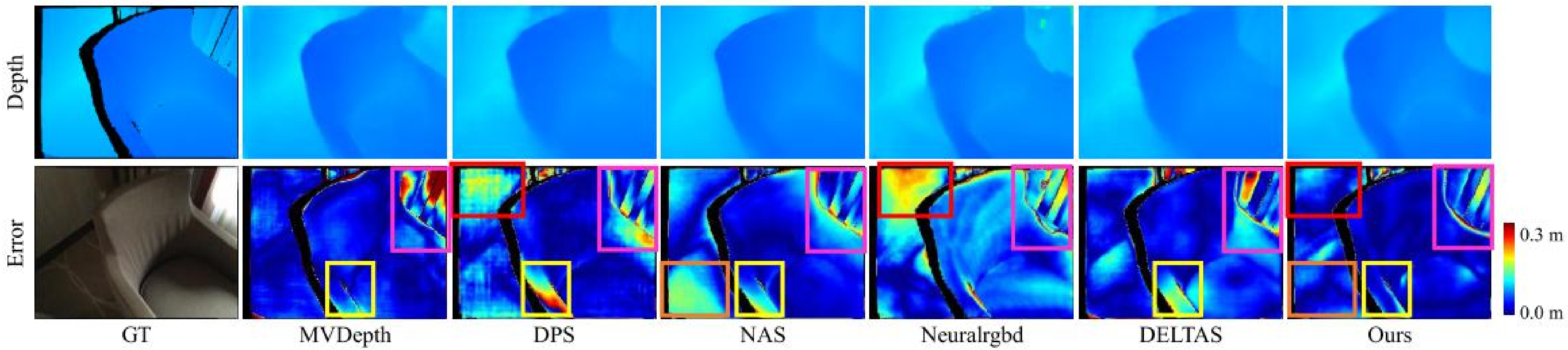


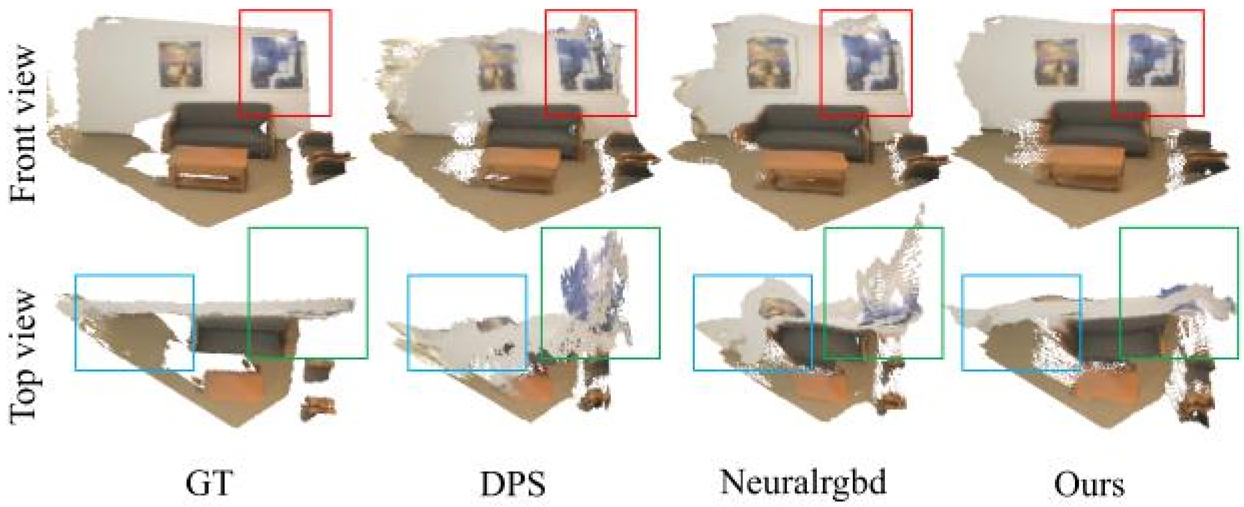
- Retrieve relevant values from a memory space storing the pairs of keys and values of N past frames.
- Slide Window, when window moves on, the memory space will be also updated accordingly.

Experiment: Depth accurary

Range	Method	ScanNet					7scenes				
		Abs Rel	Abs	Sq Rel	RMSE	$\sigma < 1.25$	Abs Rel	Abs	Sq Rel	RMSE	$\sigma < 1.25$
10m	MVDepth [32]	0.1167	0.2301	0.0596	0.3236	84.53	0.2213	0.4055	0.2401	0.5154	67.33
	MVDepth-FT	0.1116	0.2087	0.0763	0.3143	88.04	0.1905	0.3304	0.1319	0.4260	71.93
	DPS [17]	0.1200	0.2104	0.0688	0.3139	86.40	0.1963	0.3471	0.1970	0.4625	72.51
	DPS-FT	0.0986	0.1998	0.0459	0.2840	88.80	0.1675	0.2970	0.1071	0.3905	76.03
	NAS [20]	0.0941	0.1928	0.0417	0.2703	90.09	0.1631	0.2885	0.1023	0.3791	77.12
	CNM [23]	0.1102	0.2129	0.0513	0.3032	86.88	0.1602	0.2751	0.0819	0.3602	76.81
	DELTAS [30]	0.0915	0.1710	0.0327	0.2390	91.47	0.1548	0.2671	0.0889	0.3541	79.66
	Ours-EST(concat)	0.0818	0.1536	0.0301	0.2234	92.99	**0.1458**	0.2554	0.0745	0.3436	79.82
	Ours-EST(adaptive)	**0.0812**	**0.1505**	**0.0298**	**0.2199**	**93.13**	0.1465	**0.2528**	**0.0729**	**0.3382**	**80.36**
5m	Neuralrgbd [21]	0.1013	0.1657	0.0502	0.2500	91.60	0.2334	0.4060	0.2163	0.5358	68.03
	Ours-EST(concat)	0.0811	0.1469	0.0279	0.2066	93.19	**0.1458**	0.2554	0.0745	0.3435	79.82
	Ours-EST(adaptive)	**0.0805**	**0.1438**	**0.0275**	**0.2029**	**93.33**	0.1465	**0.2528**	**0.0729**	**0.3382**	**80.36**

Experiment: Depth accurary





Experiment: Temporal coherence & complexity analysis

Table 2. Comparison of temporal coherence over ScanNet dataset with depth evaluation range $0 \sim 5m$.

Metric	DPS[17]	NAS [?]	Neuralrgbd [21]	DETALS [30]	Ours
Abs	0.1887	0.1823	0.1642	0.1650	**0.1432**
Std	0.2243	0.2177	0.1848	0.1886	**0.1673**

Table 3. Memory and computation complexity analysis.

Model	Params	MACs	Memory	Time
DPS [17]	**4.2M**	442.7G	**1595M**	337ms
NAS [20]	18.0M	527.7M	1689G	212ms
Neuralrgbd [21]	<u>5.3M</u>	616.6G	2027M	195ms
DELTAS [30]	124.6M	**98.6G**	2395M	495ms
Ours-ESTM	36.2M	176.9G	1799M	**71ms**

Experiment: Ablation study

Table 4. The usefulness of *ContextNet* and epipolar transformer. We test models with various settings on SUN3D [37].

Cont.	Trans.	Inference type	Abs	Sq Rel	RMSE	$\sigma < 1.25$
✗	✗	Independent	0.3333	0.0994	0.4897	80.89
✗	✓	Joint	0.3429	0.1291	0.4927	81.36
✗	✓	ESTM	0.3319	0.1073	0.4822	81.43
✓	✗	Independent	0.3220	0.0897	0.4657	82.82
✓	✓	Joint	**0.3133**	**0.0883**	**0.4556**	**83.52**
✓	✓	ESTM	0.3137	0.0884	0.4554	83.43

Independent: without EST transfomer

Joint: with EST transfomer

ESTM: estimate depth sequentially

Table 5. The effect of different size of ESTM memory space. Experiments are done over 7scenes [29].

Memory size	Abs Rel	Abs	Sq Rel	RMSE	$\sigma < 1.25$
1	0.1530	0.2632	0.783	0.3494	79.07
2	0.1465	0.2528	0.0729	0.3382	80.36
3	**0.1460**	**0.2520**	**0.0727**	**0.3376**	**80.44**
4	0.1461	0.2521	0.0728	0.3377	80.44

Table 6. EST transformer vs RNN. We replace the EST transformer with Gated Recurrent Units in our model.

Model	ScanNet		SUN3D	
	Abs	$\sigma < 1.25$	Abs	$\sigma < 1.25$
Ours-RNN	0.1680	92.67	0.3401	82.33
Ours-ESTM	**0.1505**	**93.13**	**0.3137**	**83.52**